KM 39 (1992)
pg.113-159

Missing data handling in verschillende pakketten

Inhoud

Probleemstelling en GLIM

Jan B. Dijkstra

Samenvatting

Deze notitie dient als inleiding voor zes andere bijdragen over missing data handling in verschillende pakketten. Uitgebreide aandacht zal worden besteed aan een probleem met ontbrekende waarnemingen in de response-variabele van een lineair model. De mogelijkheden van een analyse met het pakket GLIM zal in dit geval worden behandeld. Het probleem blijkt gebruikersvriendelijk en inhoudelijk verdedigbaar oplosbaar. Aanzienlijk slechter scoort GLIM zodra de ontbrekende waarnemingen voorkomen in de predictoren van een lineair model.

1. De aanleiding

Op 24 januari 1991 hield de Sectie Statistische Programmatuur van de Vereniging voor Statistiek 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. 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 ontbrekend behandelde waarneming wel degelijk kennen) is een wat ongebruikelijk kwaliteiscriterium haalbaar: de kwadratensom voor de tien missing data (KSMD).

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)

2. 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. Bij gebrek aan voorinformatie is dit zomaar een probeersel. Een grafische representatie van deze aanpassing bestaat uit twee lijnen. De verticale 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 aanpassing gedaan met de defaultinstelling 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 (ten opzichte van wat we straks zullen zien) in deze aanpassing met slechts 3 parameters.

3. 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, KSMD=34.04) 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, een wat overdadig model). 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 determinatiecoëfficient is namelijk bij (7) 0.8187 en bij (6) 0.8364. Bovendien bevat (6) geen niet-significante predictoren.

8. De winnaar is een model met 27 intercepts en 2 hellingen (29 parameters). KSMD=10.66 en dat is gunstiger dan het resultaat bij (6). De determinatiecoëfficient 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 de volgende hoofdstukken verwezen.

4. 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 enig eenvoudig toepasbare mogelijkheid met GLIM is hierbij Listwise Deletion. Dit houdt in dat een observatie in zijn geheel genegeerd wordt zodra voor tenminste één 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 covariantiematrix van de predictoren zijn. Hierbij gaat Pairwise Deletion aanzienlijk zuiniger om met de beschikbare informatie dan Listwise Deletion. Bij de eerstgenoemde 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 datamatrix 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.

5. 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 datamatrix zou dus 220 getallen bevatten. Stel nu dat voor iedere observatie in een willekeurige predictor de waarde ontbreekt. De datamatrix 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.

6. Numerieke problemen

GLIM kent alleen Listwise Deletion (een methode die veronderstelt dat de ontbrekende waarnemingen willekeurig geplaatst zijn) zodat na het opschonen van de datamatrix zeer beproefde algoritmen kunnen worden toegepast op de resterende rijen. Enkele van de in de volgende paragrafen te behandelen pakketten bieden Pairwise Deletion (of varianten hierop). Hoewel elk element van de hierdoor berekende covariantiematrix correct geschat wordt,

is deze matrix niet noodzakelijk semi positief definiet. In tegenstelling tot bij de vorige 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).

7. Literatuur

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
Biometrica (51) 313-326

De behandeling van ontbrekende gegevens in SPSS

Lex Jansen[1]

1. Inleiding

In deze bijdrage wordt beschreven hoe SPSS omgaat met het probleem van ontbrekende waarnemingen. SPSS biedt verschillende mogelijkheden voor

[1]Centraal Bureau voor de Statistiek, Hoofdafdeling Statistische Methoden, Postbus 4481, 6401 CZ Heerlen. De in dit artikel weergegeven opvattingen zijn die van de auteur en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek

de behandeling van ontbrekende waarnemingen. Er wordt beschreven hoe in SPSS ontbrekende gegevens geïdentificeerd kunnen worden. Aan de hand van een voorbeeld zal getoond worden hoe SPSS bij regressie-analyse met ontbrekende gegevens omgaat.
SPSS is een algemeen statistisch pakket. Met dit pakket kan men gegevens manipuleren en beheren, rapporten opstellen en statistische analyses uitvoeren. Het pakket kan zowel op een mainframe (SPSS^{X}) als op PC (SPSS/PC+) geïnstalleerd worden. De analyses in deze nota zijn uitgevoerd op een PC met SPSS/PC+ versie 3.1.

2. Ontbrekende waarnemingen

SPSS maakt onderscheid tussen twee verschillende soorten ontbrekende waarnemingen, namelijk *user-missing values* en *system-missing values*. Soms is er voor een case geen informatie beschikbaar voor een bepaalde variabele. In zo'n geval kan de gebruiker een speciale code toepassen om aan te geven dat de waarde ontbreekt. Als bijvoorbeeld de leeftijd van een individu niet bekend is, kan dat worden aangegeven door op de desbetreffende plaats in de datafile bijvoorbeeld de code –1 te plaatsen. Met behulp van het MISSING VALUE commando wordt de waarde aangegeven die ontbrekende informatie identificeert. Bijvoorbeeld:

MISSING VALUE leeftijd (–1)

Dit is een user-missing value, welke door de gebruiker zelf wordt geïdentificeerd. System-missing values daarentegen worden door SPSS zelf toegekend. Ze worden in de uitvoer met een punt weergegeven. Er kunnen verschillende redenen zijn waarom SPSS een system-missing value toekent. Het kan bijvoorbeeld voorkomen dat een waarde in de datafile niet overeenkomt met het formaat dat gespecificeerd is voor de desbetreffende variabele. Ook als een variabele het resultaat is van een berekening waarbij één of meerdere variabelen met ontbrekende waarnemingen worden gebruikt, krijgt de desbetreffende uitkomst de code voor een ontbrekende waarneming. In de handleiding van SPSS/PC+ (Zie Norusis, 1988, blz. B-32) worden een aantal uitzonderingen op deze regel genoemd:

```
0*missing   =  0
0/missing   =  0
missing**0  =  1
0**missing  =  0
```

Deze uitzonderingen blijken in de praktijk niet altijd te kloppen. SPSS/PC+ kent in alle 4 de gevallen de code voor een ontbrekende waarneming toe aan de uitkomst van de expressie. Deze fout is ook in latere versies van de handleiding niet hersteld.
In de (SPSS^X)-versies 2.2 en 3.0 krijgen de expressies "`0*missing`" en "`0/missing`" echter wél de waarde 0 toegekend. De expressies "`missing**0`" en "`0**missing`" krijgen de code voor een ontbrekende waarneming toegekend.
Niet alle versies van SPSS gaan dus op dezelfde manier met deze expressies om. Een andere reden voor het toekennen van een system-missing value is dat een variabele niet is gedefinieerd. Dit kan bijvoorbeeld gebeuren bij deling door 0 of als het argument van de logaritme-functie negatief is.

3. Behandeling van ontbrekende waarnemingen

Bij meervoudige regressie kent SPSS verschillende manieren waarop er met ontbrekende waarnemingen wordt omgegaan.

Bij *listwise deletion* wordt iedere case die een ontbrekende waarneming op een variabele heeft, verwijderd. Dit is de default methode bij SPSS. Als men een grote fractie cases heeft verwijderd zal de variantie van de schatter vaak naar evenredigheid toegenomen zijn. Deze methode leunt zwaar op de veronderstelling dat de volledige cases opgevat kunnen worden als een aselecte steekproef uit de verzameling van alle cases.

Bij *pairwise deletion* wordt de berekening van de covariantie van twee variabelen gebaseerd op alle cases die volledige informatie voor die twee variabelen hebben. De covariantiematrix hoeft nu niet meer positief-definiet te zijn. In de handleiding van SPSS (Norusis, 1988, blz. B-218) wordt de gebruiker er terecht op gewezen dat hier enige voorzichtigheid geboden is. In de uitvoer wordt de gebruiker gewaarschuwd als er sprake is van een correlatiematrix die niet positief-definiet is. SPSS meldt dan: "The following

variables have impossible tolerances. Original correlation matrix may not be positive definite. Pairwise deletion may be inappropriate for these variables."

Een ander probleem met deze matrices is inconsistentie. Als bijvoorbeeld leeftijd en gewicht en leeftijd en lengte een positieve correlatie hebben, dan is het in dezelfde populatie onwaarschijnlijk dat gewicht en lengte een sterk negatieve correlatie hebben. Als echter de berekening van deze drie correlatiecoëfficiënten niet steeds op dezelfde cases gebaseerd is, kunnen de uitkomsten toch een dergelijke inconsistentie te zien geven. Ook levert pairwise deletion soms zelfs schatters met een grotere variantie dan listwise deletion, zie Verbeek (1979).

Bij *meansubstitution* worden alle cases gebruikt voor de berekeningen, waarbij het gemiddelde van een variabele ingevuld wordt voor de ontbrekende waarnemingen. Dit leidt systematisch tot te kleine varianties, covarianties en correlaties.

Een andere mogelijkheid bij SPSS is om via eigen berekeningen of met behulp van externe programmatuur 'redelijke' schattingen voor de ontbrekende waarnemingen te produceren. Met behulp van de *include-optie* kunnen deze schattingen dan bijvoorbeeld bij regressieanalyse worden gebruikt. Op deze manier wordt het mogelijk om te imputeren.

4. Een voorbeeld

Aan de hand van het reeds eerder beschreven voorbeeld van Little en Rubin (1987) zullen we nu laten zien hoe SPSS bij meervoudige regressie met ontbrekende waarnemingen omgaat. Voor ieder geslacht is de gemeten afstand op leeftijd 10 verwijderd voor die gevallen waar op leeftijd 8 een lage waarde in de waarneming optrad. We hebben hier dus te maken met een geval waarbij het ontbreken van de data met de waarde van een andere variabele samenhangt. Een eenvoudige manier om te onderzoeken of de waarnemingen volgens een willekeurig patroon ontbreken is om de waarnemingen in twee groepen te verdelen. De ene groep bevat de cases met ontbrekende waarnemingen op één of meerdere variabelen en de andere groep bevat de cases met volledige informatie. Met behulp van de SPSS-procedure T-TEST kan men nu de verdeling van de andere variabelen over de twee groepen onderzoeken.

We gaan nu een lineair model aan de data aanpassen. Het lineair model bevat de gemeten afstand als respons-variabele en de leeftijd en het individu als predictoren. Ieder individu heeft een eigen intercept en iedere sexe heeft een eigen helling. Er worden dus 29 parameters geschat. De volgende variabelen worden ingevoerd:

DIST	:	afstand van het centrum van het hersenaanhangsel tot de kaaksplijting
AGEM	:	voor een meisje de leeftijd en voor een jongen de waarde 0,
AGEJ	:	voor een jongen de leeftijd en voor een meisje de waarde 0,
Ij	:	een 0/1-variabele die de waarde 1 heeft voor individu j waarde 0 voor de overige individuen (j=1,...27).

Met de volgende commando's wordt het model aangepast met pairwise deletion:

```
MISSING VALUE DIST(-1).
IF (INDI EQ 1) I1=1.
.
.
.
IF (INDI EQ 27) I27=1.
RECODE i1 to i27 (SYSMIS=0).
REGRESSION VARIABLES= I1 TO I27, AGEM AGEJ DIST
/MISSING PAIRWISE
/DEPENDENT=DIST
/ENTER.
```

Enigszins bewerkt ziet de uitvoer er dan als volgt uit:

```
Pairwise Deletion of Missing Data
 Multiple R           .91043
 R Square             .82889
 Adjusted R Square    .75945
 Standard Error      1.46170
```

```
Analysis of Variance
                        DF  Sum of Squares  Mean Square
Regression              28       714.12779     25.50456
Residual                69       147.42323      2.13657

F =     11.93716    Signif F =  .0000

---------------------- Variables in the Equation -----------------------

Variable            B        SE B      Beta         T  Sig T
AGEJ              .71788     .08582   1.37271     8.365  .0000
I27             -2.66210    1.08555   -.16948    -2.452  .0167
.
.
I10             -6.16075    1.83463   -.39221    -3.358  .0013
AGEM              .64991     .10350   1.22474     6.279  .0000
(Constant)      17.89077    1.21587              14.714  .0000

Residuals Statistics:
             Min       Max      Mean   Std Dev     N
*PRED    16.9293   31.9325   24.1735    2.7133   108
*RESID   -5.0137   -5.1073    -.1910    1.3170    98
*ZPRED   -2.6698    2.8596     .0000    1.0000   108
*ZRESID  -3.4300    3.4941    -.1307     .9010    98
```

In de uitvoer zien we dat het model vrij goed past. In tabel 2 staan voor listwise deletion, pairwise deletion en meansubstitution voor het model de gemiddelde restkwadratensommen voor alle waarnemingen en de gemiddelde restkwadratensommen voor de ontbrekende waarnemingen.

Tabel 2. Gemiddelde restkwadratensommen

pairwise		listwise		meansubstitution	
model	missing	model	missing	model	missing
2.14	0.55	2.10	1.07	2.64	0.57

5. Conclusie

SPSS beschikt over verschillende methoden om een regressieanalyse uit te voeren in het geval dat een dataset ontbrekende waarnemingen bevat. De gebruiker dient zelf te onderzoeken wat in een concreet geval de beste methode is. Bij de interpretatie van het verkregen resultaat is voorzichtigheid geboden. SPSS vermeldt dan ook waarschuwend in de handleiding (Norusis, 1988, blz. B-218): "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."

6. Literatuur

Little, R.J.A. en D.B. Rubin, 1987,
Statistical analysis with missing data.
(John Wiley & Sons, New York)

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

Verbeek, A, 1979,
Ontbrekende scores bij het schatten van covariantiematrices.
Uit de Statistische Consultatie, VVS-Bulletin, jaargang 12,
no. 2, pp. 38-47.

Missing Data in SAS

Henk van der Knaap

1. Inleiding

SAS is een algemeen data-verwerkings en analyse-pakket. De bouwers ervan claimen dat het een 'strategic tool' is, handig in gebruik om uit ruwe data kant en klare informatie te verkrijgen.

Om dit doel te verwezenlijken heeft de gebruiker van SAS een groot scala van procedures ter beschikking. De te verwerken gegevens worden aan een procedure aangeboden in de vorm van een dataset, afkomstig uit een 'data-step'. Deze data-step is ook de programmeeromgeving waarin gegevens bewerkt kunnen worden. Verder bestaat ook de mogelijkheid om uitkomsten van een procedure door te spelen naar een volgende procedure. Voor databasemanagement heeft SAS, naast zijn eigen faciliteiten, ook voor goede communicatiemiddelen naar de vooraanstaande database pakketten gezorgd. Zo is het bijvoorbeeld niet nodig om de inhoud van een ORACLE database eerst met diverse vertaalslagen om te vormen tot een echte SAS dataset.
Dit hoofdstuk beschrijft hoe SAS omgaat met missing data in het algemeen, en met de probleemstelling van dit artikel in het bijzonder. De beschreven analyses zijn uitgevoerd op een IBM/PS2 model 80 met versie 6.04 van SAS.

2. Missing data

De standaard representatie van missing data in numerieke variabelen is de punt '.'. Om onderscheid te kunnen maken tussen verschillende oorzaken van het ontbreken van gegevens kan, door middel van een optie in een SAS programma, ook gebruik gemaakt worden van de onderscore '_' en de hoofdletters `A` t/m `Z`.
De interne ordening van een variabele met missing data is:
'_'<'.'< `A...Z` < getallen.
De waarde van een variabele is missing als bij het inlezen van de data de waarde ontbreekt, niet ingelezen of niet uitgerekend kan worden.
Bijvoorbeeld:
Er zijn 3 variabelen A, B en C en voor een observatie ontbreekt de waarde van B. Dan is de waarde van log(B) en de som van de 3 variabelen missing. SAS biedt in de vorm van functies mogelijkheden om een aantal kengetallen toch uit te rekenen over de variabelen met niet-ontbrekende waarden. Zo is in dit voorbeeld het resultaat van de functie SUM(A,B,C) voor de derde waarneming gelijk aan de som van A en C. Een en ander wordt in het volgende SAS programma weergegeven.

```
data example;
input a 1 b 3 c 5;
logb=log(b);
som=a+b+c;
```

```
sum=sum(a,b,c);
lines;
1 7 5
6 3 4
9   2
7 8 3
4 2 7
;
proc print;
run;
```

OBS	A	B	C	logB	SOM	SUM
1	1	7	5	1.94591	13	13
2	6	3	4	1.09861	13	13
3	9	.	2	.	.	11
4	7	8	3	2.07944	18	18
5	4	2	7	0.69315	13	13

3. Groeiprobleem

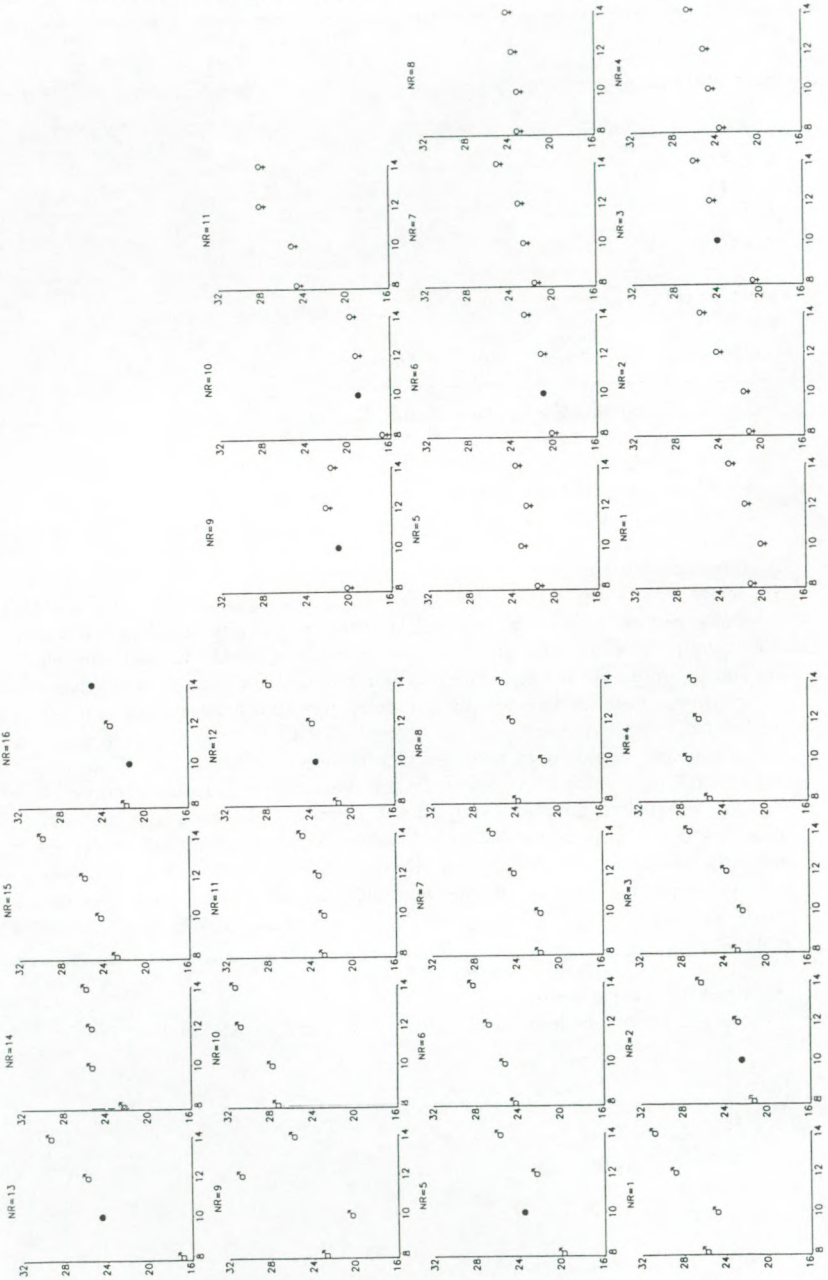
De eerste stap in een data-analyse is doorgaans een inspectie van de te analyseren gegevens. In het probleem dat Dijkstra stelt in zijn inleiding kan dat met behulp van een aantal grafieken gebeuren. De grafische uitvoer van een SAS-programma wordt weergegeven in figuur 1. Hierin zijn de te verwijderen gegevens weergegeven door een punt en de overige door het symbool voor het geslacht.

Bij het inlezen van de data kan men kiezen voor een dataset met 27 observaties en 6 variabelen (2 voor het design: geslacht en persoonsnummer, 4 analyse-veriabelen: groei op leeftijd 8, 10, 12 en 14) of een dataset met 108 observaties en 4 variabelen (3 voor het design: geslacht, persoonsnummer en leeftijd, 1 analyse-variabele: groei). De ene organisatie van de dataset kan overigens op een eenvoudige manier veranderd worden in de andere door gebruik te maken van PROC TRANSPOSE of een aantal manipulaties in een data-step.

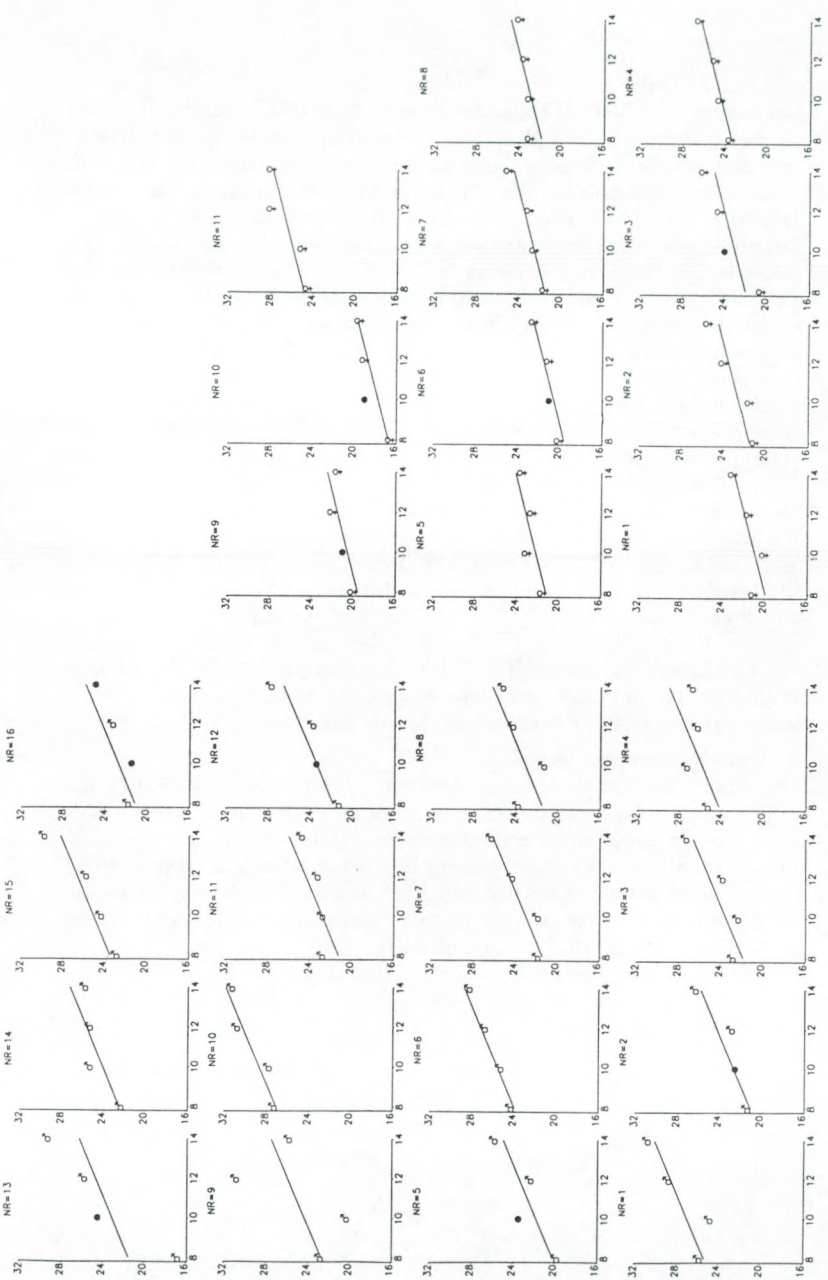
4. 4 analyse-variabelen:

Deze organisatie van de dataset ligt voor de hand als de analyse van het

Figuur 1: Overzicht van de groeicijfers.



Figuur 2: Groei per persoon als functie van de leeftijd.



probleem op multivariate wijze zal plaatsvinden.
Met behulp van PROC REG kan dan de groei op leeftijd 10 verklaard worden uit die op de andere leeftijden. Men verkrijgt op deze wijze inschattingen voor de ontbrekende gegevens met uitzondering voor jongen 16. Om voor deze observatie schattingen te verkrijgen is een extra stap nodig waarin eerst de groei op leeftijd 14 verklaard wordt uit die op leeftijd 8 en 12.
Een principale componenten analyse zou ook overwogen kunnen worden. Het nadeel van de klassieke benadering (m.b.v. PROC PRINCOMP) is dan dat alle observaties met missing data verwijderd worden. Bovendien kunnen inschattingen voor de ontbrekende gegevens slechts verkregen worden na enig programmeerwerk. De nieuwe procedure PROC PRINQUAL biedt hier echter soelaas. Deze procedure is o.a. gebaseerd op het werk van de universiteit van Leiden (De Leeuw c.s) aan optimaal schalen. Een van de mogelijke toepassingen van deze procedure is juist om op iteratieve wijze schattingen te verkrijgen voor gaten in datasets. Voorwaarde is dan wel dat het aantal lege cellen relatief klein is en willekeurig gespreid over de dataset. Aan de laatste voorwaarde wordt hier zeker niet voldaan.
Het nadeel bij elk van de tot nu toe genoemde analyse-methodes is dat op geen enkele wijze een model waarin de invloed van het geslacht en de leeftijd tegelijkertijd onderzocht kan worden. Dit zou wel in beperkte mate met een discriminant-analyse (PROC DISCRIM) of een multivariate variantie-analyse (PROC GLM) kunnen. In feite is het gestelde probleem een zogenaamd REPEATED MEASUREMENTS probleem en de REPEATED optie in PROC GLM ligt dus voor de hand. Echter, beide laatst genoemde procedures verwijderen ook de observaties met missing data.

5. 1 analyse-variabele:
Deze organisatie van de dataset ligt voor de hand als men met PROC GLM een linear model wilt onderzoeken waarin de groei gerelateerd wordt aan de leeftijd en waarbij rekening gehouden wordt met de persoon en het geslacht. In deze situatie wordt alle niet-missing informatie gebruikt en men verkrijgt inschattingen voor de lege cellen. Op deze wijze kunnen alternatieve modellen met elkaar vergeleken worden met als criteria de Residual Mean Squares zoals Dijkstra die beschrijft in zijn inleiding.
Onderstaande tabel geeft enige van deze resultaten.

Termen	Npar	Res. Mean Sq. Model	Missing
27 intercepts, 1 helling	28	2.23	1.08
27 intercepts, helling per sex	29	2.10	1.07 ***
27 intercepts, 27 hellingen	54	1.78	2.04
27 intercepts, helling per sex en kromming per sex	31	2.09	1.44
Na worteltransformatie op leeftijd:			
27 intercepts, helling per sex	29	2.14	0.96

Uit deze tabel blijkt dat een model met 27 intercepts en 27 hellingen de beste oplossing geeft v.w.b. de RMS(model) maar is voor praktisch gebruik niet aan te bevelen. Dit geldt wel voor het model 27 intercepts en 2 hellingen. Bovendien scoort dit model ook het best op het criterium RMS (missing) maar dat is in het dagelijks leven van een statistisch consulent niet een reëel criterium. Als echter naar het 'Press' kengetal gekeken wordt om de residuen te inspecteren (in de probleemstelling van dit artikel heel toepasselijk), dan zou ook voor dit model gekozen worden. De resultaten van deze analyse worden nog eens grafisch weergegeven in figuur 2. Een betere oplossing gebaseerd op RMS(missing) werd nog wel gevonden na worteltransformatie van de variabele 'leeftijd'. Deze transformatie werd gekozen na een niet-lineaire regressie analyse waarin het model de exponent van 'leeftijd' als parameter bevatte.
Er zullen echter in de praktijk gegronde redenen moeten bestaan om een dergelijk model te hanteren.

6. Opmerkingen

In de andere door Dijkstra genoemde situaties (missing data in predictor variabelen en het ontaarde geval met in bijna alle observaties een lege cel) zou met PROC PRINQUAL een analyse uitgevoerd kunnen worden. De eerder genoemde kanttekeningen zijn daarbij ook onverkort van kracht.

7. Conclusie

In het pakket SAS verwijderen alle voor dit onderwerp relevante procedures, met uitzondering van PROC PRINQUAL, de observaties met missing data. Toch kan met PROC GLM en een juiste modelkeus in een groot aantal gevallen een optimaal gebruik van alle beschikbare informatie gebruik gemaakt worden.

missing data handling in STATA

Ger Snijkers [2]

We zullen nu eens bekijken hoe STATA met ontbrekende gegevens omgaat. We bespreken de mogelijkheden van imputeren met regressie, interpolatie en wegen. Maar eerst zullen we kort een aantal kenmerken van STATA noemen. Voor een uitvoerige bespreking wordt verwezen naar Snijkers (1991).

1. Enkele kenmerken van STATA

STATA werkt vector-georiënteerd en commando-gestuurd. Het is echter ook mogelijk om menu-gestuurd te werken. Alle commando's worden interactief ondersteund met uitgebreide help-faciliteiten, die het gebruik van de (duidelijke) manual; (CRC, 1990) bijna overbodig maken.

STATA kent een aantal variabele typen. Default worden variabelen als *float* ingelezen. Een overzicht van de variabele-typen is gegeven in tabel 1, waarbij het aantal bytes is aangegeven, alsmede het waardenbereik van de typen en de waarde die aan de ontbrekende score wordt toegekend.

Tabel 1. Variabele-typen in STATA.

Variabele type	aantal bytes	minimum	maximum	missing
byte	1	-127	126	127
int	2	-32.767	32.766	32.767
long	4	-2.147.483.648	2.147.483.646	2.147.483.647
float (default)	4	$\pm 10^{-37}$	$\pm 10^{37}$	2^{128}
double	8	$\pm 10^{-99}$	$\pm 10^{99}$	2^{333}
string *)	2-80	2	80	""

*) STATA kent alleen even string-lengtes (str2, str4, ...); oneven strings worden met 1 opgehoogd.

[2] Centraal Bureau voor de Statistiek, Hoofdafdeling Statistische Methoden, Postbus 959, 2270 AZ Voorburg. De hier weergegeven opvattingen zijn die van de auteur en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek.

2. STATA en ontbrekende scores

Ontbrekende scores worden in STATA gedefinieerd als een punt (.). In een data bestand dienen ontbrekende scores voor numerieke variabelen als zodanig te zijn aangegeven (in free format). Voor string-variabelen wordt de *null string* ("") gebruikt.

De numerieke waarde die aan de ontbrekende score wordt toegekend is groter dan alle niet-ontbrekende scores (zoals we gezien hebben in tabel 1). Alleen voor string-variabelen geldt dat de ontbrekende score (de *null string*) aan alle andere karakters vooafgaat.

Elke rekenkundige operatie (+, −, *, /, ^) en elke functie (abs, exp, log, sqrt, mod, atan, cos, sin) die wordt toegepast op een ontbrekende score resulteert in een ontbrekende score. Delen door nul levert eveneens ontbrekende scores.

In de uitvoer van statistische analyses worden ontbrekende scores default niet meegenomen. Bijvoorbeeld in *summarize* en *tabulate*.

In multivariate statistische analyses is alleen *listwise deletion* mogelijk. Dit staat nergens expliciet in de manual vermeld. Voor bijvoorbeeld het berekenen van correlaties en in regressie-analyse worden dus alleen de complete records van de gekozen set variabelen gebruikt.

3. Missing data handling en gegevens-manipulatie

We zullen nu eens bekijken hoe (ontbrekende) gegevens met STATA gemanipuleerd kunnen worden.

- Imputeren met regressie-analyse

Om te beginnen moeten we natuurlijk eerst de gegevens inlezen. Vervolgens worden allerlei variabelen aangemaakt die nodig zijn bij de analyse. Dit staat in figuur 3. Nadat de gegevens zijn ingelezen van de file groei2.raw met *infile* ("lft" staat voor leeftijd, "groei" heeft betrekking op de groeicijfers, en "wgroei" bevat de bekende ontbrekende cijfers) wordt een variabele "id" gemaakt met volgnummers (met het commando *generate*, afgekort tot *gen*). Daarbij maken we gebruik van de systeem-vector *_n*, die volgnummers bevat. Daarna wordt de variabele "sex" gemaakt: een dummy met de score 0 als het volgnummer kleiner is dan 12 (meisjes), en de score 1 als het volgnummer groter dan of gelijk is aan 12 (jongens). Vervolgens worden op gelijke wijze twee leeftijds- en identificatievariabelen gemaakt voor beide sexen (==

betekent "gelijk aan"). Met het *replace*-commando kunnen in een variabele (die al bestaat) scores worden gewijzigd. De variabelen p1 tot en met p27 zijn dummy-variabelen, die in de regressie-analyse worden gebruikt om voor elke persoon een eigen intercept te kunnen schatten. Deze kunnen ook sneller worden gemaakt met het commando *tabulate id, gen(p)*. Tot slot worden de variabelen gesaved in de systeem file groei2.dta, die door vermelding van de parameter *replace* wordt overschreven als deze al bestaat. Figuur 3. Inlezen en manipuleren van de gegevens.

```
. infile lft groei wgroei using groei2.raw
. generate  id=_n
. gen       sex=(id>=12)
. gen       lftv=(sex==0)*lft
. replace   lftv=. if lftv==0
. gen       lftm=(sex==1)*lft
. replace   lftm=. if lftm==0
. gen       idv=(sex==0)*id
. gen       idm=(sex==1)*(id-11)
. gen       p1=(id==1)
  .....     .....
. gen       p27=(id==27)
. save groei2, replace
```

We gaan nu een regressie uitvoeren van groei op leeftijd, waarbij elk kind een eigen intercept krijgt en we voor de jongens en de meisjes een aparte helling bij leeftijd schatten. Hoe dit in STATA gebeurt staat in figuur 4: *reg* is de afkorting van het *regress*-commando, waarmee we hier "groei" regresseren op de variabelen "p1" tot en met "p27", "lftv" en "lftm". Verder geven we aan dat we geen algemene constante in het model willen opnemen met de parameter *noconst*. Omdat er gaten in de variabele "groei" zitten worden niet alle records in de regressie meegenomen: er vindt listwise deletion plaats; alleen de 98 complete van de in totaal 108 records worden gebruikt. Nadat de regressie is uitgevoerd wordt een variabele "pgroei" gemaakt, die de geschatte groeicijfers op grond van het model bevat. Dit doen we met het commando *predict*.

Figuur 4. Regressie: groei(i) = b_0(i) + b_1×lftv + b_2×lftm (i=1,2,...,27).

```
. reg groei p1-p27 lftv lftm, noconst (obs=98)
  Source |       SS       df       MS              Number of obs =      98
---------+------------------------------           F( 29,    69) =  954.37
   Model |  57983.9416     29  1999.44626           Prob > F      =  0.0000
Residual |  144.558362     69  2.09504873           R-square      =  0.9975
---------+------------------------------           Adj R-square  =  0.9965
   Total |  58128.50       98  593.147959          Root MSE      =  1.4474

------------------------------------------------------------------------------
Variable | Coefficient   Std. Error      t    Prob > |t|      Mean
---------+--------------------------------------------------------------------
   groei |                                                    24.17347
---------+--------------------------------------------------------------------
      p1 |   15.93478    1.305629    12.205    0.000       .0408163
     ....................
     p27 |   14.75991    1.318985    11.190    0.000       .0204082
    lftv |   .4945652    .0987904     5.006    0.000       4.530612
    lftm |   .7990088    .0831984     9.604    0.000       6.530612
------------------------------------------------------------------------------

. predict pgroei
```

De laatste stap is het imputeren van de geschatte groeicijfers. Dit is aangegeven in figuur 5. Alvorens we gaan imputeren maken we eerst een dummy "groeimis" waarin we met een 0 aangeven welke gegevens in het oorspronkelijke bestand ontbrekend waren (~= betekent "ongelijk aan"). Op die manier kunnen we altijd de imputatie ongedaan maken. Of we kunnen de variabele "groeimis" gebruiken als een variabele van gewichten in analyses, waardoor de geïmputeerde scores worden uitgesloten. Vervolgens worden dan de geschatte groeicijfers geïmputeerd in de variabele "groei". We willen nu ook nog zien wat het resultaat van deze operatie is. Daartoe maken we een grafiek met *graph*. Maar eerst maken we een hulpvariabele "pp", met een 1 voor die kinderen waarvoor gegevens ontbreken en een 0 voor alle andere. Het resultaat is te zien in figuur 6. De bekende en geïmputeerde groeicijfers zijn met een open rondje aangegeven, de bekende ontbrekende scores met een plusje. We zien dat we voor deze data op een eenvoudige manier een goede schatting hebben kunnen maken van de ontbrekende scores.

Figuur 5. Imputatie.

```
. sort id lft
. generate groeimis(groei~= .)
. replace groei=pgroei if groeimis==0
. generate pp= p3 + p6 + p9 + p10 + p13 + p16 + p23 + p24 + p27
. graph groei wgroei lft if pp==1, by(id)
```

- Lineaire interpolatie

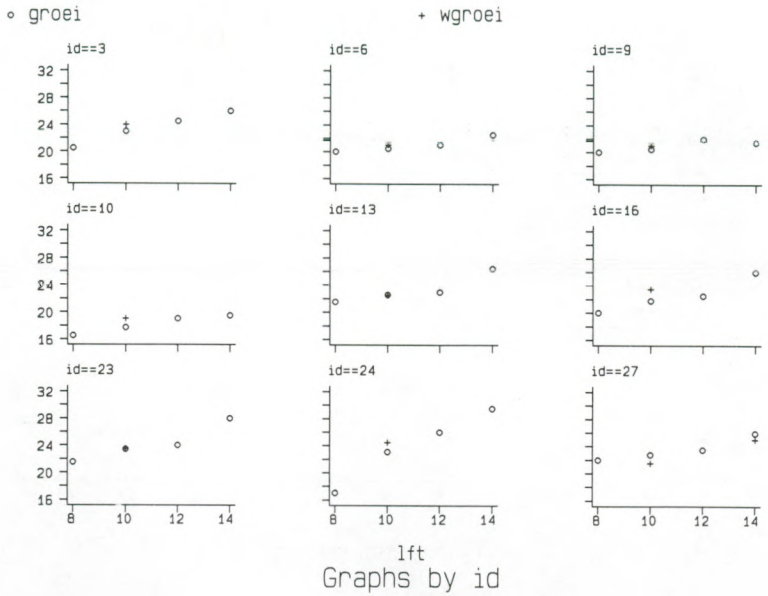
Aangezien de groeicijfers een lineair patroon laten zien, kunnen we ook overwegen de ontbrekende cijfers te imputeren met behulp van lineaire interpolatie.

Hoe dit in STATA in zijn werk gaat is te zien in figuur 7. Van de lijnvergelijking groei=a +b×lft bepalen we eerst de richtingscoëfficiënt b door het verschil tussen twee groeicijfers te delen door het verschil tussen de bijbehorende leeftijden. Daarbij maken we gebruik van de systeemvector _*n* met volgnummers. Vervolgens wordt de intercept a berekend. We hebben nu voor elk kind meerdere lijnfuncties gedefinieerd. We zijn echter alleen geïnteresseerd in die lijnfuncties waarmee we de ontbrekende groeicijfers kunnen berekenen. daartoe maken we eerst een variabele "ipgroei" met berekende groeicijfers voor alle kinderen bij 10 jaar. Voor kind 27 wordt nog een groeicijfer berekend op 14 jarige leeftijd. Vervolgens worden de berekende groeicijfers weer geïmputeerd op de lege plaatsen in de oorspronkelijke groei-vector. Het resultaat van deze bewerking is te zien in figuur 8.

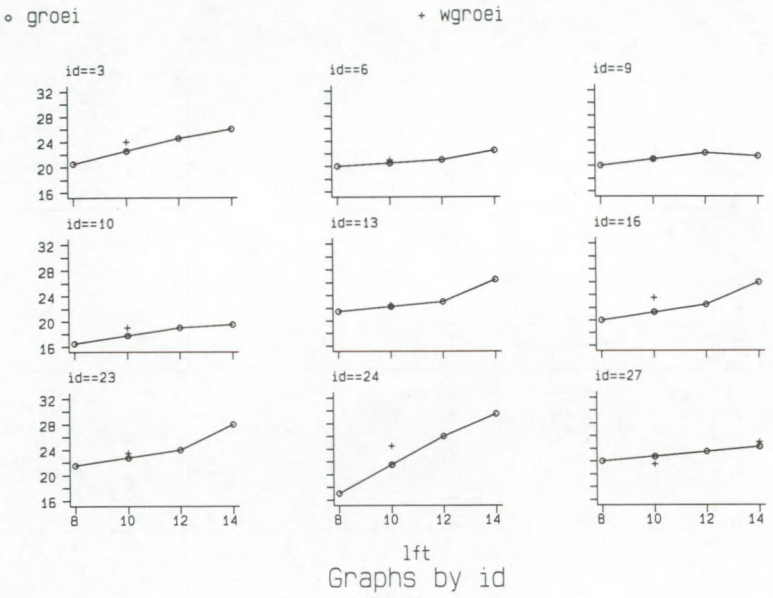
Figuur 7. Lineaire interpolatie.

```
. sort id lft
. gen b=(groei[_n+1]-groei[_n-1])/(lft[_n+1]-lft[_n-1])
. gen a=groei[_n-1] - (b*lft[_n-1])
. gen ipgroei=a + (b*lft) if (lft==10)
. replace ipgroei=a[_n-2] + (b[_n-2]*lft) if (id==27) & (lft==14)
. replace groei=ipgroei if groeimis==0
. graph groei wgroei lft if pp==1, by(id)
```

Figuur 6. Het resultaat van de imputatie met regressie.



Figuur 8. Het resultaat van de interpolatie.



- Wegen

Een laatste mogelijkheid die we bespreken om met ontbrekende scores om te gaan is wegen. We gaan dan op de lege plaatsen in het bestand geen berekende scores invullen zoals bij imputeren, maar we geven bekende gegevens een extra gewicht, waarmee de ontbrekende scores worden gecompenseerd.

Een eenvoudige manier om gewichten te bepalen voor ons voorbeeld is het representeren van de ontbrekende scores door bekende scores: we tellen sommige kinderen een aantal keren mee. We zoeken daartoe voor de kinderen met ontbrekende scores kinderen die een vergelijkbaar groeipatroon hebben. We beperken ons nu tot de groeicijfers op 10 jarige leeftijd. We maken eerst een gewichten-vector "w10" van enen. Dit is te zien in figuur 9. Vervolgens geven we gewichten aan de representanten met het commando *replace w10=<getal> in <kindnummer>*. Met behulp van deze gewichtenvector kunnen we onder andere een gewogen gemiddelde berekenen met het *summarize*-commando, we kunnen tabelleren (met *tabulatie*), of een gewogen regressie-analyse draaien. Door *=w10* op te nemen in de commando's worden de gewichten in de analyses meegenomen.

Figuur 9. Analyse met gewichten.

```
. gen w10=1
. replace w10=2 in 2
. replace w10=4 in 1
. replace w10=3 in 22
. replace w10=2 in 14
. replace w10=2 in 26
. replace w10=2 in 18
. sort sex
. by sex: sum g10 =w10
```

→ sex=0

Variable	Obs	Weight	Mean	Std. Dev	Min	Max
g10	7	11.0000	21.90909	1.896423	20	25

→ sex=1

Variable	Obs	Weight	Mean	Std. Dev	Min	Max
g10	11	16.0000	23.78125	2.153962	20.5	28

4. Conclusie

Ontbrekende gegevens kunnen in STATA eenvoudig worden gemanipuleerd. Ook kunnen statistische analyses worden uitgevoerd op variabelen met ontbrekende gegevens. Dit hebben we gezien aan de hand van het imputeren van ontbrekende gegevens op basis van lineaire regressie en lineaire interpolatie, en wegen (waarbij de gewichten ook op basis van een model kunnen worden berekend). Het is echter niet mogelijk een regressie-model te schatten voor een set van variabelen waarbij in elk record ten minste één gegeven ontbreekt (in de response- en de predictor-variabelen), aangezien STATA in multivariate analyses listwise deletion toepast.

5. Literatuur

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

Snijkers, G. J. M. E., 1991, Ontbrekende gegevens in STATA. CBS-rapport (Hoofdafdeling Statistische Methoden, Centraal Bureau voor de Statistiek, Voorburg).

Missing Data Handling in BMDP

Pierre Debets

BMDP is een algemeen statistisch pakket voor de analyse en verwerking van gegevens en is zowel beschikbaar op mainframe computers als op (IBM-compatible) PC's. Bij de start van BMDP in 1959 werd het programma in eerste instantie speciaal geschreven voor biomedisch onderzoek. De statistische programma's worden nu echter in allerlei onderzoeksgebieden gebruikt. BMDP heeft procedures voor het alledaagse statistische werk zoals frekwenties, beschrijvende statistiek, histogrammen, etc. En vele procedures voor geavanceerde statistiek zoals missing values analyse, factor-analyse, clusteranalyse, analyse van herhaalde metingen, tijdreeksen en survival analyse.

Het belangrijkste verschil met pakketten als SAS en SPSS, waar alle procedures in een programma zijn ondergebracht, is dat in BMDP voor elke procedure een apart programma bestaat. De versie van 1990 bestaat uit een reeks van 43 programma's en een Data Manager (file- en data-manipulaties), die compleet of in onderdelen verkrijgbaar zijn. Op de pc zijn deze programma's geintegreerd in een interaktief menu systeem met eigen editor, faciliteiten voor het bekijken van uitvoer, plotjes en invoer van data.

1. Missing values in BMDP.

De gebruiker heeft diverse mogelijkheden om aan te geven wat in zijn dataset de kodes voor missing data zijn. Deze worden in BMDP intern aangeduid met XMIS. Daarnaast kan men in BMDP eveneens voor de data een range opgeven. Waarden buiten deze ranges krijgen de interne kodes TOOBIG of TOOSMALL.

In het geval van transformaties worden de resultaten gelijk aan XMIS als in de berekening een van deze interne waarden voorkomt, tenzij de gebruiker expliciet aangeeft dat er toch met de oorspronkelijke waarden gerekend moet worden.

In de 'univariate' procedures zoals beschrijvende statistiek en frekwenties worden alleen geldige waarden voor een variabele gebruikt. In alle andere procedures worden alleen complete cases gebruikt in de analyse. In sommige procedures moet men hierbij bovendien nog aangeven dat die check op complete cases alleen op de analyse variabelen gedaan moet worden in plaats van op alle variabelen in een dataset.

2. De groei gegevens

Het bijschatten van de ontbrekende gegevens in de groei data volgens het in de inleiding besproken model kan in BMDP met 2 procedures: BMDP1R en BMDP5V. Onderstaand voor beide procedures de setup en een korte toelichting.

```
/input    title='model met BMDP1R'.
          variables=4.
          format=free.
/variable names are sexe, casenr, leeftijd, groei.
/transform   for i=1 to 27.
          % case|i=0.
          if (casenr=i) then case|i=1.%
```

BMDP1R is een programma voor multiple lineaire regressie voor meerdere groepen.

In de transform paragraaf worden 27 variabelen berekend voor de persoonspara-

```
        lftm=(sexe eq 1)*leeftijd.
        lftj=(sexe eq 2)*leeftijd.
/group  codes(sexe)=1,2.
        names(sexe)=meisjes, jongens.
/regres dep=groei.
        indep=lftm,lftj, case1 to case27.

        type=zero.
/print data.
/end
```

meters en de 2 variabelen voor de twee hellingscoefficienten.

Voor de ontbrekende gegevens worden op basis van het in de regres paragraaf gespecificeerde model de predicted values berekend en eventueel bewaard in een systemfile.

```
/input  title= 'model met BMDP5V'.
        variables=6.
        format=free.
        file='c:misssp.dat'.
/variable names are sexe, case, g8, g10, g12, g14.
/transform case=kase.
        if (sexe=1) then sexe=-1.
        if (sexe=2) then sexe=1.
/group  codes(sexe)=-1,1.
        names(sexe)=meisjes, jongens.
        codes(case)=1 to 27.
/design grouping=sexe,case.
        level=4.
        repeated=leeftijd.
        contrast(leeftijd)=lft.
        lft=8,10,12,14.
        dpname=groei.
        dpvar= g8, g10, g12, g14.
/model  groei='sexe + lft + sexe.lft
        + case'.
        no interc.
/print residuals.
```

BMDP5V is een programma voor een groot scala van experimentele designs o.a. met missing values.

Het model wordt beschouwd als een repeated measurement design met vier metingen van de groei per case.

De 'within-subject' factor `lft` heeft 4 waarden en wordt in het model opgenomen.

Voor de ontbrekende gegevens worden op basis van het in de model paragraaf gespecificeerde model de predicted values berekend

/end en optioneel bewaard.

3. Andere mogelijkheden in BMDP.

BMDP is een pakket waarin veel aandacht wordt geschonken aan het missing data probleem. Als men het probleem zou willen aanpakken zonder echt diep in allerlei statistische procedures te duiken dan biedt de transform paragraaf hiervoor al enkele mogelijkheden.
Men kan bijvoorbeeld voor een case de gemiddelde groei berekenen en die imputeren.

```
/TRANSFORM
    gemgroei=MEAN(g8, g10, g12, g14).
    FOR groei=g8, g10, g12, g14.
    % IF (groei EQ XMIS) THEN groei=gemgroei.  %
```

Er zijn ook 2 aparte funkties voor het vervangen van missing data door middel van interpolatie:

`REPL(g8,g10,g12,g14).`	imputeert in g10 i. h. g. van een missing value het gemiddelde van g8 en g12. Missings op g14 worden echter niet geschat.
`FILL(x1,y1,x2,y2,x3,y3,...)`	imputeert een paarsgewijze lineaire interpolatie voor de ontbrekende y waardes.

Veel meer mogelijkheden bieden de procedures 8D en AM; twee procedures speciaal bedoeld voor missing data.

BMDP8D berekent de correlaties/covarianties tussen variabelen op 4 verschillende methoden:

`COMPLETE`	gebruikt alleen volledige cases: d.w.z. cases met geldige waarden voor alle variabelen;
`CORPAIR`	gebruikt paarsgewijs die cases die voor beide variabelen een geldige waarde hebben.
`COVPAIR`	gebruikt voor de berekening van de covariantie paarsgewijs volledige cases maar bij de berekening van de standaard deviaties doen alle cases mee met geldige waarden voor de betreffende variabele.
`ALLVALUE`	gebruikt voor de berekening van de covariantie het gemiddelde voor een variabele gebaseerd op alle geldige cases.

In het manual wordt de gebruiker echter gewaarschuwd: "An important re-

quirement for valid estimation of the correlation or covariance matrix in program 8D (or program AM) is that missing data be missing completely at random." D.w.z. dat de 'volledige' cases beschouwd kunnen worden als een random steekproef. Als controle hierop geeft 8D

- de gemiddelden en varianties voor een variabele gebaseerd op paarsgewijs complete cases met alle andere variabelen

  MEANS OF COLUMN VARIABELE OVER CASES WHERE ROW VARIABELE IS PRESENT

  ---

	g8	g10	g12	g14
g8	22.1852	23.6111	24.6481	26.1346
g10	23.3333	23.6111	25.5556	26.6667
g12	22.1852	23.6111	24.6481	26.1346
g14	22.1923	23.6111	24.6923	26.1346

  De gemiddelden in een kolom zouden aan elkaar gelijk moeten zijn. Voor g8, g12 en g14 veranderen ze nogal bij de berekening voor alleen geldige cases op g10. Een soortgelijke tabel verkrijgt men voor de varianties.

- de t-toets en de variantie ratio voor alle paren van variabelen, waarbij een van de variabelen de groeperings variabele is (geldige waarde afwezig/aanwezig). In dit geval alleen interessant voor g10.

	g8	g10	g12	g14	In een cel staan:
g10	-4.49	0.00	-2.89	-1.29	T-waarde
	14.9	0.0	20.9	10.2	aantal vrijheidsgraden
	1.198	0.000	0.554	2.017	var ratio afwezig/aanw
	9 18	0 18	9 18	8 18	n afwezig n aanwezig

  De significante T-values duiden aan dat er in dit geval geen sprake is van "missing completely at random".

BMDPAM is bedoeld voor het schatten van de ontbrekende gegevens. Ook hier het advies om eerst naar de data te kijken en te bepalen of dit "missing at random" is. AM geeft hierbij als hulpmiddelen:

- correlaties tussen de gedichotomiseerde variabelen (wel of niet geldige value) die gelijk aan nul zouden moeten zijn.

	g10	g14
g10	1.000	
g14	0.277	1.000

- de 'patronen' van missing data

CASE LABEL	CASE NO.	NO OF MISS. VARS.	GROUP	WT.	sexe	g8	g10	g12	g14
3	3	1					M		
6	6	1					M		
9	9	1					M		
10	10	1					M		
2	13	1					M		
5	16	1					M		
12	23	1					M		
13	24	1					M		
16	27	2					M		M

De correlatiematrix kan berekend worden volgens drie methoden: COMPLETE, ALLVALUE en ML. De eerste twee zijn gedefinieerd als bij 8D echter wordt de matrix resulterend uit ALLVALUE nog "gesmoothed" zodat deze positief definiet is. De ML (maximum likelihood) methode is door Beale en Little ontwikkeld speciaal voor het schatten van de covariantiematrix in het geval van ontbrekende gegevens. Deze methode is te prefereren omdat de schattingen van de matrix consistent en efficiënt zijn ook al is niet voldaan aan de voorwaarde "missing completely at random" maar slechts aan de minder strenge voorwaarde "missing at random". Het ontbreken van gegevens behoeft in dat geval niet onafhankelijk te zijn van de geobserveerde variabelen, wel moet het missing data mechanisme van een variabele onafhankelijk zijn van het ontbreken van gegevens op andere variabelen. De eerder be-

schreven opties van BMDPAM kunnen gebruikt worden bij de controle van deze assumptie.

Voor het imputeren van de ontbrekende gegevens bij COMPLETE en ALLVALUE kan men kiezen uit de methodes:

MEAN imputeert het gemiddelde
SINGLE regressieschatter op basis van de hoogst gecorreleerde variabele
TWOSTEP regressieschatter op basis van max. 2 var. stepwise geselecteerd
STEP regressieschatter op basis van de variabelen stepwise geselecteerd
REGR regressieschatter op basis van alle variabelen

De maximum likelihood methode werkt echter alleen in kombinatie met de REGR methode.

Het gaat te ver om hier uitgebreid op de verschillende methoden en hun resultaten in te gaan. Hiervoor verwijzen we naar Little en Rubin (1987).

De schattingen voor de missing values kunnen in een pakket als BMDP uiteraard bewaard worden.

4. Conclusie

Met deze mogelijkheden biedt BMDP voldoende mogelijkheden voor het schatten en imputeren van ontbrekende gegevens voor het in de inleiding omschreven probleem van ontbrekende gegevens in de response variabele in een lineair model. Ook de andere 2 gevallen met missing values in de predictor variabelen geven geen problemen in BMDP.

5. Literatuur

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

Analyses van data met missende waarden bij SYSTAT

Jan Raatgever

1. Het pakket SYSTAT.

SYSTAT 4.0 is een algemeen statistisch pakket. Het pakket wordt geleverd samen met SYGRAPH; een mooi en uitgebreid programma met statistische grafieken. Het programma is modulair d.w.z. het bestaat uit een aantal modules die één voor één geladen moeten worden. De stuurtaal is en blijft steeds heel eenvoudig en inzichtelijk. De commando's werken in principe interactief, maar kunnen ook in een commando file opgeslagen worden. De manual is goed leesbaar en heeft soms meer weg van een statistiekboek dan van een naslagwerk. Daarom zijn dingen wel eens wat moeilijk te vinden temeer omdat een Command Reference ontbreekt, maar het nodigt wel meer uit tot lezen dan een puur naslagwerk. De lezer wordt steeds voorzien van voorbeelden en goed onderbouwde raadgevingen en waarschuwingen. Het pakket is geschikt voor statistici en diegenen die veel statistiek gebruiken en open staan voor de raadgevingen in het boek. Sinds kort is SYSTAT 5.0 beschikbaar. Deze versie heeft tal van verbeteringen en aanvullingen. Het programma is nu naast commando-gestuurd ook menu-gestuurd en er is nu gelukkig wel een Command Reference in de manual. Voor de hier beschreven analyses is echter nog de oude versie gebruikt.

2. Definitie van missende waarden in SYSTAT

Missende waarden worden intern gecodeerd als een waarde die kleiner is dan alle getallen die mogelijk zijn in SYSTAT: `-1.0E36`. Zij worden voorgesteld door een punt (.) als ze numeriek zijn en door een blank " " als het character data betreft. Per module wordt aangegeven hoe met missende waarden gerekend wordt.

3. De analyse van het groeimodel

Het inlezen van de data file gaat met de module DATA als volgt:

```
DATA
GET 'MISSYS.DAT'
SAVE MYSSYS
```

```
INPUT SEX ID AGE GROWTH
QUIT
```

De cases worden zo niet per individu, maar per meetpunt van GROWTH ingelezen. Het aantal cases is 108.
Er wordt een SYSTAT-file aangemaakt met de naam MISSYS.SYS. Om het effect van de leeftijd binnen SEX te kunnen schatten moeten we twee AGE-variabelen maken, die gelijk zijn aan de AGE voor de betreffende sexe en gelijk zijn aan nul voor de andere sexe. Ook wordt zo'n dummy variabele gemaakt van AGE voor ieder individu apart. Als we de leeftijd als een categorische variabele willen beschouwen, moeten we nog een nieuwe variabele maken: AGECAT, die de waarden 1, 2, 3 en 4 moet hebben voor de vier leeftijden die voorkomen.
Eén en ander gebeurt in de module DATA:

```
DATA
USE MISSYS
SAVE MISSDUM
IF SEX=1 THEN LET AGEJ=AGE
IF SEX=0 THEN LET AGEJ=0
IF SEX=1 THEN LET AGEM=0
IF SEX=0 THEN LET AGEM=AGE
DIM AGEP(27)
FOR I=1 TO 27
IF I=ID THEN LET AGEP(I)=AGE
ELSE LET AGEP(I)=0
NEXT
IF AGE=8 THEN LET AGECAT=1
IF AGE=10 THEN LET AGECAT=2
IF AGE=12 THEN LET AGECAT=3
IF AGE=14 THEN LET AGECAT=4
RUN
QUIT
```

Hierna hebben we de SYSTAT file MISSDUM met de variabelen: ID, SEX, AGEM, AGEJ, AGE, AGECAT, GROWTH en AGEP(27).
De variabele ID moet altijd als categorisch beschouwd worden, dit kan met het commando:

CATEGORY ID=27

binnen het programma MGLH. De module MGLH (Multivariate General Linear Hypothesis) wordt gebruikt om het model te schatten. SYSTAT genereert nu zelf k–1 (=26) variabelen die een niveauverschil ten opzichte van een constante term in het model vertegenwoordigen. Hierdoor worden 27 niveaus (één voor ieder kind) onderscheiden in het lineaire model. De schatting van het model in MGLH gaat nu als volgt: bijvoorbeeld voor het model met gelijke helling voor alle meisjes en voor alle jongens met verschillende intercepts voor ieder kind:

```
MGLH
USE MISSDUM
SAVE MODEL8/RESID
CATEGORY ID=27
MODEL GROWTH=CONSTANT+AGEM+AGEJ+ID
ESTIMATE
QUIT
```

De SAVE opdracht maakt een SYSTAT file waarin o.a. de residuen en de estimated waarden weggeschreven worden. Om de kwadratensom van de verschillen tussen de geschatte waarden en de (stiekem toch) bekende invulling voor de missende waarden te berekenen, moeten we weer terug naar de module DATA. Allereerst moet nog een SYSTAT file gemaakt worden uit een data file waar de missende waarden door de goede waarden zijn vervangen (MISSC.DAT):

```
DATA
SAVE MISSC
GET 'MISSC.DAT'
INPUT SEX ID AGE GROWTH
RUN
QUIT
```

Daarna worden de SYSTAT files MISSC.SYS en bijvoorbeeld MODEL1.SYS aan elkaar gekoppeld en wordt de kwadratensom uitgerekend:

```
DATA
USE MISSC MODEL1
HOLD
```

```
LET N=N+1
IF N=10 OR N=22 OR N=34,
OR N=38 OR N=50 OR N=62 OR N=90 OR N=94,
OR N=106 OR N=108 THEN FOR
  LET KWAD=(GROWTH-ESTIMATE)*(GROWTH-ESTMATE)
  LET SUMKWAD=KWAD+SUMKWAD
NEXT
IF EOF THEN PRINT 'de kwadatensom is:', SUMKWAD
RUN
```

De truc hier is: het gebruik van het HOLD commando dat de waarden van een (vorige) case vasthoudt.
Voor een aantal modellen is op deze wijze de kwadratensom uitgerekend. Steeds wordt ook het MODEL-commando vermeld, zoals dat in de module MGLH gegeven wordt:

Model1: Gemiddelde per combinatie van leeftijd en geslacht berekenen.

```
CATEGORY AGECAT=4
MODEL GROWTH=CONSTANT+AGECAT+SEX
```

Model2: Twee verschillende lijnen met ieder een eigen intercept en helling.

```
MODEL GROWTH=CONSTANT+AGEJ+AGEM+SEX
```

Model3: Twee lijnen met gemeenschappelijke helling en ieder een apart intercept.

```
MODEL GROWTH=CONSTANT+AGE+SEX
```

Model4: Ieder individu zijn eigen intercept. Leeftijd, leeftijd in het kwadraat en leeftijd tot de derde macht en interacties daarvan met geslacht.

```
CATEGORY ID=27
MODEL GROWTH=CONSTANT+AGE+AGE*AGE+,
   AGE*AGE*AGE+AGE*SEX+,
   AGE*AGE*SEX+AGE*AGE*AGE*SEX+ID
```

Model5: Ieder individu zijn eigen helling en intercept.

```
CATEGORY ID=27
MODEL GROWTH=CONSTANT+AGEP1+AGEP2+,
  ..+AGEP27+ID
```

Model6: Ieder individu zijn eigen intercept. Verder de leeftijd in het kwadraat en de interactie hiervan met het geslacht.

```
CATEGORY ID=27
```

```
        MODEL GROWTH=CONSTANT+AGE*AGE+,
          AGE*AGE*SEX+ID
```

Model7: 27 intercepts en één enkele helling.

```
        CATEGORY ID=27
        MODEL GROWTH=CONSTANT+AGE+ID
```

Model8: 27 intercepts en per sexe een helling.

```
        CATEGORY ID=27
        MODEL GROWTH=CONSTANT+AGEJ+AGEM+ID
```

De volgende tabel geeft de resultaten van deze modellen in de vorm van de kwadratensom voor de missende waarden en de kwadratensom (residuele som) van de 98 niet missende punten. De volgorde van de modellen is van een hoge residuele kwadratensom naar een lagere. De modellen worden in het algemeen in deze volgorde ook steeds complexer.

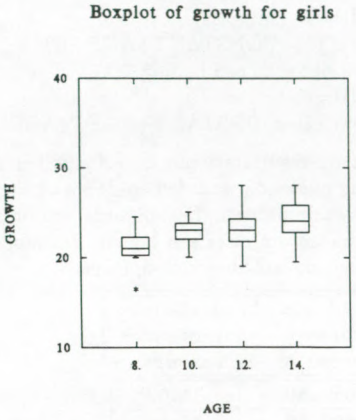
Model	aantal para-meters	residuele kwadra-tensom	kwadratensom voor missen-de waarden
model3	3	509.30	34.07
model1	8	508.48	36.16
model2	4	495.71	35.65
model7	28	156.20	10.82
model8	29	144.56	10.66
model6	29	140.94	13.02
model4	33	136.47	20.47
model5	54	78.46	20.36

Uit de tabel zien we dat al te nauwkeurig fitten van het model met 98 meetpunten weer resulteert in het onnauwkeurig schatten van de 10 missende waarden. Het optimum ligt bij model8. In de figuren 10 en 11 zijn Box plots gegeven van de verdeling van GROWTH voor iedere leeftijd en per sexe. Deze figuren suggereren dat een tweede- of derdegraads kromme afhankelijk van de leeftijd en verschillend per geslacht, goed zou kunnen fitten. Dit komt ook uit de resultaten naar voren (model6 en model4).

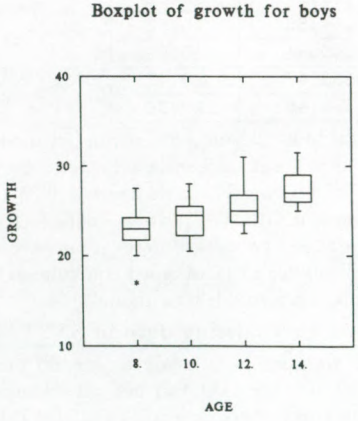
4. Aanpak van analyses met missing data in SYSTAT

De aanpak van missende waarden in de analyse verschilt per module. In de module MGLH wordt gebruik gemaakt van het verwijderen van een complete case als een van de variabelen missend is (LISTWISE deletion). De

Figuur 10: Box plots van GROWTH voor vier leeftijden voor de 11 meisjes



Figuur 11: Box plots van GROWTH voor vier leeftijden voor de 16 jongens



module CORR kan een correlatiematrix berekenen volgens LISTWISE en PAIRWISE, d.w.z. alle waarden die bekend zijn van twee variabelen gebruiken om de correlatie uit te rekenen ongeacht of er voor andere variabelen missende waarden voorkomen. Het is mogelijk een model te fitten in de module MGLH uitgaande van een correlatie- of covariantiematrix afkomstig van de module CORR. Op deze wijze kan dus MGLH ook met PAIRWISE deletion werken. Voor testen van de effecten kan gecorrigeerd worden door de '/N=' optie bij het MODEL-commando. Hier kan bijvoorbeeld als N gekozen worden: het laagste aantal niet missende waarden dat voorkomt bij alle paren variabelen.

Bij de module CORR wordt voor het gebruik van PAIRWISE deletion een waarschuwing in de manual gegeven dat als er een te groot percentage missende waarden voorkomt (bijvoorbeeld 20%), de subsets van de cases per correlatie al wezenlijk van elkaar kunnen verschillen. Daarom worden die pakketten gehekeld die deze methode als een standaard methode bij regressie gebruiken. Aangeraden wordt de data in zo'n geval te schatten, bijvoorbeeld m.b.v. Cluster Analyse van SYSTAT. De MATRIX-optie van het JOIN commando hiervan geeft een rangschikking van de cases zodanig dat ontbrekende waarden van een variabele uit de waarden van de twee sterk gelijkende "buren" geschat kunnen worden.

Hetzelfde JOIN/MATRIX commando kan ook toegepast worden op een file waarin de missende gelijk gemaakt zijn aan 1 en de niet missende waarden 0 gemaakt zijn. De uitvoer geeft dan inzicht in de patronen (afhankelijkheden) in de verdeling van de missende waarden.

SURFOX

Jan Hooft van Huijsduijnen

SURFOX is een door AB-onderzoek te Delft ontwikkeld nederlandstalig pakket voor het analyseren, bewerken en verwerken van bestanden en enquêtes.

In 1986 is gestart met de ontwikkeling van SURFOX, omdat bestaande statistische pakketten zeer traag werkten bij grote bestanden en daardoor niet geschikt bleken voor interactief gebruik.
SURFOX kan bestanden aan tot 65000 records en 1000 variabelen. Ook bij bestanden van deze grootte verschijnen tabellen en analyse-resultaten binnen enkele seconden op het scherm.

SURFOX is in 1989 op de markt gekomen voor PC's (beperkt in het aantal records en variabelen), PRIME-computers en computers draaiend onder UNIX. Het bevat in grote lijnen dezelfde functionaliteit als van bijvoorbeeld het basispakket van SPSS en heeft daarbij een aantal extra's zoals modules voor het maken van wegingsvariabelen, het bijschatten van 'missing values' en het opstellen, invoeren en verwerken van complexe enquêtes. Het pakket is menugestuurd, waardoor geen commando's behoeven te worden geleerd. Er bestaat tevens een batch-versie van SURFOX.

1. Oplossing probleemstelling

Alvorens in te gaan op de uitwerking van het gestelde probleem zijn twee zaken binnen SURFOX van belang.

- Allereerst betreft dit het schaal-niveau van de variabelen. Variabelen worden door SURFOX in eerste instantie ingedeeld in nominale, ordinale, interval- en ratio-variabelen op basis van het aantal verschillende waarden en het type waarden (gehele of gebroken getallen). De gebruiker kan uiteraard deze indeling veranderen. Getoonde statistics houden rekening met dit schaalniveau. Zo wordt geen gemiddelde gegeven van nominale variabelen.

- Surfox maakt onderscheid tussen geldige waarden, NVT's (Niet van Toepassing) en TOGA's (Ten Onrechte Geen Antwoord). TOGA's komen overeen met de bekende 'missing values'. Deze kunnen binnen SURFOX worden bijgeschat. NVT's zijn terechte 'missing values' en kunnen daardoor niet worden bijgeschat. De namen TOGA en NVT zijn interne koderingen van SURFOX. Er wordt altijd rekening gehouden bij berekeningen, presentatie en analyses met deze 'missing values' en NVT's. Het toekennen van een 'missing value' of NVT binnen SURFOX volgt of automatisch uit de routing van een vragenlijst, of door het expliciet toekennen van deze waarden door de gebruiker binnen

IF-ELSEIF-ELSE berekeningen.

Voor het bijschatten van de 'missing values' in het gestelde probleem wordt gebruik gemaakt van de regressie-module van SURFOX. Het volgende scherm verschijnt op de terminal:

```
ANALYSE VARIABELE ------------------> groei
VERKLARENDE VARIABELEN -------------> lftm, lftj, case
SELECTIE ---------------------------->
WEEGFACTOR -------------------------->
```

Er hoeft geen selectie te worden gemaakt; records met NVT's en 'missing values' van de te verklaren variabele worden automatisch buiten de berekeningen gehouden. Omdat 'case' nominaal is, wordt voor elke verschillende waarde van deze variabele evenveel dummy-variabelen aangemaakt. De variabelen 'lftm' en 'lftj' zijn interval-variabelen. SURFOX maakt van elk van beide variabelen binnen de module regressie twee nieuwe variabelen; een met de geldige waarden en met een 0 op de plaats van de NVT's en een tweede variabele met een 1 op de plaats van de NVT's en de rest 0.
Eventuele 'missing values' van de verklarende variabelen worden op dezelfde manier als de NVT's als verklarende dummy-variabele toegevoegd. SURFOX kent daardoor binnen regressies geen 'listwise deletion'.

Na de invoer van de variabelen verschijnt er een scherm, waarin stapsgewijs variabelen kunnen worden toegevoegd aan of verwijderd uit het model. Geschatte waarden en/of de residuen kunnen worden bewaard onder een nieuwe variabele-naam.

2. BIJSCHATTEN van MISSING VALUES (algemeen)

In de praktijk komen de meeste 'missing values' voor in steekproeven. Omdat op geaggregeerd niveau uitspraken moeten worden gedaan over de doelpopulatie is het probleem van 'missing values' hier wel groter dan bij kleine bestanden, waar het analyseren van verbanden centraal staat. Unit non-respons (geen enkele vraag is door de respondent beantwoord) kan worden gecorrigeerd door ophoogfactoren. Voor item non-respons ('missing values') zal de onderzoeker per variabele een uitspraak moeten doen. Doet hij/zij er niets mee, dan wordt er impliciet van uitgegaan dat de non-respons verdeeld is volgens de respons op het moment dat de gegevens gegeneraliseerd worden. *Niet bijschatten is ook bijschatten.* Meestal is er sprake van selectieve

non-respons en geeft het weglaten van 'missing values' uit de tabellen en berekeningen een vertekend beeld van de werkelijkheid.
In onderstaande tabel staat een overzicht van de belangrijkste methoden waarmee 'missing values' kunnen worden bijgeschat, gespecificeerd naar het type verklarende of te verklaren variabelen. De methoden PMM en Random Hotdeck komen in de volgende paragraaf aan de orde. De overige methoden trachten op record-niveau een zo'n goed mogelijke schatting te maken, met het gevolg dat de verdeling niet altijd goed zal zijn.

Verklarende variabelen	Afhankelijke variabele	
	continu	discreet
continu	Regressie	Discriminant-analyse
discreet	Anova	Logit-analyses
continu/ discreet	Regressie PMM	Random Hotdeck

Aan regressies als methode voor het bijschatten van 'missing values' kleven de volgende belangrijke bezwaren:

- regressies zijn eigenlijk alleen geschikt voor ratio-variabelen. In veel steekproeven zijn nominale variabelen in de meerderheid (geslacht, opleiding, type woning, beroepsklasse, burgerlijke staat, regio etc...);

- regressie-schatters kunnen buiten het domein van de steekproef treden. Het is goed mogelijk dat bijvoorbeeld leeftijden worden geschat van -6 of 140;

- een van de belangrijkste nadelen van regressie-modellen is de onderschatting van de variantie. In onderstaande tabel staan de schattingsresultaten van een regressie-model naast de oorspronkelijke waarden. De waarden zijn in klassen weergegeven. De correlatie-coëfficiënt bedroeg 0.76. Ondanks deze relatief hoge R2 tenderen de geschatte waarden naar het gemiddelde.

Herstelkosten (in guldens)	Oorspronkelijk (aantal records)	Geschat (aantal records)
0 tot 500	64	1
500 tot 1000	43	0
1000 tot 2000	162	164
2000 tot 3000	612	444
3000 tot 5000	624	609
5000 tot 10000	768	895
10000 tot 20000	624	977
20000 tot 30000	215	88
30000 tot 40000	51	0
40000 en meer	15	0
	----	----
Totaal	3178	3178

Bron: AB-onderzoek, Koppeling KWR aan WBO, februari 1985

De onderschatting van de variantie zal voorkomen bij elke schattingsmethode waarbij getracht wordt op record-niveau een zo goed mogelijke schatting te maken.

3. BIJSCHATTEN binnen SURFOX

Bovengenoemde bezwaren van regressies worden in de literatuur onderkend en er zijn diverse oplossingen gevonden. Een ervan is voor elk record een randomtrekking te doen uit de normale verdeling van de storingsterm en deze op te tellen bij de geschatte waarde. Er wordt dan bewust afgezien van de beste schatter op record-niveau. Deze oplossing onderschat de variantie in veel mindere mate, maar heeft als bezwaar dat waarden buiten het domein van de steekproef of doelpopulatie kunnen vallen.

Predictive mean matching

Een oplossing van Little (1988) komt aan de meeste bezwaren tegemoet voor het bijschatten van ratio-variabelen. De door hem geïntroduceerde methode 'predictive mean matching' (PMM) is ook in SURFOX ingebouwd. De basis van de methode vormen de geschatte waarden van de afhankelijke variabele voor alle records met geldige waarden en 'missing values'. Voor elk record met een 'missing value' wordt gezocht naar een donor-record, waarvan de geschatte waarde zo dicht mogelijk ligt bij de geschatte waarde van het ongeldige record.

Vervolgens wordt van het donor-record de oorspronkelijke waarde genomen en geïmputeerd voor de 'missing value'. Waarden blijven hierdoor binnen het domein van de steekproef (wat bij kleine steekproeven weer een bezwaar kan zijn) en de variantie wordt in veel mindere mate onderschat.

Random hotdeck
PMM is voor interval en ratio-variabelen een goede imputatie-techniek. Voor nominale en ordinale variabelen is binnen SURFOX een andere oplossing gebouwd. Dit is de 'random-hotdeck methode' (RHM). Met deze methode wordt wederom niet getracht de beste schatter op record-niveau te vinden, maar is de verdeling van de bijgeschatte variabele belangrijker. Dit kan worden geïllustreerd door het volgende voorbeeld. Wanneer alleen bekend is dat 70% van de personen in een huurwoning woont en 30% in een koopwoning en u zou van 100 personen moeten schatten of ze in een huur- of koopwoning wonen, kunt u twee dingen doen:

- alle 100 personen als huurder aanmerken. U zit dan bij 70 personen goed. Toch is het duidelijk dat dit geen goede schatting is;
- u kunt ook willekeurig 70 personen tot huurder verklaren en 30 personen tot koper. U zult dan in ongeveer 0.7x70+0.3x30=58% gevallen goed zitten. Minder dan in het eerste geval, maar de verdeling is in ieder geval goed.

Deze tweede methode is uit te breiden met een volgende variabele. Stel dat bekend is dat 80% van de personen met een gouden ring in een koopwoning woont, dan kan door het waarnemen van een gouden ring een betere schatting worden gemaakt op recordniveau met handhaving van een goede verdeling. Binnen SURFOX kunnen maximaal 10 verklarende variabelen opgenomen. Dit mag elk type variabele zijn. Ratio-variabelen moeten dan in klassen worden ingedeeld.
Verschillende statistics zijn opgenomen om de mate van verklaring te meten. Dit zijn onder andere de verklaarde entropie en de PRU-maat (oftewel de 'uncertainty coëfficiënt'). Het te bouwen model is een verzadigd log-lineair model, wat betekent dat alle interacties tussen de verklarende variabelen worden meegenomen.

Binnen SURFOX wordt dit bijschatten 'klonen' genoemd, omdat getracht wordt voor elke record met een 'missing value' een donor-record te zoeken,

waarin de verklarende variabelen zoveel mogelijk overeenkomen. De geldige waarde uit het donor-record wordt vervolgens geïmputeerd. Ook is het mogelijk om voor meerdere 'missing values' binnen een record de geldige waarden van een donorrecord te imputeren.

Er wordt nog veel onderzoek gedaan naar het bijschatten van 'missing values'. Rubin heeft in diverse publicaties en boeken het belang aangetoond van 'multiple imputatie' (1987). Als basis dienen hiervoor imputatie-technieken zoals de in SURFOX bestaande PMM en RHM. De statistics die worden berekend met een bijgeschatte variabele (bijgeschat m.b.v. een 'single' imputatie-techniek) zijn nooit zuiver. Door 'multiple imputation' kunnen praktisch zuivere betrouwbaarheidsintervallen van de statistics worden berekend (Little, Rubin, 1989).

4. Literatuur

(1987) Little, R.J.A. & D.B. Rubin
Statistical Analysis with Missing Data.
New York, J. Wiley

(1988) Little, R.J.A.
Missing-Data Adjustments in Large Surveys.
Journal of Business & Economic Statistics,
July 1988, Vol. 6 No.3, pag 287-296

(1989) Little, R.J.A. & D.B. Rubin
The analyses of Social Science Data with Missing Values.
Sociological methods and research, Vol 18 Nos 2 & 3,
November 1989/February 1990, pag. 292-326.

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