Ontbrekende waarnemingen en statistische pakketten

TUE-RC 85499

Eindhoven University of Technology

Computing Centre Note 52

Ontbrekende Waarnemingen en Statistische Pakketten

Jan B. Dijkstra

Samengesteld voor de najaarsvergadering van de SWS op 13 november 1991 in Utrecht.

Ontbrekende Waarnemingen en Statistische Pakketten

Jan B. Dijkstra

Samenvatting

Aan de hand van een voorbeeld betreffende groeigegevens zijn zeven statistische pakketten vergeleken door even zovele onderzoekers. De dataset bevatte ontbrekende waarnemingen in de response-variabele van een door trial and error te vinden lineair model. Als referentie wordt met GLIM nagegaan hoe hinderlijk de missing data zijn bij het vinden van de meest geschikte aanpassing. Daarna worden zes andere pakketten hiermee vergeleken. Voor dit onderdeel krijgt geen enkel pakket een onvoldoende.
Een tweede onderzoek in deze studie betreft ontbrekende waarnemingen in de predictorvariabelen van een lineair model. Deze situatie resulteert in problemen die niet middels ieder pakket op gebruikersvriendelijke en inhoudelijk verdedigbare wijze kunnen worden opgelost.
Het derde en laatste onderdeel betreft een kunstmatig voorbeeld. Een data-matrix is voor 90.9 procent gevuld en toch is het niet met ieder pakket mogelijk de parameters te schatten van een eenvoudig model dat bij deze waarnemingen hoort.

De aanleiding

Op 24 januari 1991 hield de Sectie Statistische Programmatuur van de Vereniging Voor Statistiek in Utrecht haar jaarvergadering. Zoals gebruikelijk was er naast een huishoudelijk deel als lokkertje voor de leden een lezingencyclus georganiseerd rond een centraal thema. Dit keer was dat de missing data handling in verschillende pakketten. Onderstaande tabel geeft het programma weer.

Software	Spreker
GLIM en Probleem	Jan B. Dijkstra
SPSS	Lex Jansen
SAS	Henk van der Knaap
STATA	Ger Snijkers
BMDP	Pierre Debets
SYSTAT	Jan Raatgever
SURFOX	Jan Hooft van Huijsduijnen

De volgende twee tabellen betreffen gegevens van meisjes en jongens op de leeftijden van 8, 10, 12 en 14 jaar. De gegevens zijn afkomstig van Potthoff en Roy (1964) en opgenomen in het handboek voor missing data handling van Little en Rubin (1987). De getallen geven de afstand van het centrum van de hypofyse (hersenaanhangsel) weer tot aan het punt waarop de bovenkaak zich splitst. Omdat van ieder kind metingen

bekend zijn op vier equidistante punten in de tijd zijn dit in feite groeigegevens. Voor het nu volgende experiment worden de data tussen haakjes als ontbrekend beschouwd.

Meisje	8	10	12	14
1	21	20	21.5	23
2	21	21.5	24	25.5
3	20.5	(24)	24.5	26
4	23.5	24.5	25	26.5
5	21.5	23	22.5	23.5
6	20	(21)	21	22.5
7	21.5	22.5	23	25
8	23	23	23.5	24
9	20	(21)	22	21.5
10	16.5	(19)	19	19.5
11	24.5	25	28	28

Jongen	8	10	12	14
1	26	25	29	31
2	21.5	(22.5)	23	26.5
3	23	22.5	24	27.5
4	25.5	27.5	26.5	27
5	20	(23.5)	22.5	26
6	24.5	25.5	27	28.5
7	22	22	24.5	26.5
8	24	21.5	24.5	25.5
9	23	20.5	31	26
10	27.5	28	31	31.5
11	23	23	23.5	25
12	21.5	(23.5)	24	28
13	17	(24.5)	26	29.5
14	22.5	25.5	25.5	26
15	23	24.5	26	30
16	22	(21.5)	23.5	(25)

In dit voorbeeld zijn er alleen ontbrekende waarnemingen in de gemeten afstand die hier opgevat kan worden als de response-variabele in een lineair model met de leeftijd, sexe en het individuele rangnummer van de persoon als predictoren. Verschillende aanpassingen zullen worden geprobeerd. In deze uitzonderlijke situatie (waarin we de waarde van de als ontbrekend behandelde waarneming wel degelijk kennen) is een wat ongebruikelijk kwaliteitscriterium haalbaar: de kwadratensom voor de tien missing data (KSMD).

Behandeling met GLIM

Veronderstel dat de ontbrekende waarnemingen gecodeerd zijn met nul of negatieve waarden. Die keuze is redelijk omdat elke echte waarneming positief dient te zijn. We beginnen met een zeer eenvoudige aanpassing: de afstand Y wordt in een lineair model voorspeld uit de sexe en de leeftijd. Een grafische

representatie van deze aanpassing bestaat uit twee lijnen. De vertikale as bevat de afstand en de horizontale de leeftijd. Iedere sexe heeft zijn eigen intercept, maar de helling is gemeenschappelijk, zodat de lijnen evenwijdig zijn. In GLIM-taal krijgen we het volgende:

```
$CALC W=(Y>0)
$WEIGHT W
$YVAR Y
$FIT SEX+AGE
```

De vector W bevat nu enen voor de waarnemingen en nullen voor de missing data. Met deze gewichten wordt de aanpasing gedaan met de default-instelling van GLIM: een lineair model met normale fouten en constante variantie die geschat moet worden. Na deze fit bevat de systeem-vector %FV (fitted values) de aangepaste waarden. Niet alleen voor de waarnemingen, maar ook voor de missing data. Hiermee kan het kriterium KSMD worden berekend. Dit heeft de matige waarde van 34.04 in deze aanpassing met slechts 3 parameters.

Alternatieven

1. Gemiddelde voor iedere sexe en leeftijd (8 parameters) geeft KSMD=39.76. Hiervan viel overigens ook weinig te verwachten omdat Little en Rubin de gaten in de data-matrix niet helemaal aselect hadden gekozen: Missing data werden vooral in het tiende levensjaar gegenereerd als de waarde in het achtste levensjaar laag uitviel. Daarom kan meer worden verwacht van modellen die aan ieder individu een eigen intercept toekennen.
2. Twee lijnen met eigen intercept en helling (4 parameters) geeft KSMD=35.61 en dat is conform de verwachting niet veel beter.
3. Het eenvoudiger model met alleen sexe en leeftijd uit de vorige paragraaf (3 parameters) is iets minder slecht.
4. Ieder individu krijgt zijn eigen intercept. De leeftijd wordt lineair, kwadratisch en tot de derde macht meegenomen. Van alle drie deze termen wordt de interactie met de sexe beschouwd (27+3+3=33 parameters). Dit geeft KSMD=20.55. Een sprong voorwaarts, maar we zijn er nog niet.
5. Ieder individu krijgt een eigen helling en intercept (27+27=54 parameters). KSMD verbetert maar weinig tot 20.35.
6. Op (4) is een achterwaartse eliminatie toegepast met onbetrouwbaarheid 5 procent. Dit resulteert in 27 intercepts, de leeftijd in het kwadraat en interactie hiervan met sexe (29 parameters). Een enorme verbetering voor KSMD; de waarde is nu 13.03.
7. Ieder individu heeft een eigen intercept, maar allen hebben dezelfde helling (28 parameters). Tot verrassing van de auteur dezes leverde dit een verbetering op: KSMD=10.79. Indien hier niet de waarde van de missing data bekend zou zijn (en KSMD dus niet berekend kon worden) dan zou men niet gauw voor deze aanpassing kiezen in vergelijking met (6). De determinatiecoefficient is namelijk bij (7) 0.8187 en bij (6) 0.8364. Bovendien bevat (6) geen niet-significante prediktoren.
8. De winnaar is een model met 27 intercepts en 2 hellingen (29 parameters). KSMD=10.66 en dat is gunstiger dan de waarde bij (6). De determinatiecoefficient is echter 0.8322 en dat is minder goed dan het resultaat bij (6). Zonder voorkennis omtrent de waarde van de missing data zou men dus voor (6) kunnen kiezen in plaats van voor (8). Maar het lineaire model is eenvoudiger dan het

kwadratische en de winst van de laatste maar gering. De preferentie voor (8) is dus ook verdedigbaar zonder gebruikmaking van KSMD.

Er zijn nog betere aanpassingen mogelijk (niet-lineaire regressie, Box-Cox transformaties en dergelijke). De winst is echter gering en de motivering aanvechtbaar. Voor andere meningen hierover wordt de lezer naar het uitgebreide artikel in Kwantitatieve Methoden verwezen.

De predictoren

Ideaal in regressie-situaties is dat de predictorwaarden gekozen worden en vervolgens zeer nauwkeurig ingesteld, zodat missing data eigenlijk alleen maar in de response-variabele kunnen voorkomen. De praktijk gooit echter vaak roet in het eten. Veel regressies worden uitgevoerd met gemeten predictoren, zodat ook daar ontbrekende waarnemingen te vrezen zijn. De enige eenvoudig toepasbare mogelijkheid met GLIM is hierbij listwise deletion. Dit houdt in dat een observatie in zijn geheel genegeerd wordt zodra voor tenminste een variabele (predictor of response) een missing data code ontmoet wordt. Uiterst klungelig bij GLIM is dat deze mogelijk alleen middels WEIGHT gerealiseerd kan worden. Daardoor ontstaan ook in dit geval aangepaste waarden voor de missing data. Deze worden echter berekend op basis van de gekozen codering. En dat is onverdedigbaar.

Een eerste stap bij de aanpassing van een regressie-model kan de berekening van de covariantie-matrix van de predictoren zijn. Hierbij gaat pairwise deletion aanzienlijk zuiniger om het de beschikbare informatie dan listwise deletion. Bij de eerstegenoemde methode heeft het ontbreken van een waarneming voor variabele V1 geen invloed op de berekening van de covariantie tussen V2 en V3. Een nog slimmere variant op Pairwise Deletion is afkomstig van Boas (1967). Deze zal hier echter niet worden besproken. Beide methoden uit deze alinea zijn helaas niet beschikbaar in GLIM.

Sommige pakketten hebben allerlei technieken om de gaten in een data-matrix te vullen met schattingen (eventueel met kunstmatig toegevoegde ruis). GLIM heeft op dit gebied niets te bieden, maar de ervaren gebruiker kan dit soort vultechnieken zonder al te veel moeite zelf in GLIM-taal coderen.

Samenvattend kan men zeggen dat in GLIM wel degelijk mogelijkheden aanwezig zijn voor de behandeling van ontbrekende waarnemingen, maar de gebruiker moet zelf wel veel programmeren als de gaten in de predictor-ruimte zitten. Tekenend voor GLIM is het ontbreken van de entry Missing Data in de index.

Een ontaard geval

We komen nu toe aan de derde situatie. Denk aan een response-variabele en 10 predictoren. Veronderstel dat er 20 waarnemingen zijn. Een volledige data-matrix zou dus 220 getallen bevatten. Stel nu dat voor iedere observatie in een willekeurige predictor de waarde ontbreekt. De data-matrix bevat dan dus 20 gaten en is derhalve voor 90.9 procent gevuld. Toch kan GLIM hier niets mee aanvangen tenzij men gebruik maakt van een zelf te programmeren vultechniek. Verderop zullen we zien dat andere pakketten in dit opzicht aanzienlijk gunstiger scoren.

Numerieke problemen

GLIM kent alleen Listwise Deletion zodat na het opschonen van de data-matrix zeer beproefde algoritmen kunnen worden toegepast op de resterende rijen. Enkele van de in volgende paragrafen te behandelen pakketten bieden Pairwise Deletion (of varianten hierop). Hoewel elk element van de hierdoor berekende covariantie-matrix correct geschat wordt, is deze matrix niet noodzakelijk semi positief definiet. In

tegenstelling tot bij de volledige dataset en bij Listwise Deletion kunnen negatieve eigenwaarden ontstaan en daar hebben methoden als Choleski Decompositie nogal veel moeite mee. Faciliteiten voor missing data handling kunnen dus de verdere analyse behoorlijk in de wielen rijden. En dat hoeft niet alleen te resulteren in een abortie van het rekenproces; onjuiste antwoorden behoren ook tot de mogelijkheden (onderschatting restvariantie, te hoge t-waarden of F-waarden en dergelijke).

SPSS

SPSS kent user-missing en system-missing values. Voor de eerste categorie kan de gebruiker een code kiezen en de tweede ontstaat in een SPSS sessie en wordt in de uitvoer met een punt meegegeven. De verschillende versies van SPSS zijn hierbij niet erg consistent. Zo geldt voor een mainframe 0*missing=0 en voor de PC 0*missing=missing.

Naast de gebruikelijke listwise deletion kent SPSS ook pairwise deletion en mean substitution. Als bij pairwise deletion de resulterende correlatie-matrix niet positief definiet is, dan wordt in de uitvoer een passende waarschuwing afgedrukt. Het ontbreken van deze waarschuwing wil niet zeggen dat alles in orde is. Verbeek (1979) merkte reeds op dat pairwise deletion kan resulteren in grotere varianties voor de schatters dan listwise deletion.

Voor mean substitution wordt in de manual gewaarschuwd omdat dit kan leiden tot te kleine varianties en dergelijke. Meer algemeen geeft de manual de volgende (zeer terechte) waarschuwing: Missing-value treatment problems should not be treated lightly. You should always select a missing-value treatment based on careful examination of the data and not leave the choices up to system defaults.

Alle drie de problemen uit de inleiding kunnen met SPSS worden aangepakt. Voor een automatisme dienen de gaten random verdeeld te zijn. Dit kan in veel gevallen met de procedure T-TEST worden geverifieerd.

SAS

SAS kent de gebruikelijke mogelijkheden om met missing data om te gaan. Evenals bij SPSS worden deze in de uitvoer weergegeven met een punt. De mogelijkheden voor data-handling zijn bij SAS zeer flexibel. Bij het voorbeeld van de groeigegevens kan met aan twee strukturen denken: (1) het model waarbij voor elke leeftijd een aparte variabele is gereserveerd en (2) het model waarbij de leeftijd een der predictorvariabelen is. Deze strukturen zijn met de procedure TRANSPOSE eenvoudig in elkaar over te voeren.

Grafische mogelijkheden zijn bij SAS goed vertegenwoordigd. Een exploratief onderzoek van de struktuur van de missing data is in beide bovengenoemde data-representaties eenvoudig uit te voeren.

De procedure PRINQUAL (Gebaseerd op werk van De Leeuw c.s. uit Leiden) kan gebruikt worden om op iteratieve wijze schattingen voor de gaten in een data-matrix te krijgen. Het aantal lege cellen moet dan wel klein zijn, en ze moeten willekeurig over de data-matrix verspreid zijn. Bij het probleem van de groeigegevens is hieraan niet voldaan. Wel liggen hier mogelijkheden voor de twee andere problemen.

Een aantrekkelijke variant op de regressie-aanpak uit de inleiding voor het probleem van de groeigegevens wordt gevormd door een macro voor Box-Cox transformaties op de response-variabele. Ook kan de waarde van KSMD worden gereduceerd door van niet-lineaire regressie gebruik te maken. Omdat voor beide aanpakken echter geen inhoudelijk verdedigbare argumenten lijken te bestaan, zullen ze hier verder buiten beschouwing blijven.

STATA

STATA hanteert uitsluitend listwise deletion bij de afhandeling van ontbrekende waarnemingen bij multivariate analyses. Net als bij GLIM is het na aanpassing van een regressie-model ook hier mogelijk om aangepaste waarden te berekenen voor alle instel-punten (dus ook voor die punten waarbij de waarde van de response-variabele ontbreekt).

Binnen de predictor-ruimte kan men op enigszins gekunstelde wijze gaten vullen op basis van regressieanalyses of (gewogen) gemiddelden. Deze vullingen kan men een tijdelijke status geven; voor verdere analyses kan men ze altijd weer opheffen.

De grafische mogelijkheden van STATA zijn aantrekkelijk. Het is bijvoorbeeld vrij eenvoudig om te zoeken naar een volledige observatie die zo veel mogelijk lijkt op een observatie waaraan de waarde voor een zekere variabele ontbreekt. Dit kan nuttig zijn bij het vullen van de gaten.

Bij het eerste probleem biedt STATA dezelfde mogelijkheden als GLIM. Bij het tweede en derde probleem biedt STATA geen volledig uitgewerkte methoden. De gebruiker kan wel allerlei mogelijkheden zelf construeren.

BMDP

Hoewel BMDP in 1959 oorspronkelijk geintroduceerd werd voor bio-medische analyses, is het inmiddels uitgegroeid tot een algemeen statistisch pakket dat geen slecht figuur slaat in vergelijking met SPSS, SAS en STATA. Naast de gebruikelijke mogelijkheden voor de definitie van user-missing of system-missing values kan men bij BMDP ook nog ranges opleggen aan in te voeren waarden. TOOBIG en TOOSMALL zijn dan bijzondere coderingen voor missing data.

Met TRANSFORM kan men eenvoudig het gemiddelde voor missing data invullen. Interpolatie kan zelfs op meerdere manieren. Naast listwise deletion kent BMDP nog drie alternatieven: (1) gewone pairwise, (2) een variant die de covarianties paarsgewijs berekent maar de standaarddeviaties per variabele afzonderlijk en zo compleet mogelijk en (3) een nog vollediger vorm waarbij de gemiddelden per variabele worden berekend en hierop de covarianties worden gebaseerd.

De problemen uit de inleiding zijn dus alle drie met BMDP goed aan te pakken. Terecht is er in de manual een waarschuwing: An important requirement for valid estimation of the correlation or covariance matrix is that the data be missing completely at random. Er zijn faciliteiten om dit na te gaan.

Na pairwise deletion of varianten hierop is de matrix niet meer noodzakelijk positief definiet. Desgewenst kan deze echter worden getransformeerd tot de dichtstbij gelegen positief definiete matrix.

Soms wordt niet voldaan aan de eis van missing completely at random, maar wel aan de zwakkere eis van missing at random (het ontbreken van gegevens in de ene variabele mag dan afhankelijk zijn van waarden in andere variabelen, maar niet van hun missing data mechanisme). Voor de maximum likelihood methode van Beale en Little biedt BMDP in dat geval ook een programma.

SYSTAT

Dit is weer een algemeen statistisch pakket met de gebruikelijke mogelijkheden voor missing data handling. Voor numerieke variabelen wordt ook hier een punt in de uitvoer gegeven voor een ontbrekende waarde (of spaties voor nominale variabelen). Grafische analyse vooraf is met SYGRAPH zeer goed mogelijk en de faciliteiten zijn ruim bemeten.

Naast listwise deletion is ook pairwise deletion aanwezig met de inmiddels gebruikelijke waarschuwingen in de manual. Het tweede en derde probleem uit de inleiding kunnen dus worden aangepakt. Niet alleen het mogelijk niet positief definiet zijn van de covariantie-matrix wordt gesignaleerd, maar ook wordt gewezen op het probleem van het ongedefinieerde aantal vrijheidsgraden bij de F-toets en de t-toetsen. De gebruiker moet hier zelf een keuze maken (bijvoorbeeld kan men doen alsof het aantal cases gelijk is aan het laagste aantal niet-missende waarden uit de verzameling van alle paren variabelen).

Aantrekkelijker lijkt de mogelijkheid om gaten te vullen door te zoeken naar sterk gelijkende buren (voor wat betreft de variabelen met niet-ontbrekende waarnemingen). In SYSTAT kan men hiervoor gebruik maken van de module voor Cluster Analyse.

SURFOX

SURFOX is een door AB-Onderzoek te Delft ontwikkeld nederlandstalig pakket voor het analyseren van bestanden en het verwerken van enquetes. Het project loopt sinds 1986 en in 1989 kwam een PC-versie op de markt. Hoewel het programma menu-gestuurd is, bestaat er tevens een batch-versie.

Ten aanzien van ontbrekende gegevens maakt SURFOX onderscheid tussen NVT (niet van toepassing) en TOGA (ten onrechte geen antwoord). De laatste categorie vorm het onderwerp van deze notitie.

Bij regressie maakt SURFOX (als enig programma in dit overzicht) geen gebruik van listwise deletion voor de behandeling van missing data. Door automatisch te construeren dummy-variabelen krijgen de betreffende observaties bij de berekeningen een aparte rol toebedeeld.

Het vullen van gaten in een datamatrix kan desgewenst worden uitgebreid met een pseudo-random storingsterm uit de normale verdeling. Er wordt dan afgezien van de beste schatter op record-niveau, maar de onderschatting van de restvariantie (een bezwaar dat aan de meeste vultechnieken kleeft) kan hiermee adequaat worden gecompenseerd.

Voor continue variabelen biedt SURFOX Predictive Mean Matching van Little (1988) waarin een record met missing data wordt gecompleteerd op basis van het hier zoveel mogelijk op gelijkende record zonder missing data. Voor nominale variabelen is gekozen voor Random Hotdeck (de verdeling van de bijgeschatte variabele prevaleert hierbij boven een zo goed mogelijke schatting op record-niveau).

Samenvatting

Bij het eerste voorbeeld (de groeigegevens voor 11 meisjes en 16 jongens) werd door elke onderzoeker en met elk pakket vrij vlot de conclusie bereikt dat 27 intercepts en 2 hellingen de geschiktste aanpassing weergeven. De uitwerking met SAS is als een bijlage bij deze notitie opgenomen. De residu-analyse geeft aanleiding tot enige onrust; toetsing op normaliteit levert een overschrijdingskans op van 0.0022 en dat is te weinig. Grafisch onderzoek (stem-and-leaf plot, boxplot, normal probability plot en een plotje van de residuen tegen de aangepaste waarden) demonstreert dat enkele uitschieters hiervan de oorzaak zijn. Het blijkt dat de groei van jongen nummer 9 tussen het 10e en 12e levensjaar teveel afwijkt, en dat de groei van jongen nummer 13 te geprononceerd is over de hele periode van 8 tot 14 jaar. Gelukkig hebben deze uitschieters niet veel invloed op de uiteindelijke aanpassing (de grootste Cook-statistic bij deze uitschieters is 0.3112 voor jongen nummer 13 op 8-jarige leeftijd en dat is veilig kleiner dan de gebruikelijke grens van 1).

Voor het tweede en derde probleem kan men gebruik maken van pairwise deletion. SPSS, BMDP en SYSTAT hebben hier standaard faciliteiten voor. De grootste bezwaren tegen deze aanpak worden alleen door

BMDP redelijk goed ondervangen.

Voor vultechnieken moet men vaak zelf programmeren. Elk pakket biedt mogelijkheden, maar werkelijk handig gaan de eenvoudigste technieken alleen met SPSS, BMDP en SURFOX. Kwalitatief aantrekkelijk lijken de vultechnieken die gebruik maken van zoveel mogelijk overeenstemmende complete observaties. In STAT, SYSTAT en SURFOX gaat dit eenvoudig. Met SPSS, SAS en BMDP is het ook mogelijk, maar het vergt wel enig handmatig ingrijpen van de gebruiker.

Litteratuur

NAG Central Office (1987) The GLIM Release 3.77 Manual (Edition 2)
The Numerical Algorithms Group, Oxford

Boas, J. (1967) A note on the estimation of the covariance between two random variables using extra information on the separate variables
Statistica Neerlandica 21

Little, R.J.A. and D.B. Rubin (1987) Statistical analysis with missing data
Wiley series in Probability and Mathematical Statistics, New York

Potthoff, R.F. and S.N. Roy (1964) A generalized multivariate analysis of variance model useful especially for growth curve problems
Biometrika (51) 313-326

Norusis, M.J. (1988) SPSS/PC+ V2.0 Base Manual
SPSS Inc. Chicago

Verbeek, A. (1979) Ontbrekende scores bij het schatten van covariantie-matrices
VVS Bulletin (12-2) 38-47

J.B. Dijkstra, L. Jansen, H. van der Knaap, G, Snijkers, P. Debets, J. Raatgever, J. Hooft van Huijsduijnen (1991) Missing data handling in verschillende pakketten
Verschijnt binnenkort in Kwantitatieve Methoden

SAS/STAT Guide for Personal Computers (1987) Version 6 Edition
SAS Institute, Cary

Computing Resource Center (1990) STATA 2.1 Reference Manual
Los Angeles, California

Snijkers, G.J.M.E. (1991) Ontbrekende gegevens in STATA
CBS-rapport, Voorburg

BMDP (1985) Statistical Software Manual
University of California Press, Berkeley

Little, R.J.A. (1988) Missing-Data Adjustments in Large Surveys
Journal of Business and Economic Statistics (6-3) 287-296

Litle, R.J.A. and D.B. Rubin (1989) The Analyses of Social Science Data with Missing Values
Sociological Methods and Research (18-2,3) 292-326

Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys
John Wiley and Sons, New York

```
OPTIONS PAGESIZE=55 NONUMBER NODATE;
DATA EEN;
TITLE'GROWTH DATA FOR 11 GIRLS AND 16 BOYS';
INPUT Y SEX IDENT AGE;
CARDS;
21 0 1 8
21 0 2 8
20.5 0 3 8
23.5 0 4 8
21.5 0 5 8
20 0 6 8
21.5 0 7 8
23 0 8 8
20 0 9 8
16.5 0 10 8
24.5 0 11 8
20 0 1 10
21.5 0 2 10
. 0 3 10
24.5 0 4 10
23 0 5 10
. 0 6 10
22.5 0 7 10
23 0 8 10
. 0 9 10
. 0 10 10
25 0 11 10
21.5 0 1 12
24 0 2 12
24.5 0 3 12
25 0 4 12
22.5 0 5 12
21 0 6 12
23 0 7 12
23.5 0 8 12
22 0 9 12
19 0 10 12
28 0 11 12
23 0 1 14
25.5 0 2 14
26 0 3 14
26.5 0 4 14
23.5 0 5 14
22.5 0 6 14
25 0 7 14
24 0 8 14
21.5 0 9 14
19.5 0 10 14
28 0 11 14
26 1 1 8
21.5 1 2 8
23 1 3 8
25.5 1 4 8
```

20 1 5 8
24.5 1 6 8
22 1 7 8
24 1 8 8
23 1 9 8
27.5 1 10 8
23 1 11 8
21.5 1 12 8
17 1 13 8
22.5 1 14 8
23 1 15 8
22 1 16 8
25 1 1 10
. 1 2 10
22.5 1 3 10
27.5 1 4 10
. 1 5 10
25.5 1 6 10
22 1 7 10
21.5 1 8 10
20.5 1 9 10
28 1 10 10
23 1 11 10
. 1 12 10
. 1 13 10
25.5 1 14 10
24.5 1 15 10
. 1 16 10
29 1 1 12
23 1 2 12
24 1 3 12
26.5 1 4 12
22.5 1 5 12
27 1 6 12
24.5 1 7 12
24.5 1 8 12
31 1 9 12
31 1 10 12
23.5 1 11 12
24 1 12 12
26 1 13 12
25.5 1 14 12
26 1 15 12
23.5 1 16 12
31 1 1 14
26.5 1 2 14
27.5 1 3 14
27 1 4 14
26 1 5 14
28.5 1 6 14
26.5 1 7 14
25.5 1 8 14
26 1 9 14
31.5 1 10 14
25 1 11 14

```
28 1 12 14
29.5 1 13 14
26 1 14 14
30 1 15 14
. 1 16 14
PROC GLM;
      CLASS IDENT SEX;
      MODEL Y=IDENT*SEX AGE SEX*AGE;
      OUTPUT OUT=TWEE P=FIT R=RES COOKD=COOK STUDENT=F;
RUN;
PROC UNIVARIATE NORMAL PLOT DATA=TWEE;
      VAR RES;
PROC PLOT DATA=TWEE;
      PLOT RES*FIT/VREF=0;
PROC PRINT;
RUN;
```

GROWTH DATA FOR 11 GIRLS AND 16 BOYS

GENERAL LINEAR MODELS PROCEDURE
CLASS LEVEL INFORMATION

CLASS	LEVELS	VALUES
IDENT	16	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
SEX	2	0 1

NUMBER OF OBSERVATIONS IN DATA SET = 108

NOTE: DUE TO MISSING VALUES, ONLY 98 OBSERVATIONS CAN BE USED IN THIS ANALYSIS.

GROWTH DATA FOR 11 GIRLS AND 16 BOYS

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F
MODEL	28	716.9926583	25.6068807	12.22	0.0001
ERROR	69	144.5583621	2.0950487		
CORRECTED TOTAL	97	861.5510204			

R-SQUARE	C.V.	ROOT MSE	Y MEAN
0.832211	5.987673	1.447428	24.1734694

SOURCE	DF	TYPE I SS	MEAN SQUARE	F VALUE	PR > F
IDENT*SEX	26	471.2593537	18.1253598	8.65	0.0001
AGE	1	234.0927835	234.0927835	111.74	0.0001
AGE*SEX	1	11.6405211	11.6405211	5.56	0.0213

SOURCE	DF	TYPE III SS	MEAN SQUARE	F VALUE	PR > F
IDENT*SEX	26	352.6846478	13.5647941	6.47	0.0001
AGE	1	210.1557702	210.1557702	100.31	0.0001
AGE*SEX	1	11.6405211	11.6405211	5.56	0.0213

GROWTH DATA FOR 11 GIRLS AND 16 BOYS

UNIVARIATE PROCEDURE

VARIABLE=RES

MOMENTS

N	98	SUM WGTS	98
MEAN	0	SUM	0
STD DEV	1.220775	VARIANCE	1.490292
SKEWNESS	0.153562	KURTOSIS	4.577216
USS	144.5584	CSS	144.5584
CV	.	STD MEAN	0.123317
T:MEAN=0	0	PROB>|T|	1.0000
SGN RANK	-26.5	PROB>|S|	0.9258
NUM ^= 0	98		
W:NORMAL	0.948657	PROB<W	0.0022

QUANTILES(DEF=5)

100% MAX	5.075991	99%	5.075991
75% Q3	0.647026	95%	1.727974
50% MED	-0.01087	90%	1.272026
25% Q1	-0.70099	10%	-1.04901
0% MIN	-4.5033	5%	-1.57599
		1%	-4.5033
RANGE	9.579295		
Q3-Q1	1.348018		
MODE	-0.36957		

EXTREMES

LOWEST	OBS	HIGHEST	OBS
-4.5033(	57)	1.727974(	107)
-3.82599(	69)	1.772026(	55)
-2.02203(	96)	2.522026(	52)
-1.95099(	61)	3.202643(	105)
-1.57599(	68)	5.075991(	85)

MISSING VALUE	.
COUNT	10
% COUNT/NOBS	9.26

GROWTH DATA FOR 11 GIRLS AND 16 BOYS

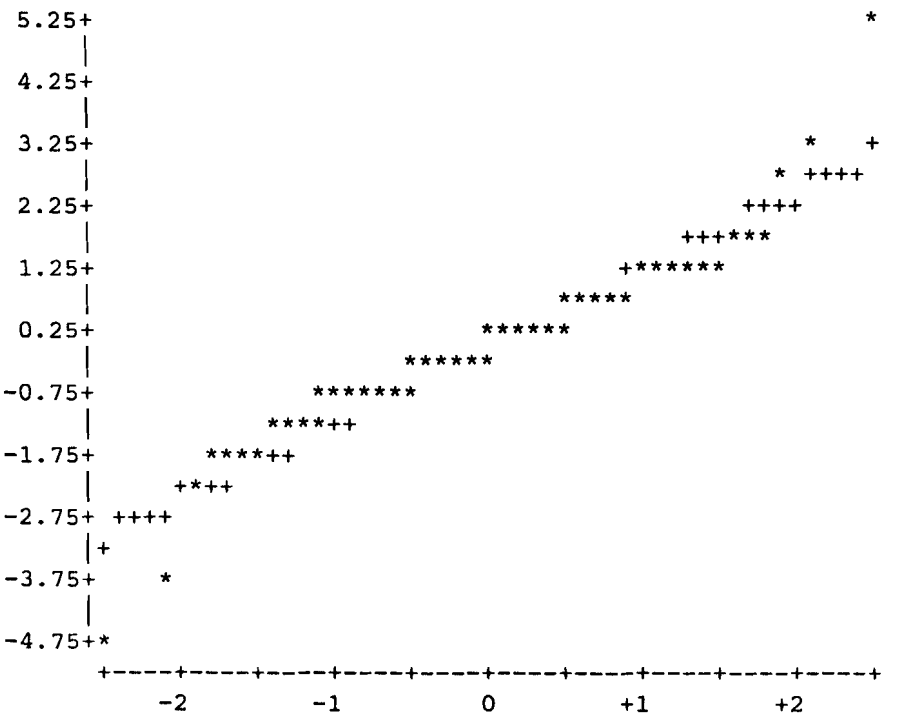
UNIVARIATE PROCEDURE

VARIABLE=RES

STEM	LEAF	#	BOXPLOT
5	1	1	*
4			
4			
3			
3	2	1	0
2	5	1	|
2			|
1	778	3	|
1	00011113344	11	|
0	55555555566778999	17	+------+
0	00111111233444	14	| + |
-0	44444332222221000	17	*------*
-0	99999988776665555	17	+------+
-1	320000000	9	|
-1	655	3	|
-2	00	2	|
-2			
-3			
-3	8	1	0
-4			
-4	5	1	0

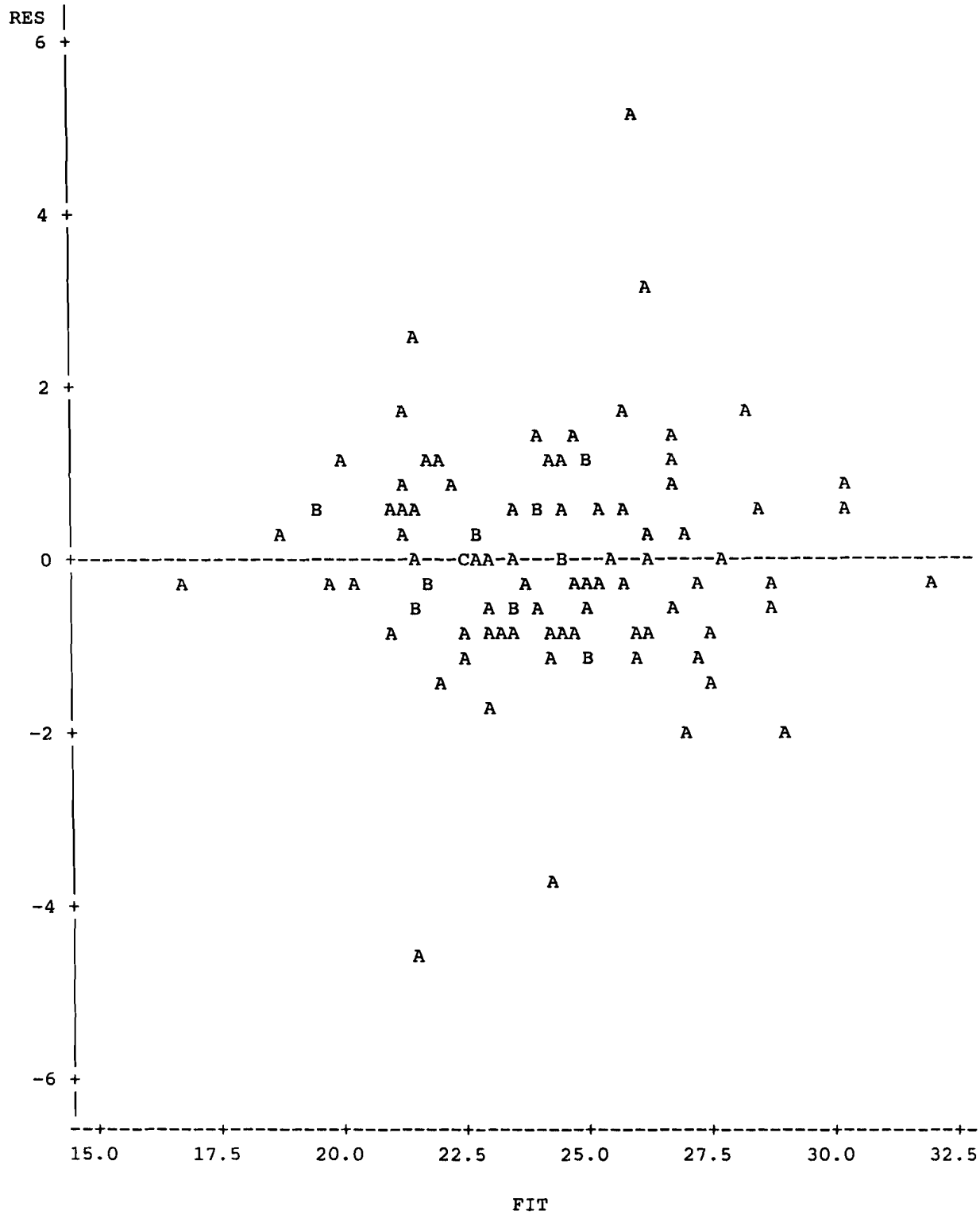
----+----+----+----+

NORMAL PROBABILITY PLOT



GROWTH DATA FOR 11 GIRLS AND 16 BOYS

PLOT OF RES*FIT. LEGEND: A = 1 OBS, B = 2 OBS, ETC.



NOTE: 10 OBS HAD MISSING VALUES.

GROWTH DATA FOR 11 GIRLS AND 16 BOYS

OBS	Y	SEX	IDENT	AGE	FIT	RES	COOK	F
1	21.0	0	1	8	19.8913	1.10870	0.011780	0.91028
2	21.0	0	2	8	21.5163	-0.51630	0.002555	-0.42391
3	20.5	0	3	8	22.0181	-1.51812	0.038634	-1.33753
4	23.5	0	4	8	23.3913	0.10870	0.000113	0.08924
5	21.5	0	5	8	21.1413	0.35870	0.001233	0.29450
6	20.0	0	6	8	19.5181	0.48188	0.003893	0.42456
7	21.5	0	7	8	21.5163	-0.01630	0.000003	-0.01339
8	23.0	0	8	8	21.8913	1.10870	0.011780	0.91028
9	20.0	0	9	8	19.5181	0.48188	0.003893	0.42456
10	16.5	0	10	8	16.6848	-0.18478	0.000572	-0.16280
11	24.5	0	11	8	24.8913	-0.39130	0.001467	-0.32128
12	20.0	0	1	10	20.8804	-0.88043	0.005849	-0.70457
13	21.5	0	2	10	22.5054	-1.00543	0.007627	-0.80460
14	.	0	3	10	23.0072	.	.	.
15	24.5	0	4	10	24.3804	0.11957	0.000108	0.09568
16	23.0	0	5	10	22.1304	0.86957	0.005705	0.69587
17	.	0	6	10	20.5072	.	.	.
18	22.5	0	7	10	22.5054	-0.00543	0.000000	-0.00435
19	23.0	0	8	10	22.8804	0.11957	0.000108	0.09568
20	.	0	9	10	20.5072	.	.	.
21	.	0	10	10	17.6739	.	.	.
22	25.0	0	11	10	25.8804	-0.88043	0.005849	-0.70457
23	21.5	0	1	12	21.8696	-0.36957	0.001030	-0.29574
24	24.0	0	2	12	23.4946	0.50543	0.001927	0.40447
25	24.5	0	3	12	23.9964	0.50362	0.003170	0.42681
26	25.0	0	4	12	25.3696	-0.36957	0.001030	-0.29574
27	22.5	0	5	12	23.1196	-0.61957	0.002896	-0.49581
28	21.0	0	6	12	21.4964	-0.49638	0.003080	-0.42066
29	23.0	0	7	12	23.4946	-0.49457	0.001845	-0.39578
30	23.5	0	8	12	23.8696	-0.36957	0.001030	-0.29574
31	22.0	0	9	12	21.4964	0.50362	0.003170	0.42681
32	19.0	0	10	12	18.6630	0.33696	0.001419	0.28556
33	28.0	0	11	12	26.8696	1.13043	0.009642	0.90463
34	23.0	0	1	14	22.8587	0.14130	0.000191	0.11602
35	25.5	0	2	14	24.4837	1.01630	0.009899	0.83442
36	26.0	0	3	14	24.9855	1.01449	0.015466	0.88057
37	26.5	0	4	14	26.3587	0.14130	0.000191	0.11602
38	23.5	0	5	14	24.1087	-0.60870	0.003551	-0.49976
39	22.5	0	6	14	22.4855	0.01449	0.000003	0.01258
40	25.0	0	7	14	24.4837	0.51630	0.002555	0.42391
41	24.0	0	8	14	24.8587	-0.85870	0.007066	-0.70502
42	21.5	0	9	14	22.4855	-0.98551	0.014595	-0.85541
43	19.5	0	10	14	19.6522	-0.15217	0.000348	-0.13209
44	28.0	0	11	14	27.8587	0.14130	0.000191	0.11602
45	26.0	1	1	8	25.3530	0.64703	0.003715	0.52672
46	21.5	1	2	8	21.0033	0.49670	0.003786	0.43235
47	23.0	1	3	8	21.8530	1.14703	0.011677	0.93375
48	25.5	1	4	8	24.2280	1.27203	0.014360	1.03551
49	20.0	1	5	8	20.1700	-0.16997	0.000443	-0.14795
50	24.5	1	6	8	23.9780	0.52203	0.002419	0.42496
51	22.0	1	7	8	21.3530	0.64703	0.003715	0.52672

GROWTH DATA FOR 11 GIRLS AND 16 BOYS

OBS	Y	SEX	IDENT	AGE	FIT	RES	COOK	F
52	24.0	1	8	8	21.4780	2.52203	0.05645	2.05308
53	23.0	1	9	8	22.7280	0.27203	0.00066	0.22145
54	27.5	1	10	8	27.1030	0.39703	0.00140	0.32320
55	23.0	1	11	8	21.2280	1.77203	0.02787	1.44254
56	21.5	1	12	8	21.8366	-0.33664	0.00174	-0.29303
57	17.0	1	13	8	21.5033	-4.50330	0.31125	-3.91994
58	22.5	1	14	8	22.4780	0.02203	0.00000	0.01793
59	23.0	1	15	8	23.4780	-0.47797	0.00203	-0.38910
60	22.0	1	16	8	21.1520	0.84802	0.02564	0.83973
61	25.0	1	1	10	26.9510	-1.95099	0.02846	-1.55986
62	.	1	2	10	22.6013	.	.	.
63	22.5	1	3	10	23.4510	-0.95099	0.00676	-0.76034
64	27.5	1	4	10	25.8260	1.67401	0.02095	1.33841
65	.	1	5	10	21.7680	.	.	.
66	25.5	1	6	10	25.5760	-0.07599	0.00004	-0.06076
67	22.0	1	7	10	22.9510	-0.95099	0.00676	-0.76034
68	21.5	1	8	10	23.0760	-1.57599	0.01857	-1.26004
69	20.5	1	9	10	24.3260	-3.82599	0.10946	-3.05897
70	28.0	1	10	10	28.7010	-0.70099	0.00367	-0.56046
71	23.0	1	11	10	22.8260	0.17401	0.00023	0.13912
72	.	1	12	10	23.4347	.	.	.
73	.	1	13	10	23.1013	.	.	.
74	25.5	1	14	10	24.0760	1.42401	0.01516	1.13853
75	24.5	1	15	10	25.0760	-0.57599	0.00248	-0.46052
76	.	1	16	10	22.7500	.	.	.
77	29.0	1	1	12	28.5490	0.45099	0.00152	0.36058
78	23.0	1	2	12	24.1993	-1.19934	0.01791	-1.01594
79	24.0	1	3	12	25.0490	-1.04901	0.00823	-0.83871
80	26.5	1	4	12	27.4240	-0.92401	0.00638	-0.73877
81	22.5	1	5	12	23.3660	-0.86601	0.00934	-0.73358
82	27.0	1	6	12	27.1740	-0.17401	0.00023	-0.13912
83	24.5	1	7	12	24.5490	-0.04901	0.00002	-0.03918
84	24.5	1	8	12	24.6740	-0.17401	0.00023	-0.13912
85	31.0	1	9	12	25.9240	5.07599	0.19267	4.05837
86	31.0	1	10	12	30.2990	0.70099	0.00367	0.56046
87	23.5	1	11	12	24.4240	-0.92401	0.00638	-0.73877
88	24.0	1	12	12	25.0327	-1.03267	0.01328	-0.87476
89	26.0	1	13	12	24.6993	1.30066	0.02107	1.10177
90	25.5	1	14	12	25.6740	-0.17401	0.00023	-0.13912
91	26.0	1	15	12	26.6740	-0.67401	0.00340	-0.53889
92	23.5	1	16	12	24.3480	-0.84802	0.02564	-0.83973
93	31.0	1	1	14	30.1470	0.85297	0.00646	0.69437
94	26.5	1	2	14	25.7974	0.70264	0.00701	0.60531
95	27.5	1	3	14	26.6470	0.85297	0.00646	0.69437
96	27.0	1	4	14	29.0220	-2.02203	0.03629	-1.64605
97	26.0	1	5	14	24.9640	1.03598	0.01524	0.89246
98	28.5	1	6	14	28.7720	-0.27203	0.00066	-0.22145
99	26.5	1	7	14	26.1470	0.35297	0.00111	0.28734
100	25.5	1	8	14	26.2720	-0.77203	0.00529	-0.62848
101	26.0	1	9	14	27.5220	-1.52203	0.02056	-1.23902
102	31.5	1	10	14	31.8970	-0.39703	0.00140	-0.32320

GROWTH DATA FOR 11 GIRLS AND 16 BOYS

OBS	Y	SEX	IDENT	AGE	FIT	RES	COOK	F
103	25.0	1	11	14	26.0220	-1.02203	0.00927	-0.83199
104	28.0	1	12	14	26.6307	1.36931	0.02662	1.17962
105	29.5	1	13	14	26.2974	3.20264	0.14562	2.75898
106	26.0	1	14	14	27.2720	-1.27203	0.01436	-1.03551
107	30.0	1	15	14	28.2720	1.72797	0.02650	1.40668
108	.	1	16	14	25.9460	.	.	.