7. SECONDAIRE ANALYSE 'VAN JAAR TOT JAAR' MET BEHULP VAN NIET-LINEAIRE MULTIVARIATE TECHNIEKEN

Jan de Leeuw; Ineke Stoop

Samenvatting

Met behulp van descriptieve multivariate technieken wordt nagegaan in hoeverre de assumpties van lineariteit en additiviteit opgaan in een eerder door Dronkers geanalyseerde subset van de 'Van Jaar tot Jaar' gegevens. Er blijken verschillende interessante niet-lineaire verbanden te vinden te zijn, die de multivariate structuur van de variabelen verduidelijken. We benadrukken overigens dat onze technieken geen functionele of causale relaties tussen variabelen aantonen, of zelfs maar aannemelijk maken.

0. Inleiding

0.1. De getallenberg

De wetenschapsopvatting van de psychometrische of sociometrische tak van de sociale wetenschappen is traditioneel een vorm van naïef empirisme. Zoals bekend zijn de sociale wetenschappen er in honderd jaar empirisch onderzoek nog nauwelijk in geslaagd om stabiele functionele verbanden te vinden, die zich in de eerste plaats tot niet-triviale deelgebieden beperken, en die zich in de tweede plaats in interessante en maatschappelijk belangrijke situaties toe laten passen. De remedie ligt volgens de psychometrici voor de hand. We moeten meer gegevens verzamelen, meer variabelen meten, meer tests construeren, meer analytische technieken ontwikkelen, en grotere stapels computer-output fabriceren. Daarnaast moeten we ook betere gegevens verzamelen, betere tests construeren, betere technieken ontwikkelen, en betere variabelen meten. De recente herrijzing van de psychometrische genetica en van verwante biologisch georiënteerde benaderingen kan het beste begrepen worden als men zich op dit naïef empiristisch standpunt stelt. Psychobiologen meten meer variabelen, en hebben maar al te vaak de indruk dat ze betere variabelen meten.

Het zou voor de hand liggen in dit artikel het werk van Jensen, Eysenck of Jencks als voorbeeld van de psychometrische aanpak te behandelen. We kiezen echter een wat actueler voorbeeld. In Buikhuisen (1978) wordt vastgesteld dat de criminologie er niet in geslaagd is misdadig gedrag afdoende te verklaren uit factoren die samenhangen met onderwijs, opvoeding en inkomen. In psychometrische termen:

slechts een relatief klein percentage van de variantie in crimineel gedrag kan 'verklaard' worden uit 'omgevings'-variabelen (l.c. pag. 13). En omdat er maar weinig variantie verklaard wordt, kan de criminologie ook geen effectieve preventieve maatregelen voorstellen, met als direct gevolg dat criminaliteit niet afneemt, en dat het aantal recidivisten groot blijft. De verklaring voor het falen van de criminologie in dit opzicht moet echter niet gezocht worden in het feit dat de sociale wetenschappen er niet in geslaagd zijn de criminogene factoren in de omgeving te ontdekken, maar in het feit dat de sociale wetenschappen altijd de biologische componenten verwaarloosd hebben. Deze redenering suggereert natuurlijk, dat het benadrukken van verschillen in omgeving faalt, omdat de ware oorzaken in de delinquent zelf liggen. De man heeft een misdadige persoonlijkheid, hij scoort hoog op de criminaliteitsschaal (l.c. pag. 14), omgevingsvariabelen werken hoogstens faciliterend of inhiberend (l.c. pag. 16-17). De redenering komt overeen met die van Jensen (1969), die het falen van compensatieprogramma's verklaart uit de genetisch bepaalde domheid van de doelgroepen en met die van Jencks (1972), die tot de conclusie komt dat iemands inkomen voor het grootste gedeelte bepaald wordt door dom geluk. In tijden van financiële schaarste wordt het bovendien extra pijnlijk dat de sociaal wetenschappelijke aanpak van maatschappelijke problemen weinig succes heeft. Om de geldstroom nog enigszins in stand te houden of om die geldstroom te beknotten en af te buigen blijkt het effectief om de schuld te geven aan de proefpersonen of aan oncontroleerbare zaken als geluk. De groteske opmars van grauwe ganzen en stekelbaarzen, die bekend staat als sociobiologie, kan alleen maar begrepen worden vanuit het falen der sociale wetenschappen, en dit falen is het directe gevolg van de atheoretische en anti-experimentele instelling van veruit de belangrijkste tak van die sociale wetenschappen.

Een wetenschap in nood wendt zich tot de biologie. Voor de zoveelste maal overigens. De constatering van Jensen, Eysenck, Buikhuisen en anderen dat de sociale wetenschappen niets van biologische variabelen moeten hebben is historisch gezien volkomen onjuist. In de criminologie gaf tot ongeveer 1910 de criminele antropologie van Lombroso wetenschappelijk de toon aan. Lombroso en zijn school besteedden veel aandacht aan biologische variabelen. De man die het meeste bijdroeg aan de val van Lombroso ~~van~~ was Charles Goring, een discipel van Karl Pearson, en minstens even biologisch georiënteerd als de latere Lombroso. Tussen de twee wereldoorlogen werd de toon aangegeven door de Duitse criminologie, zich zichzelf aanduidde als Kriminalbiologie of als Psychobiologische Verbrecherkunde. Pas na de tweede wereldoorlog kwam de nadruk duidelijk op de omgeving te liggen, hoewel

sommige invloedrijke theorieën over crimineel gedrag zoals die van Eysenck (1964) nog steeds expliciet psychobiologisch zijn. Dezelfde tendens vinden we in de 'mental test' beweging. Dat biologische factoren veel belangrijker zijn dan culturele stond voor Galton, Cattell. Thorndike en Terman als een paal boven water. Pas ver na de eerste wereldoorlog begon men aan onderzoeken over veranderlijkheid en beïnvloedbaarheid van het IQ. De resultaten waren helemaal niet indrukwekkend en verre van eenduidig. Mede daardoor is een sterk biologisch georiënteerde vleugel blijven bestaan, en met name toonaangevende figuren als Burt, Eysenck en Cattell staan nog precies op hetzelfde standpunt als Galton (volgens hen natuurlijk met aanzienlijk meer bewijsmateriaal dan Galton had).

Op zichzelf is de constatering dat het percentage 'verklaarde' variantie groter wordt wanneer we genetische, psychofysiologische en neuropsychologische variabelen toevoegen natuurlijk niet onjuist. De bewering is triviaal, omdat bij gebruik van meer predictoren de predictie per definitie niet slechter wordt. Een nogal afschrikwekkend voorbeeld van zo'n psychometrische 'theorie', die in feite alleen maar bestaat uit een immense hoeveelheid metingen, vinden we in Broman e.a. (1975). Daar wordt het IQ op 4-jarige leeftijd voorspeld met behulp van meer dan 150 prenatale, perinatale en postnatale variabelen. Uit het feit dat meer dan 70% van de variantie ''onverklaard' achterblijft, zou Jencks ongetwijfeld concluderen dat IQ op 4-jarige leeftijd voornamelijk bepaald wordt door geluk. Het lijkt erop dat Buikhuisen's befaamde pseudo-formule C $f(P_i, S_i)$ alleen maar een voorbeeld is van dit 'hoe meer, hoe beter' principe, en zijn pleidooi voor multidisciplinaire en interdisciplinaire aanpak wijst ook in die richting. Pas als Buikhuisen variabelen als het activatieniveau en de aan- of afwezigheid van bepaalde stoffen in het bloed terloops bespreekt alsof het factoren zijn waarvan de samenhang met criminaliteit al overduidelijk aangetoond is, pas dan komt de sociobiologische aap uit de mouw. Uit het alleszins moderne en uitgebalanceerde overzicht van Rutter en Madge (1976) krijgen we een betere indruk over het feitelijk belang van biologische variabelen. En zelfs in het biocriminologische boek van Mednick en Christiansen (1977), wordt op een aanzienlijk voorzichtiger manier met biologisch bewijsmateriaal omgesprongen.

0.2. Correlatie en causatie

Het zwakke punt in de psychometrische en aanverwante theorieën zit hem natuurlijk in het gebruik van het woord 'verklaren', een woord dat in meer respectabele wetenschappen gebruikt wordt als een synoniem voor het vinden van invariante functionele verbanden. Wat de psychometrici vinden zijn correlaties, of verschillen

tussen groepen van bijvoorbeeld criminelen en niet-criminelen. De constatering dat iedere toekomstige theorie over crimineel gedrag tenminste deze verschillen moet kunnen verklaren lijkt voor de hand te liggen. Maar vooralsnog worden de psychofysiologische verschillen bijvoorbeeld soms wel en soms niet gevonden, vooralsnog zijn de resultaten van tweeling- en adoptiestudies op het gebied van criminaliteit of intelligentie onoverzichtelijk of zelfs verdacht, en is, om kort te gaan, iedere vorm van invariantie in de gevonden resultaten ver te zoeken. De hypothese dat de werking van het autonome zenuwstelsel bij psychopaten afwijkt, en de eerste pogingen om dit experimenteel vast te stellen, vinden we al bij Lombroso. De meetinstrumenten zijn verbeterd, de steekproeven zijn aanzienlijk groter en aanzienlijk aselecter en de databewerkingstechnieken zijn geperfectioneerd. Maar dat is alleen maar technologische vooruitgang, de theorievorming ontbreekt nog steeds, en de resultaten zijn nog steeds verre van eenduidig.

In de hoogtijdagen van de factor analyse beweerden psychometrici als Cattell dat factor analyse een unieke techniek was, perfect voor psychobiologisch onderzoek. In de eerste plaats omdat er geen enkele theorie nodig was om het toe te kunnen passen, je hoefde alleen maar correlaties uit te kunnen rekenen (en zelfs dat is nu niet meer nodig). En in de tweede plaats omdat er uit een toepassing van factor analyse automatisch een theorie te voorschijn kwam. Nogmaals een afschrikwekkend voorbeeld van naïef empirisme: een theorie wordt beschouwd als een handige ordening van een zo groot mogelijk aantal feiten. Door middel van factor analyse kan deze ordening bovendien ook nog geheel mechanisch geschieden. De wetenschapsfilosofie die hier achter steekt is het duidelijkst geformuleerd in het boek van Pearson (1892), en komt overeen met de ideeën van negentiende eeuwse fysici als Mach, Maxwell, en Kelvin. Het zijn dezelfde fysici waaraan Eysenck zo graag de motto's van zijn boeken ontleent, en het is dezelfde Pearson die de moderne statistiek op poten zette, die zich druk bezighield met de Eugenetica, en die veel moeite deed om de erfelijkheid van intelligentie aan te tonen. De laatste aanwinst in het Galton-Pearson scala van data analytische technieken zijn de zogenaamde 'causale' modellen. Zoals uiteengezet in het boek van Pearson met name in de latere edities van 1900 en 1911, is correlatie meer een fundamenteel begrip dan causatie. Causatie is het grensgeval van perfecte correlatie, en in de biologie , psychologie, en sociologie zijn de correlaties meestal niet perfect. Pearson's fenomenalisme ontkende in feite het bestaan van functionele verbanden, het enige met werkelijkheidswaarde is de contingentie van zintuigelijke ervaringen. In de geest van Pearson moeten we 'causale modellen' dan ook be-

grijpen als technieken om correlaties op een eenvoudige manier te beschrijven. Pearson, en na hem de factor analytici, hebben het voortdurend over het Scheermes van Occam. In de context waarin zij het gebruiken betekent dit dat de meest eenvoudige wetenschappelijke theorie ook de juiste wetenschappelijke theorie is. Dit is een logisch gevolg van het instrumentalistische uitgangspunt, maar een duidelijk antwoord op de vraag wat nu eigenlijk met 'eenvoudig' bedoeld wordt is nooit gegeven. In het kader van parametrische statistiek kunnen ad hoc antwoorden gegeven worden (vrijheidsgraden), maar het recente IQ debat laat overtuigend zien dat aantoonbaar onjuiste, maar zeer eenvoudige, statistische modellen de beschikbare gegevens goed beschrijven. Er zijn veel onzinnige modellen met een prima fit. De effectiviteit van 'causale' of pad-modellen in de genetica is gebaseerd op het feit dat Mendeliaanse assumpties tot lineariteit van regressie leiden, terwijl de waarde van Mendeliaanse assumpties aangetoond is in experimenteel en toegepast genetisch onderzoek. De psychometrische genetica probeert de redenering om te draaien: uit de lineariteit van regressie, uit de regressie naar het gemiddelde, uit de correlaties van IQ's van familieleden, volgt dat het Mendeliaanse model opgaat. Maar deze redenering gaat niet op. En op dezelfde manier betekent een grote padcoëfficiënt geen belangrijk causaal verband, betekent een h^2 van .80 niet dat intelligentie grotendeels erfelijk bepaald is, en betekent een verklaarde variantie van 20% niet dat inkomen voor 80% wordt bepaald door geluk. En als criminelen psychofysiologisch verschillen van niet-criminelen dan betekent dat niet dat criminaliteit biologisch bepaald is. Het gaat in de wetenschap om veronderstelde functionele verbanden, daaruit afgeleide predicties en daarop gebaseerde falsificatiepogingen. Causaliteit is een onderdeel van een wetenschappelijke theorie, functionele verbanden worden niet aangetoond door mechanische manipulatie van gegevens.

0.3. En dit dan?

Na deze inleiding lijkt het misschien wat vreemd dat dit artikel een grote hoeveelheid verse correlaties de wereld instuurt, een stapeltje computer output toevoegt aan de bestaande stapel, en multivariate technieken uit de Galton-Pearson school toepast op gegevens die duidelijk iets te maken hebben met maatschappelijk relevante problemen. Ons excuus is dat de 'Van Jaar tot Jaar' gegevens interessant genoeg zijn om zo volledig mogelijk op statistische verbanden onderzocht te worden. De feitelijke constatering van een statistisch verband impliceert niets over eventuele causale relaties, maar aannemen van causale relaties impliceert wel degelijk iets over het al dan niet bestaan van statistische verbanden. Boven-

dien kunnen descriptieve en exploratieve studies in een later stadium tot theorievorming leiden. Het is natuurlijk juist dat factor analytici en multivariate analysten van de meer bescheiden school al tientallen jaren zeggen dat hun explorations in een later stadium tot theorievorming en tot confirmatief onderzoek moeten leiden, en dat er van die onderzoekingen in dat latere stadium dan nooit meer iets vernomen wordt. Dat is heel jammer, maar wij geloven niet dat het een reden is om geheel met exploratief onderzoek te stoppen, of dat het een reden is om bestaande onderzoeken niet meer te verbeteren. Het door het I.T.S. verzamelde materiaal is veruit het interessantste dat op het ogenblik beschikbaar is in Nederland en het lijkt interessant om na te gaan in hoeverre diverse statistische aannamen als lineariteit, additiviteit, en multivariate normaliteit voor dit materiaal opgaan. Als blijkt dat deze aannamen niet al te absurd zijn, dan worden analyses als die van Dronkers (1979) meer waardevol. We willen overigens nogmaals benadrukken, dat we een presentatie als die van Dronkers (1979) of Dronkers en Jungbluth (1979) riskant vinden, omdat het gebruik van functionele terminologie ons in dit stadium te veel lijkt te suggereren, met name aan die soorten gebruikers die de methodologische voorbehouden oninteressant vinden en niet de moeite waard om rekening mee te houden. In De Leeuw (1978) wordt dit standpunt nader toegelicht.

De niet-lineaire technieken die we gebruiken vallen in twee categorieën uiteen. In het eerste geval (behandeld in paragraaf 1) bestuderen we homogeniteit van de 25 belangrijkste variabelen uit het 'Van Jaar tot Jaar' onderzoek, dat wil zeggen we gaan na hoe ze met elkaar samenhangen, en in hoeverre ze tot één enkele schaal gereduceerd kunnen worden. Bovendien wordt nagegaan in hoeverre lineariteit en multivariate normaliteit volgehouden kunnen worden. De benodigde theorie wordt in 1.1. uiteengezet, de resultaten worden besproken in 1.2. In paragraaf 2 behandelen we niet-lineaire technieken om het bereikte onderwijsniveau te prediceren. De in 2.1. besproken theorie maakt duidelijk dat we in dit geval ook additiviteit kunnen onderzoeken. De resultaten worden in 2.2. besproken, en in paragraaf 3 tenslotte worden de resultaten van de paragrafen 1 en 2 vergeleken en samengevat.

Wij danken het Instituut voor Toegepaste Sociologie voor het ter beschikking stellen van het materiaal, de heer Schrik van het Steinmetz Archief voor het behulpzaam zijn bij het ontcijferen van de tape, en Jaap Dronkers van het SISWO voor aanmoediging en inspiratie.

1. Homogeniteit

1.1. Theorie

1.1.1. Definitie

Stel $\underline{x}_1,\ldots,\underline{x}_m$ zijn stochastische grootheden (random variables). We nemen aan dat de verwachte waarden $E(\underline{x}_j)$ gelijk aan nul zijn voor alle j, en dat alle varianties $\sigma_j^2 = E(\underline{x}_j^2)$ bestaan. Definieer

$$\underline{x}_\cdot = \frac{1}{m} \sum_{j=1}^{m} \underline{x}_j .$$

We kunnen nu $\underline{x}_\cdot$ gebruiken als een variabele die de informatie in alle m variabelen $\underline{x}_j$ samenvat. Om te meten hoe goed $\underline{x}_\cdot$ die informatie samenvat, gebruiken we de bekende relatie

$$\sum_{j=1}^{m} E(\underline{x}_j^2) = m\, E(\underline{x}_\cdot^2) + \sum_{j=1}^{m} E\{(\underline{x}_j - \underline{x}_\cdot)^2\}$$

De term links schrijven we als $T(\underline{x}_1, \ldots,\underline{x}_m)$, en noemen we de *totale variantie*. De eerste term aan de rechterkant is de variantie *tussen* variabelen, geschreven als $B(\underline{x}_1,\ldots,\underline{x}_m)$, en de tweede term is de variantie *binnen* variabelen, geschreven als $W(\underline{x}_1,\ldots,\underline{x}_m)$. Als $W(\underline{x}_1,\ldots,\underline{x}_m)$ nul is, dan kunnen we de $\underline{x}_j$ vervangen door $\underline{x}_\cdot$ zonder verlies van 'informatie', dat wil zeggen zonder verlies van variantie. De $\underline{x}_j$ heten in dat geval *perfect homogeen*.

Als maat voor homogeniteit kiezen we

$$\lambda(\underline{x}_1,\ldots,\underline{x}_m) = \frac{B(\underline{x}_1,\ldots,\underline{x}_m)}{T(\underline{x}_1,\ldots,\underline{x}_m)} .$$

Deze maat ligt altijd tussen nul en één, hij is gelijk aan nul als alle variantie binnen is, dat wil zeggen als de variabelen niets gemeenschappelijks hebben, en hij is gelijk aan één als alle variantie tussen is, dat wil zeggen in het geval van perfecte homogeniteit.

We kunnen een eenvoudiger formule voor λ geven door de covarianties $\sigma_{j\ell} = E(\underline{x}_j\underline{x}_\ell)$ te verzamelen in een matrix C, en de varianties σ_j^2 in een diagonale matrix D. We vinden dan

$$\lambda(\underline{x}_1,\ldots,\underline{x}_m) = \frac{1}{m}\,\frac{e'Ce}{e'De}$$

waarbij e een vector is met al zijn m elementen gelijk aan +1. Als de $\underline{x}_j$ ongecorreleerd zijn, dus als $\sigma_{j\ell} = 0$ voor alle $j \neq \ell$, dan geldt $\lambda = 1/m$. Omdat λ tussen nul en één ligt, zeggen sociale wetenschappers dikwijls dat $\underline{x}_\cdot$ een percenta-

ge λ van de variantie in de $\underline{x}_j$ verklaart.
Zoals in de inleiding al aangeduid werd, is dit woordgebruik misleidend: een gemiddelde is geen wetenschappelijke wet. Niettemin is het middelen van variabelen een veel voorkomende techniek om tot een simpele schaal te komen, denk bijvoorbeeld aan subtests, rapportcijfers en multiple choice tentamens.

1.1.2. Lineaire transformatie

De maat λ heeft enige voor de hand liggende nadelen. In de eerste plaats is λ afhankelijk van de schaling van de variabelen. In de sociale wetenschappen zijn de eenheden waarin we een variabele meten echter arbitrair, en zijn lineaire transformaties van de variabele even zo goede representaties van die variabele. De voor de hand liggende oplossing van dit probleem is om de gestandariseerde variabelen $\underline{z}_j = \underline{x}_j/\sigma_j$ te definiëren en vervolgens de homogeniteit van de $\underline{x}_j$ te definiëren als $\lambda(\underline{z}_1,\ldots,\underline{z}_m)$. Als we de correlatie matrix $R = D^{-1/2}CD^{-1/2}$ van de $\underline{x}_j$ gebruiken, dan geldt

$$\lambda(\underline{z}_1,\ldots,\underline{z}_m) = \frac{1}{m^2}\, e'Re.$$

We kunnen dit ook nog anders schrijven door een vector u met elementen σ_j^{-1} te gebruiken. Dit geeft

$$\lambda(\underline{z}_1,\ldots\underline{z}_m) = \frac{1}{m}\,\frac{u'Cu}{u'Du}.$$

In deze vorm werd de homogeniteitsmaat als voorgesteld door Galton. Er zijn bovendien enige voor de hand liggende maar hier niet nader uit te werken verbanden met betrouwbaarheidsschattingen zoals Cronbach's α. In eenvoudige pogingen om tot sets van variabelen met een hogere homogeniteit te komen verwijderen we soms variabelen, wat neerkomt op het kiezen van $u_j = 0$ in de laatste formule, en veranderen we soms variabelen van teken, wat neerkomt op kiezen van $u_j = -\sigma$. Dit zijn allemaal voorbeelden van lineaire transformaties van de $\underline{x}_j$ die het doel hebben de homogeniteit groter te maken. Het ligt nu voor de hand om je af te vragen hoe groot we de homogeniteit maximaal kunnen maken met behulp van lineaire transformaties.

Definieer

$$\lambda(\underline{x}_1,\ldots\underline{x}_m;a) = \frac{1}{m}\,\frac{a'Ca}{a'Da}.$$

De oorspronkelijke gedefinieerde λ is het speciale geval waarin $a = e$, en $\lambda(\underline{z}_1,\ldots.\underline{z}_m)$ is het speciale geval waarin $a = u$. In het algemeen geldt dat als $\underline{v}_j = a_j\underline{x}_j$, dan $\lambda(\underline{v}_1,\ldots\underline{v}_m) = \lambda(\underline{x}_1,\ldots\underline{x}_m;a)$. Uit bekende resultaten in de matrix algebra volgt dat de grootst mogelijke waarde van $\lambda(\underline{x}_1,\ldots\underline{x}_m;a)$ gelijk is aan de grootste eigenwaarde van R/m, we vinden deze waarde door a te kiezen als $D^{-1/2}v$,

waarbij v een bij die grootste eigenwaarde behorende eigenvector is. Als $\underline{x}_j$ ongecorreleerd zijn, geldt nog steeds $\lambda(\underline{x}_1,\ldots\underline{x}_m;a) = 1/m$, onafhankelijk van hoe we a kiezen.

1.1.3. Niet-lineaire transformatie

De volgende stap ligt nu voor de hand. Als $\psi_1,\ldots,\psi_m$ arbitraire transformaties zijn, dan kunnen we $\underline{t}_j = \psi_j(\underline{x}_j)$ definiëren, en de homogeniteit van $\underline{t}_j$ berekenen. De notatie die we gebruiken is $\lambda(\underline{x}_1,\ldots,\underline{x}_m;\psi_1,\ldots,\psi_m) = \lambda(\underline{t}_1,\ldots,\underline{t}_m)$. We kunnen nu ook proberen de ψ_j te kiezen en op zo'n manier dat λ zo groot mogelijk wordt. We moeten hierbij twee gevallen onderscheiden. In het meest eenvoudige geval zijn de $\underline{x}_j$ discreet, en nemen de k_j verschillende waarden aan. Zonder verlies van algemeenheid kunnen we veronderstellen dat deze waarden de integers $1,\ldots,k_j$ zijn. We kunnen nu de indicator functies $g_\mu(\underline{x}_j)$ definiëren als $g_\mu(\underline{x}_j) = 1$ als $\underline{x}_j = \mu$, en $g_\mu(\underline{x}_j) = 0$ als $\underline{x}_j \neq \mu$(waarbij $1 \leq \mu \leq k_j$). Iedere $\psi_j(\underline{x}_j)$ kan nu geschreven worden als een lineaire combinatie van de $g_\mu(\underline{x}_j)$, en het vinden van de coëfficiënten van de optimale lineaire combinaties komt weer neer op het oplossen van een eigenwaardeprobleem (nu van de orde $\Sigma k_j \times \Sigma k_j$). De indicator functies zijn in het dicrete geval een handige basis, maar we kunnen ook andere bases kiezen, zoals bijvoorbeeld discrete orthogonale polynomen.

Het tweede geval dat we moeten onderscheiden is datgene waarin de $\underline{x}_j$ continue variabelen zijn. Het is nu niet langer mogelijk een eindig aantal basisfuncties te vinden, waarvoor geldt dat iedere $\psi_j(\underline{x}_j)$ een lineaire combinatie van die basisfuncties is. We hebben oneindig veel van deze basisfuncties nodig, en het eigenwaarde probleem wordt dus ook oneindig groot. Zelfs met de tegenwoordige generatie computers zijn oneindig grote eigenwaarde problemen lastig op te lossen en wat we dus willen doen is een eindig aantal basisfuncties vinden waarvan we verwachten dat ze de $\psi_j(\underline{x}_j)$ goed benaderen. Vroeger gebruikte men hier bij voorkeur orthogonale polynomen voor, tegenwoordig zijn polynoomsplines het meest populair. Door de graad van de polynoom of de spline te verhogen, kan men de benadering beter maken.

Als we de overeenkomst tussen continue en discrete variabelen willen benadrukken, dan is het handig geen polynomen of splines te gebruiken maar stapfuncties. Gebruik van een eindige basis van stapfuncties komt overeen met het discretiseren van de continue variabelen in intervallen, en met het gebruiken van de indicatoren van die intervallen als basis. Een fijnere discretisering betekent dan in het algemeen een betere benadering. Bij secundaire analyse is men dikwijls gedwongen stapfuncties te gebruiken, omdat de gegevens in een eerdere fase al gediscreti-

seerd zijn. Dit is ook bij 'Van Jaar tot Jaar' gegevens het geval.
We bespreken kort een belangrijk speciaal geval. Stel de $\underline{x}_j$ zijn multinormaal. Als we als benaderende deelruimte polynomen of splines gebruiken, dan vinden we altijd als uitkomst $\psi_j(\underline{x}_j) = v_j\underline{x}_j$, met v_j het j-de element van de eigenvector behorende bij de grootste eigenwaarde van de correlatiematrix R. De maximale waarde van $m\lambda$ is ook gelijk aan die grootste eigenwaarde. Met andere woorden: voor multinormaal verdeelde grootheden heeft niet-lineaire optimalisatie van homogeniteit dezelfde resultaten als lineaire optimalisatie. Als we in plaats van polynomen of splines echter stapfuncties kiezen, dan vinden we de stapfunctie die $v_j\underline{x}_j$ het beste benadert als beste functie. De maximale homogeniteit is wat kleiner dan de grootste eigenwaarde. Als er echter transformaties ϕ_j bestaan zodanig dat $\underline{s}_j = \phi_j(\underline{x}_j)$ gezamenlijk multinormaal zijn, dan is de optimale functie $\psi_j(\underline{x}_j) = v_j\phi_j(\underline{x}_j)$. Als de verdelingen afwijken van normaliteit, bijvoorbeeld doordat ze scheef zijn, zal optimalisatie van homogeniteit op hetzelfde neerkomen als transformatie naar normaliteit. In sommige gevallen is het best mogelijk dat stapfuncties betere benaderingen geven dan polynomen of splines van een lage orde, bijvoorbeeld als we discretiseren en de volgorde van de categorieën stiekum veranderen.

1.1.4. Details

Tot nu toe hebben we onze technieken steeds gedefinieerd in termen van de variabelen $\underline{x}_j$, dat wil zeggen in termen van de populatie, en niet in termen van observaties op die variabelen, dat wil zeggen in termen van een steekproef. Als we een steekproef hebben , zoals in dit 'Van Jaar tot Jaar' geval, dan vullen we voor alle theoretische verwachte waarden zoals covariantie en celwaarschijnlijkheden gewoon hun geobserveerde schatters in, en berekenen op grond daarvan de maximale homogeniteit. Omdat de geobserveerde verwachte waarden bij een groot aantal observaties naar de theoretische waarden zullen convergeren, zal ook de maximale homogeniteit naar zijn theoretische waarde convergeren.

In de tweede plaats hebben we ons in deze opmerkingen steeds beperkt tot het vinden van één enkele transformatie van de variabelen. In vele toepassingen is men geinteresseerd in meerdere transformaties, die overeen komen met andere stationaire waarden van de homogeniteitscoëfficiënt, dat wil zeggen met andere eigenwaarden van de product moment matrixen behalve de grootste. Dit leidt direct tot lineaire en niet-lineaire vormen van principale componenten analyse. In onze analyse van de 'Van Jaar tot Jaar' gegevens beperken we ons over het algemeen tot de eerste dimensie, in de eerste plaats omdat dat het eenvoudigste is, in de tweede plaats omdat die eerste dimensie nogal dominant is (zie verderop).

1.2. Resultaten

1.2.1. Ongewogen en gewogen summatie

In de appendix staat een lijst van 25 variabelen die we in onze analyses gebruiken. Deze lijst komt overeen met die in Dronkers(1979). We verwijzen ook naar Dronkers voor nadere informatie over de categorieën van deze variabelen. In de appendix van Dronkers vinden we een correlatiematrix, berekend door de categorienummers als scores te gebruiken. We hebben zelf ook zo'n correlatiematrix berekend, hij staat in tabel 3a. Voor deze berekening hebben we SPSS gebruikt, met paarsgewijze weglating in het geval van ontbrekende gegevens. Er zijn wat verschillen tussen onze matrix en die van Dronkers, met name voor variabelen 9:DWO en 22:EXT. De verschillen voor 22:EXT begrijpen we. Het aantal extra-curriculaire activiteiten varieert van nul tot vijf, Dronkers heeft nul gecodeerd als nul en wij hebben nul gecodeerd als zes. Dit laatste om te laten zien dat dit soort ernstige vergissingen in de codering door onze niet-lineaire technieken 'gecorrigeerd' wordt. De verschillen voor 9:DWO begrijpen we niet, maar dat wordt nog uitgezocht. Overigens is in tabel 3a (en 3b) negativiteit van een correlatie aangegeven door onderstrepen, onderstrepen van de naam van een variabele geeft aan dat we deze variabele van teken hebben veranderd, om zoveel mogelijk positieve correlaties te krijgen.

Op basis van 3a hebben we twee schalen uitgerekend. De eerste door simpele summatie van gestandariseerde, en eventueel dus van teken veranderde, variabelen. De homogeniteit is .1484, de correlaties van de afzonderlijke variabelen met de afgeleide schaal staan in kolom 1 van tabel 1. De tweede schaal is berekend met optimale gewogen summatie, de homogeniteit is .1989, de correlaties met de variabelen staan in kolom 2 van tabel 1. Als we dezelfde berekeningen toepassen op de correlatiematrix van Dronkers vinden we een geoptimaliseerde homogeniteit van .2154. De correlaties van de Dronkers-schaal met de variabelen wijken hoogstens .03 van de door ons gegeven correlaties af, behalve bij 20:AOS, 21:LLS en 22:EXT Daar vinden we uit de Dronkers-matrix .43, .34, en .54. Voor een belangrijk deel is dit verschil ongetwijfeld het gevolg van de 'foutieve' codering van 22:EXT di wij gebruikt hebben. Zoals uit tabel 1 blijkt is het effect van wegen dat de hog correlaties omhoog gaan, terwijl de lage dikwijls nog iets lager worden. De schaal correleert zeer sterk met 14:ADV. 18:PRE, 19:TON, en 25:EIN.

1.2.2. Niet-lineaire transformatie: correlaties

Met het programma HOMALS hebben we de schaal na optimale transformatie van de variabelen uitgerekend. De homogeniteit is .2416, merk op dat dit maar een relatief kleine winst is als we nagaan dat het aantal vrij te kiezen gewichten van

25 (het aantal variabelen) naar 139 (het totaal aantal categorieën) gaat. We hebben ook de correlatie matrix na transformering uitgerekend, die staat in tabel 3b. Deze matrix kan in verdere analyses gebruikt worden, we presenteren hem als een betere schatting van de correlatie matrix als de schatting in de appendix van Dronkers. Bij het vergelijken van 3a en 3b valt onder andere op, dat 3b één-dimensionaler is dan 3a, in de zin dat de grootste eigenwaarde van 3b groter is dan de grootste van 3a, terwijl de overige eigenwaarden bij 3a dichter bij elkaar liggen dan bij 3b. De tweede principale component bij 3a ('homogeniteit' gelijk aan .0780) correleert het meeste met 20:AOS, 21:LLS, en 22:EXT, de tweede principale component bij 3b ('homogeniteit gelijk aan .0632) correleert (grofweg) positief met de eerste 12 variabelen en negatief met de laatste 13, dat wil zeggen de tweede principale component van 3b contrasteert gezins- en L.O.-variabelen met variabelen die betrekking hebben op het secundair onderwijs.

Als we op basis van 3b een optimale gewogen schaal uitrekenen, moeten we theoretisch dezelfde homogeniteit vinden als met HOMALS. Omdat ontbrekende gegevens anders behandeld worden is er een klein verschil: tabel 3b geeft een optimale homogeniteit van .2386. De correlaties van de variabelen met de optimale schaal vinden we in kolom 3 van tabel 1, de correlaties zoals HOMALS ze vind in kolom 1 van tabel 2. We zien dat niet-lineaire transformatie de correlaties in het algemeen hoger maakt, ook de lage, en dat de meeste winst geboekt wordt op 20:AOS, 21:LLS, en 22:EXT. De onjuiste codering wordt gecorrigeerd. Opgemerkt moet ook worden dat het kwadraat van de in kolom 3 van tabel 1 gegeven correlaties gebruikt kan worden als schatting van de variantie van de variabele, tezamen met tabel 3b, in eventuele verdere multivariate analyse op normaliteits assumpties.

1.2.3. Niet-lineaire transformatie: de transformaties

In figuur 4.1-4.25 hebben we de door HOMALS gevonden transformaties geplot. Op de X-as staat reeds het categorienummer, dat wil zeggen de in eerdere analyses gebruikte scores, en op de Y-as de door HOMALS gevonden scores. Een gedetailleerde analyse van de transformaties zou te veel plaats in dit artikel innemen, we hebben geprobeerd de resultaten op zo'n manier te presenteren dat de geinteresseerde lezer zonder al te veel moeite zelf de transformaties in detail kan bekijken. Een paar algemene opmerkingen lijken op zijn plaats. In de eerste plaats corrigeert HOMALS dikwijls voor scheefheid van de marginalen, we vinden dan monotone transformaties die convex of concaaf zijn, afhankelijk van de aard van de scheefheid. Een duidelijk voorbeeld is 07:ASO. In de tweede plaats corrigeert HOMALS voor te zware staarten. U-vormige verdelingen komen in de 'Van Jaar tot Jaar' gegevens niet voor, maar 05:URB bijvoorbeeld is bijna rechthoekig, wat tot

gevolg heeft dat de transformatie de beide uiteinden relatief ver weg zet. De algemene vorm van de transformatie kan dus dikwijls begrepen worden uit de marginalen, tezamen genomen met het idee dat HOMALS naar multinormaliteit tracht te transformeren. De meest regelmatige transformaties van numerieke variabelen vinden we dus bij afgeronde stanine-scores, zoals hier 16:BIL, 17:BIM, en 18:PRE.

Interessanter is welke transformaties we vinden voor typische nominale variabelen zoals 01:BVA, 02:OPV, 03:OPM, 14:ADV, 19:TON, en 25:EIN. De transformaties die we vinden zijn over het algemeen monotoon met de categorienummers, en wijken op het eerste gezicht weinig van lineair af. Dit bevestigt globaal de aanname van lineairiteit waarop eerdere analyses gebaseerd waren. Bij een meer gedetailleerde analyse zijn er echter een groot aantal belangwekkende afwijkingen van lineariteit te vinden. We kunnen ze hier onmogelijk allemaal bespreken, maar we noemen enige belangwekkende voorbeelden. In 01:BVA lijkt het juist om categorie 2 (geschoolde handarbeid) en categorie 5 (boeren en tuinders) tezamen te nemen, terwijl het tevens zinvol lijkt categorie 3 (uitvoerende hoofdarbeid) en categorie 4 (zelfstandige middenstand) van plaats te verwisselen.
In 02:OPV en 03:OPM lijkt het beter om categorie 2 (vakcursussen) en categorie 3 (LBO) te verwisselen of samen te nemen, om categorie 4 (ULO/MULO) en categorie 5 (MBO) te verwisselen of samen te nemen, en misschien is het zelfs beter om categorie 6 (VHMO) en categorie 7 (HBO/WO) samen te nemen. Het spreekt vanzelf dat dit soort aanbevelingen overigens nader onderzocht moeten worden, met andere technieken en/of ander materiaal, en dat de voorgestelde rangordening van beroepsprestige of onderwijsprestige in principe alleen relevant is voor onderzoek naar schoolsucces. Hetzelfde geldt voor de omcoderingen van 25:EIN die uit onze HOMALS analyse volgen. De aanbevelingen hier zijn: neem van de indeling zoals gegeven in Collaris en Kropman (1979, tabel 4B1, pag. 334) de categorieën 4, 5 en 6 samen, neem 7,8 samen, en neem 9, 10 samen. Dit leidt nauwelijks tot informatieverlies, en levert volgens HOMALS bij benadering een intervalschaal op. Hoewel het voor deze gegevens niet nodig is, lijkt het bovendien voor de hand te liggen ook 2 en 3 samen te nemen.

1.2.4. Niet-lineaire transformaties: jongens en meisjes

We hebben ook aparte HOMALS analyses gedaan voor jongens en meisjes, om een nadere vergelijking met Dronkers en Jungbluth (1979) mogelijk te maken. De homogeniteit is .2470 voor meisjes en .2487 voor jongens, een weinig indrukwekkend verschil. De correlaties met de schalen staan voor meisjes in kolom 2 en voor jongens in kolom 3 van tabel 2. Hoewel de overeenkomst tussen de twee kolommen groot is, zijn er ook duidelijke verschillen. In het algemeen correleren de gezins- en

LO-variabelen hoger met de schaal voor meisjes, terwijl de variabelen die meer direct met het secundair onderwijs samenhangen hoger correleren met de schaal voor jongens. Daarnaast zijn er een aantal detail verschillen. Dat 16:BIM en 17:BIL aanzienlijk hoger correleren bij jongens kan aan het feit liggen, dat deze tests voor jongens en meisjes apart genormeerd zijn. Ons resultaat kan dan opgevat worden als bewijsmateriaal dat, althans voor de doeleinden waarvoor wij de test gebruiken, deze aparte normering niet erg succesvol is. Anderzijds kan het natuurlijk ook betekenen dat beroepsinteresse (geoperationaliseerd als BIM of BIL) voor jongens sterker samenhangt met de rest van de schoolloopbaan dan voor meisjes. Dat 21:LLS en 22:EXT voor meisjes aanzienlijk hoger correleren met de schaal kan komen door specifieke eigenschappen van het LHNO: veel kleine scholen met weinig extra-curriculaire activiteiten (of Collaris en Kropman, 1978 tabel 7.63, pag. 252). Dat 20:AOS hoger correleert bij jongens is een VHMO effect (of Collaris en Kropman, 1978, tabel 5.70, blz. 130-132, blz. 113-114, en tabel 5.37 en 5.38). De hypothesen van Dronkers en Jungbluth (1979) zeggen niets over differentiële effecten van of op kenmerken van het secundair onderwijs, terwijl hun hypothese 5.6.k voorspelt dat BIL en BIM voor meisjes belangrijker zullen zijn dan voor jongens. We zullen de diverse hypotheses echter niet in details met onze resultaten vergelijken, de geinteresseerde lezer kan dit zelf doen, en we komen bovendien in een later hoofdstuk nog op dit soort vergelijkingen terug. De belangwekkende algemene trend, dat ouders en gezinskenmerken voor meisjes een grotere rol spelen, wordt ook door Dronkers en Jungbluth voorspelt. Uit vergelijking van kolom 2 en 3 van tabel 2 blijkt echter dat de verschillen, hoewel consistent, niet zeer groot zijn.

Deze laatste constatering wordt bevestigd door de plots in figuur 5.1-5.25, waar we de optimale transformaties tegen elkaar uitgezet hebben (X-as mannen, Y-as vrouwen). In de meeste gevallen liggen de punten op of om de lijn X = Y, in andere gevallen zijn er afwijkingen, maar die zijn dikwijls nogal onregelmatig, en treden dan ook op in de 'onbelangrijke' variabelen. Voor 16:BIM en 17:BIL, en ook voor 20:AOS, 21:LLS en 22:EXT, vinden we de verschillen die we ook al uit de correlaties met de schalen afgeleid hadden, maar in weinig gevallen lijkt er sprake te zijn van interessante afwijkingen van lineariteit. Uit het oogpunt van onderzoek naar systematische verschillen tussen jongens en meisjes in de schoolloopbaan is dit jammer, maar uit het oogpunt van vertrouwen in onze transformaties en bevestiging van de aanname van lineariteit is het dat natuurlijk niet.

2. Voorspelbaarheid

2.1. Theorie

2.1.1. Definitie

Vanwege de overeenkomsten met de definitie van homogeniteit gaan we in dit hoofdstuk wat vlugger. Stel $\underline{x}_0,\underline{x}_1,..,\underline{x}_m$ zijn gecentreerde stochastische grootheden. In dit hoofdstuk gaat het erom in hoeverre we $\underline{x}_0$ uit $\underline{x}_1,\ldots,\underline{x}_m$ kunnen *voorspellen* waarbij we weer moeten beseffen dat ons woordgebruik misleidend is. 'Voorspellen' is, evenals 'verklaren', een Galton-Pearson term, die een tendentieuze methodologische interpretatie aan een eenvoudige meetkundige operatie geeft. We veronderstellen in dit hoofdstuk dat $\sigma_j^2 = E(\underline{x}_j^2) = 1$ voor alle j=0,...,m. Bij de voorspelling maken we gebruik van *gewichten* $b_1,\ldots,b_m$, de ongewogen situatie komt overeen met $b_j = 1$ voor alle j. Om te meten hoe goed onze 'voorspelling' lukt voor een bepaalde keuze van gewichten definiëren we

$$\omega(\underline{x}_0;\underline{x}_1,\ldots,\underline{x}_m;b_1,\ldots,b_m) = E(\underline{x}_0 - \sum_{j=1}^{m} b_j\underline{x}_j)^2.$$

Als we de matrix R met elementen $E(\underline{x}_j\underline{x}_\ell)$, en de vector r met elementen $E(\underline{x}_0\underline{x}_j)$ definiëren, dan geldt

$$\omega(\underline{x}_0;\underline{x}_1,\ldots,\underline{x}_m;b_1,\ldots,b_m) = 1 - 2b'r + b'Rb.$$

Veronderstel dat R inverteerbaar is, en definieer $\hat{b} = R^{-1}r$. We vinden dan de identiteit

$$1 - 2b'r + b'Rb = 1 - r'R^{-1}r + (b - \hat{b})'R(b - \hat{b}).$$

Dat wil zeggen als

$$\omega(\underline{x}_0;\underline{x}_1,\ldots,\underline{x}_m) = \min_b \omega(\underline{x}_0;\underline{x}_1,\ldots,\underline{x}_m;b_1,\ldots,b_m),$$

dan

$$\omega(\underline{x}_0;\underline{x}_1,\ldots,\underline{x}_m) = 1 - r'R^{-1}r.$$

Merk op dat

$$E(\underline{x}_0, \sum_{j=1}^{m} \hat{b}_j\underline{x}_j) = E\{(\sum_{j=1}^{m} \hat{b}_j\underline{x}_j)^2\} = r'R^{-1}r,$$

zodat de correlatie tussen $\underline{x}_{0_{1/2}}$ en de optimale lineaire 'voorspelling' van $\underline{x}_0$ gelijk is aan $(r'R^{-1})$. Dit noemt men de *multipele correlatie*.

2.1.2. Niet-lineaire transformatie

Evenals in 1.1.3. zijn $\psi_0,\psi_1,\ldots,\psi_m$ arbitraire transformaties. We veronderstellen, zonder verlies van algemeenheid , dat $E\{\psi_j(\underline{x}_j)\} = 0$ en $E\{\psi_j(\underline{x}_j)^2\} = 1$ voor alle j: Definieer nu

$$\omega(\underline{x}_0;\underline{x}_1,\ldots,\underline{x}_m;b_1,\ldots,b_m;\psi_0,\psi_1,\ldots,\psi_m) = E\{\psi_0(\underline{x}_0) - \sum_{j=1}^{m} b_j\psi_j(\underline{x}_j)\}^2.$$

Het probleem is nu om deze maat te minimaliseren over $b_1,\ldots,b_m$ *en* over $\psi_0, \psi_1,\ldots,\psi_m$. Voor discrete variabelen met een eindig aantal mogelijke waarden levert dit geen enkel probleem op. We kiezen een basis van indicatoren, en lossen een canonisch correlatieprobleem op. Als sommige van de $\underline{x}_j$ continue variabelen zijn, moeten we meer bewust een benaderende basis uitkiezen, hetzij door te discretiseren, hetzij door polynomen of splines te gebruiken.

Evenals bij het niet-lineair maximaliseren van homogeniteit is het geval waarin alle $\underline{x}_j$ multinormaal verdeeld zijn van speciaal belang. We vinden hier weer dat optimale niet-lineaire en optimale lineaire voorspelbaarheid hetzelfde zijn, dat wil zeggen dat de optimale niet-lineaire transformaties lineair blijken te zijn. Overeenkomstige resultaten gelden als de $\underline{x}_j$ niet multinormaal zijn, maar door aparte transformaties op ieder van de variabelen naar multinormaliteit getransformeerd kunnen worden. Het is overigens niet zo dat de multinormale verdeling het enige geval is, waarin we de uitkomst theoretisch kunnen afleiden. Soortgelijke resultaten kunnen afgeleid worden voor de Guttman-schaal en voor de simplex. Het is duidelijk dat onze niet-lineaire technieken nooit helemaal los komen van de multinormaalverdeling als model, om de eenvoudige reden we kleinste kwadraten blijven gebruiken en alleen maar op (product) momenten van de eerste en tweede orde letten.

2.1.3. Interacties

We raken wat verder af van de multinormaalverdeling, en we komen wat dichter bij log-lineaire analyse en verwante technieken, als we interacties in ons model opnemen. We gaan nu uit van de meer algemene maat

$$\omega(\underline{x}_0;\underline{x}_1,\ldots,\underline{x}_m;\psi,\phi) = E\{\psi(\underline{x}_0) - \phi(\underline{x}_1,\ldots,\underline{x}_m)\}^2.$$

Niet alleen de lineariteit (van de transformaties), maar ook de additiviteit (van de combinatieregel) is nu uit de formules verdwenen. De enige lineariteit die overblijft is die van de verliesfunctie. Desondanks is het nog steeds zo dat voor multinormaal verdeelde grootheden deze meest algemene vorm van niet-lineaire voorspelbaarheid nog steeds op hetzelfde neerkomt als de meest eenvoudige lineaire vorm.

Technisch levert het minimaliseren van de meer algemene maat weinig nieuwe problemen op. We moeten nog steeds een of andere eindige basis kiezen voor de ruimte van toegelaten functies. In het discrete geval is die ruimte nog eindig-dimen-

sionaal, de dimensionaliteit is gelijk aan het aantal cellen van de bij $\underline{x}_1,\ldots,\underline{x}_m$ behorende m-dimensionale kruistabel (contingency table). De dimensionaliteit kan dus zeer groot worden, zodat we zelfs in het discrete geval moeten denken aan benaderende deelruimtes.

We hadden op precies dezelfde manier onze niet-lineaire homogeniteitsanalyse kunnen generaliseren, om op die manier de gegevens te onderzoeken op aan- of afwezigheid van hogere orde interacties (alweer: zoals bij log-lineaire analyse). Het verschil met log-lineaire analyse (zoals op de 'Van Jaar tot Jaar' gegevens toegepast door Dessens en Janssen (1979) is, dat wij globale statistische informatie zoals tests van hypotheses en schattingen van parameters over het algemeen niet erg informatief vinden. Wij zijn geïnteresseerd in meer gedetailleerde informatie over de multivariate verdelingen, de benadering is in zekere zin te vergelijken met het gebruik van contrasten in de variantieanalyse, waarbij wij onze contrasten ook nog 'optimaal' kiezen.

2.2. Resultaten

2.2.1. Globale resultaten

We gebruiken in onze eerste reeks analyses variabele 01 t/m 24 om 25:EIN te voorspellen. Dit kan in de eerste plaats lineair gebeuren. In tabel 6 vinden we de gewichten $b = R^{-1}r$ en de multipele correlatie $r'R^{-1}r$. We hebben zowel onze correlatiematrix uit tabel 3a als de matrix uit de appendix van Dronkers (1979) gebruikt. Onze laatste kolom in tabel 6 lijkt daarom sprekend op de laatste kolom in tabel 1 van Dronkers (1979). De multipele correlaties zijn over het algemeen nogal hoog, met name als we ze vergelijken met multipele correlaties in de Amerikaanse literatuur die ook zoiets als 25:EIN willen voorspellen, maar als we ons realiseren dat de predictoren met het hoogste gewicht 14:ADV, 19:TON, en 24:INS zijn, dan blijkt de hoge correlatie nogal triviaal. Het bereikte eindniveau wordt in de eerste plaats 'bepaald' door de eerste schoolkeuze na het l.o., en hangt bovendien sterk samen met het advies van de onderwijzer, de score op de Nederlandse Onderwijs Differentiatie Test, en de medewerking van de ouders. Als we dit wat anders rangschikken, dan komen we tot het 'causale model': prestatiescore, advies van de onderwijzer bepalen de keuze van het voortgezet onderwijs (dat wil zeggen als de ouders zich erbij neerleggen), de keuze van het voortgezet onderwijs bepaalt grotendeels het behaalde eindniveau op het voortgezet onderwijs. Zoals Dessens en Janssen al eerder opmerkten, is er met dit causale model op beleidsniveau uitermate weinig te beginnen.

We hebben dezelfde multipele regressies nogmaals berekend met het programma CANALS, wat tegelijkertijd de variabelen niet-lineair transformeert. De regres-

siegewichten en de gekwadrateerde multipele correlatie staan in kolom 2 van tabel 7, de transformaties staan in figuur 9.1 t/m 9.25. De correlaties van de getransformeerde variabelen 1 t/m 24 met de getransformeerde variabele 25 staan in kolom 2 van tabel 8. Het is duidelijk dat de multipele correlatie onbehoorlijk hoog wordt. We zien dat 20:AOS, 21:LLS, en 22:EXT hiervan een belangrijke oorzaak zijn. De transformaties van met name 25:EIN en 19:TON laten zien, dat de regressie voornamelijk bepaald wordt door de groep van 60 tot 70 mensen, die geen vervolgonderwijs volgt (en daardoor ook niet mogen genieten van extra-curriculaire activiteiten). De regressie-analyse verandert hierdoor min of meer automatisch in een discriminant- analyse, waarin de kleine groep van vroege schoolverlaters zo goed mogelijk onderscheiden wordt van de grote groep van doorleerders.

Zowel de lineaire als de niet--lineaire regressie analyse leveren min of meer triviale oplossingen op. Omdat we de oorzaak van deze degeneratie begrijpen, kunnen we ze door manipulaties erger of minder erg maken. Als we bijvoorbeeld alle mensen die op 19:TON of 25:EIN in categorie 1 (geen verder onderwijs na l.o.) scoren uit de analyse verwijderen, dan komt CANALS tot een gekwadrateerde multipele correlatie van slechts .82. Een andere voor de hand liggende truc is om 25:EIN te hercoderen, gebruik makend van de HOMALS scores, en uitsluitend lineair te transformeren (de 24 predictoren mogen wel niet-lineair). De gewichten staan in tabel 7, kolom3, en de correlaties in tabel 8, kolom 3. De transformaties staan in figuur 10.1 t/m 10.25. We zien dat de transformaties aanzienlijk minder springerig worden, de analyse gaat meer lijken op de lineaire analyses. Een laatste mogelijkheid is om de HOMALS correlaties in tabel 3b in de multipele regressie te gebruiken. De gewichten en correlaties staan in kolom 1 van tabel 7 en 8, de oplossing lijkt nog iets meer op de lineaire oplossing in tabel 6.

Dat de volledig niet-lineaire CANALS min of meer uit de rails loopt heeft aanwijsbare oorzaken in de gegevens, zoals we gezien hebben. Maar de moeilijkheden ontstaan *ook* door de aard van multipele regressieproblemen. In homogeniteitsanalyse proberen we ieder van de variabelen uit iedere andere variabele te voorspellen. Er is daardoor een soort ingebouwde kruisvalidatie aan de gang, die de gevonden transformaties stabiel maakt. In multipele regressie voorspellen we maar één enkele variabele uit alle andere, kleine afwijkingen in het criterium kunnen daardoor grote gevolgen hebben. Als we dit combineren met het gebruik van een kleinste-kwadraten verliesfunctie, zodat we voornamelijk op de grote fouten in de voorspelling letten, dan is het verklaarbaar waarom CANALS eerder op speci-

fieke kleine afwijkingen let als HOMALS. In de 'Van Jaar tot Jaar' gegevens vinden we met HOMALS over het algemeen monotoniciteit (en dikwijls zelfs lineariteit) van de scores. Dit wijst op weinig afwijkingen van normaliteit. Met CANALS daarentegen vinden we een duidelijke groep van uitbijters, die afwijken van het 'normale' patroon. De groep bestaat slechts uit 4% van de steekproef, maar CANALS gebruikt die groep op een (triviaal) multipele correlatie dicht bij één te construeren. Dit is een speciaal geval van een bekend psychometrisch dilemma. De psychometrici hebben immers al sinds Spearman geprobeerd hun correlaties zo hoog mogelijk te maken, omdat het evangelie van Pearson causaliteit identificeert met perfecte correlatie, en wetenschappelijk succes met verklaarde variantie. De moderne computer-programma's (met name de niet-lineaire en de zogenaamde niet-metrische) leiden dikwijls rechtstreeks naar de psychometrische hemel met perfecte correlaties, aanzienlijk sneller en efficiënter als de meer bejaarde trucs, zoals modderen met de diagonaal van de correlatiematrixs (factor analyse) of modderen met de hele correlatiematrixs (correctie voor meetonbetrouwbaarheid). Ongelukkigerwijs blijken de perfecte correlaties in de psychometrische hemel te horen bij triviale oplossingen uit de psychometrische hel. Vandaar het dilemma.

2.2.2. Stapsgewijs

We verdelen nu de variabelen in een aantal subsets, en we doen regressie analyses met een toenemend aantal variabelen. Onze indeling komt voor een groot deel overeen met die van Dronkers (1979), maar we hebben wat kleine verschuivingen aangebracht op grond van onze eerdere analyses. Het belangrijkste is natuurlijk, dat de drie laatste groepen zijn (ADV, KGS, PRE), (TON) en (AOS, LLS, EXT). In tabel 11a staan de regressiegewichten berekend op basis van de correlaties van Dronkers, in tabel 11b de regressiegewichten berekend uit de HOMALS-correlaties in 3b. De patronen in de twee tabellen lijken erg op elkaar. We zien dat de regressiecoëfficiënten sterk afhangen van de predictoren die we gebruiken, en dat met name analyses één tot en met vijf (respectievelijk 5, 9, 14, 15, 17 predictoren) nogal dramatisch verschillen van analyses zes, zeven en acht (met 20, 21, 24 predictoren). De verschillen binnen één tot en met vijf zijn relatief klein, en ook de verschillen tussen tabel 11a en 11b zijn voor analyses één tot en met vijf niet al te groot. Pas als we de 'latere' variabelen toevoegen, treden de grootste instabiliteiten op. De regressiecoëfficiënten van de eerste 17 variabelen worden platgeslagen. Er zijn een groot aantal psychometrici, die uitsluitend naar de laatste kolom van 11a en 11b kijken, en daaruit concluderen dat sociaal-economisch milieu geen enkel effect heeft op schoolprestatie. Als het al mogelijk zou zijn dit soort conclusies te trekken op basis van regressie-

analyse, dan laten onze eerdere analyses zien dat deze conclusie onzinnig is. Zoals eerder betoogd in Jaspars en De Leeuw (1979) hebben lineaire modellen de eigenschap dat de conclusies op basis van submodellen dikwijls verschillen van de conclusies op basis van het volledige model. Als men een bepaalde conclusie wil goedpraten met behulp van empirsiche gegevens, dan is er dikwijls wel een submodel te vinden dat op de vooroordelen van de onderzoekers past. Vanwege de waanzinnige hoeveelheid gegevens die door psychometrici in de afgelopen 100 jaar verzameld zijn, is het dikwijls ook nog mogelijk de juiste gegevens bij het juiste model te kiezen. Dit wordt duidelijk geïllustreerd in het werk van Jensen en (huizenhoog boven alles uittorenend) Eysenck. Het enige precedent in de geschiedenis van de wetenschap zijn de 'onderzoekers' die, dikwijls op basis van 'secundaire analyse', grote wetenschappelijke en occulte waarheden ontdekten in de afmetingen van de pyramiden.

Het is bekend dat schattingen van regressiecoëfficiënten instabiel zijn als er multicollineariteit is, en dat de gebruikelijke kleinste kwadraten schattingen statistisch verbeterd kunnen worden als er veel predictoren zijn. In het geval van de 'Van Jaar tot Jaar' gegevens lijkt het alsof alle interessante informatie gevonden is als we de eerste 17 predictoren gebruiken, daarna worden de analyses zowel oninteressant als instabiel. We illustreren dit ook nog met een aantal CANALS analyses, gepresenteerd in tabel 12. Omdat we in niet-lineaire analyses alle categorieën als predictoren gebruiken, gaan de opmerkingen over grote aantallen predictoren en daarbij behorende multicollineariteit hier eens te meer op. Het is daarom prettig, dat de niet-lineaire analyse met 17 predictoren redelijk veel op de lineaire analyses lijkt. Ook de transformatie van de afhankelijke variabele 25:EIN loopt hier nog vrijwel zoals in HOMALS.

2.2.3. Jongens en meisjes

De volledige analyse met 24 predictoren hebben we nogmaals apart gedaan voor meisjes en jongens. De resultaten staan in tabel 13a en tabel 13b (respectievelijk regressiegewichten en correlaties). We zien in deze tabellen dat dezelfde degeneraties optreden als in de analyse over de totale steekproef, maar dat deze degeneraties de regressiegewichten aanzienlijk erger treffen dan de correlaties. Bij de correlaties vinden we dezelfde differentiële effecten als bij HOMALS: in het algemeen correleren de 'vroege' variabelen bij vrouwen hoger met de schaal, terwijl de 'latere' variabelen hoger correleren bij mannen. Ook vinden we, zoals bij HOMALS, dat OPM een betere predictor is bij jongens, terwijl OPV een betere predictor is bij meisjes. De transformaties zijn geplot in figuur 14.1-14.25. Er zijn weinig duidelijke systematische effecten te ontdekken, als we de plots vergelijken

met die uit HOMALS in figuur 5.1-5.25 dan valt de grote mate van instabiliteit duidelijk op.

2.2.4. Interacties

We hebben gezien dat bij niet-lineaire voorspelling, door toename van het aantal predictoren, de instabiliteit van de regressiegewichten een ernstig probleem kan worden. Het probleem wordt alleen maar ernstiger als we de interacties toelaten, omdat dit in feite neerkomt op een verdere uitbreiding van het aantal predictoren. Over het algemeen zullen we voor een interactie-analyse dus maar een relatief klein aantal variabelen kunnen gebruiken, te meer omdat bij een groot aantal ook het aantal observaties en de interpretabiliteit van hogere-orde interacties een rol gaan spelen (zoals bij log-lineaire analyse). We zoeken wat interessante kruistabellen uit, en onderzoeken ze op interactie. Voor de 'Van Jaar tot Jaar' gegevens illustreren we dit soort analyses met een voorbeeld waarin BVA en PRE gebruikt worden om ADV te voorspellen. Voor de gelegenheid zijn de drie variabelen ieder teruggebracht tot drie niveaus: hoog - midden - laag. We ontlenen de tabel aan een eerdere versie van het artikel van Dessens en Janssen (1979). De data staan in tabel 15a, orthogonale polynomen op de drie variabelen staan in 15b, de orthogonale polynomen op BVA en PRE worden gebruikt om een basis voor alle functies op BVA x PRE te construeren, deze staat in 15c. Iedere functie in de basis is het product van een BVA-polynoom met een PRE-polynoom, waarbij we met product uitwendig product bedoelen: 3-vector met 3-vector geeft 3 x 3 matrix, geeft 9-vector. We doen vier verschillende analyses. In de eerste gebruiken we vector 2 t/m 8 in de basis 15c, en een polynoom van de tweede graad op ADV. Dit is een volledige niet-lineaire analyse met interactie term. Als we de interactietermen weglaten krijgen we een niet-lineaire additieve analyse. Dit is analyse twee, waarvoor we vectoren 2, 3, 4, 7 uit de basis 15c gebruiken. In de derde analyse is lineair en additief in de predictoren, niet-lineair in ADV, we gebruiken vectoren 2 en 4 in de basis. Voor de vierde analyse gebruiken we ook 2 en 4, maar voor ADV gebruiken we nu alleen de polynoom van de eerste graad. Onze analyse is nu volledig lineair. De geschatte transformaties voor de vier analyses, tezamen met de gekwadrateerde multipele correlaties, staan in 15d. Aan de multipele correlaties is al te zien dat het effect van niet-lineariteit en niet-additiviteit gering is, wat geen verwondering hoeft te wekken omdat deze drie variabelen bij de HOMALS analyse zich in dit opzicht ook al keurig gedroegen. De transformaties van BVA x PRE worden in tabel 15e nogmaals weergegeven, maar nu ontbonden in additieve en eventuele interactie componenten. We zien dat BVA wel degelijk een effect op ADV heeft, maar dat de interactie te verwaarlozen is. Zo-

als Dessens en Janssen ook al opmerkten, is er een mogelijke interactie tussen milieu en *hoge* prestatiescore, maar over het algemeen genomen zijn de hoofdeffecten toch overheersend.

01	BVA	.51	.57	.62
02	OPV	.60	.63	.64
03	OPM	.48	.49	.51
04	AKG	.26	.22	.23
05	URB	.27	.15	.18
06	INT	.35	.39	.34
07	ASO	.28	.20	.22
08	OOA	.31	.24	.23
09	DWO	.19	.10	.09
10	BMB	.47	.45	.42
11	KLS	.22	.13	.15
12	DLO	.34	.41	.36
13	LL6	.16	.06	.12
14	ADV	.69	.78	.81
15	KGS	.61	.65	.64
16	BIL	.27	.21	.24
17	BIM	.43	.41	.43
18	PRE	.71	.79	.79
19	TON	.71	.83	.88
20	AOS	.23	.09	.47
21	LLS	.26	.13	.39
22	EXT	.24	.12	.60
23	ASL	.21	.15	.26
24	INS	.13	.07	.05
25	EIN	.71	.82	.85
	HOM	.15	.20	.24

tabel 1:
correlaties met afgeleide schaal uit SPSS/APL voor ongewogen, gewogen, en getransformeerde summatie scores.

01	BVA	.63	.63	.62
02	OPV	.65	.66	.63
03	OPM	.51	.50	.53
04	AKG	.24	.26	.24
05	URB	.18	.20	.16
06	INT	.35	.37	.31
07	ASO	.22	.24	.20
08	OOA	.22	.24	.24
09	DWO	.09	.10	.06
10	BMB	.41	.43	.40
11	KLS	.15	.18	.12
12	DLO	.38	.38	.38
13	LL6	.12	.13	.14
14	ADV	.81	.81	.82
15	KGS	.65	.63	.67
16	BIL	.24	.19	.34
17	BIM	.43	.37	.53
18	PRE	.80	.79	.81
19	TON	.89	.90	.89
20	AOS	.46	.36	.60
21	LLS	.37	.49	.32
22	EXT	.60	.66	.55
23	ASL	.26	.31	.23
24	INS	.10	.00	.10
25	EIN	.86	.87	.85
	HOM	.24	.25	.25

tabel 2:
correlaties met afgeleide schaal uit HOMALS voor totaal, vrouwen, mannen.

01	BVA																								
02	OPV	.56																							
03	OPM	.38	.46																						
04	AKG	.06	.11	.07																					
05	URB	.00	.11	.13	.19																				
06	INT	.20	.23	.14	.10	.10																			
07	ASO	.10	.14	.14	.05	.01	.06																		
08	OOA	.16	.16	.12	.01	.06	.05	.14																	
09	DWO	.00	.01	.03	.03	.09	.02	.01	.16																
10	BMB	.23	.28	.24	.13	.11	.12	.13	.19	.12															
11	KLS	.00	.08	.07	.01	.20	.03	.02	.02	.02	.07														
12	DLO	.12	.13	.09	.09	.05	.14	.01	.06	.02	.14	.01													
13	LL6	.01	.01	.02	.02	.18	.00	.03	.04	.06	.03	.07	.04												
14	ADV	.30	.35	.25	.12	.05	.25	.07	.09	.06	.23	.05	.32	.02											
15	KGS	.28	.31	.24	.08	.17	.16	.09	.15	.09	.23	.13	.19	.09	.43										
16	BIL	.08	.10	.04	.08	.04	.02	.00	.07	.02	.09	.03	.03	.00	.12	.09									
17	BIM	.17	.21	.15	.12	.10	.12	.01	.07	.02	.15	.02	.11	.04	.29	.18	.40								
18	PRE	.32	.34	.25	.13	.11	.24	.12	.13	.08	.25	.12	.33	.05	.67	.63	.12	.25							
19	TON	.39	.41	.29	.20	.05	.32	.10	.10	.05	.29	.07	.36	.04	.73	.48	.12	.29	.65						
20	AOS	.06	,10	.07	.03	.04	.05	.05	.05	.01	.01	.01	.10	.02	.16	.08	.04	.07	.09	.06					
21	LLS	.00	.11	.06	.00	.14	.05	.04	.03	.05	.06	.03	.11	.05	.15	.11	.07	.01	.11	.06	.49				
22	EXT	.07	.05	.05	.02	.01	.07	.06	.02	.01	.04	.01	.07	.01	.15	.09	.03	.09	.10	.04	.35	.31			
23	ASL	.05	.04	.07	.01	.02	.04	.36	.00	.00	.11	.02	.00	.03	.08	.08	.06	.04	.11	.10	.00	.03	.00		
24	INS	.00	.02	.02	.02	.03	.08	.06	.03	.00	.03	.00	.08	.01	.04	.03	.01	.00	.02	.07	.05	.02	.04	.05	
25	EIN	.40	.41	.31	.16	.04	.32	.10	.14	.02	.29	.07	.41	.03	.64	.46	.12	.26	.63	.76	.05	.03	.01	.08	.18

Tabel 3a: correlaties integer scores uit SPSS

01	BVA																								
02	OPV	.66																							
03	OPM	.46	.49																						
04	AKG	.14	.12	.09																					
05	URB	.16	.12	.15	.17																				
06	INT	.17	.24	.15	.11	.10																			
07	ASO	.15	.17	.15	.03	.01	.08																		
08	OOA	.18	.17	.11	.02	.08	.05	.15																	
09	DWO	.03	.04	.04	.05	.08	.04	.01	.15																
10	BMB	.28	.29	.25	.14	.11	.12	.16	.19	.10															
11	KLS	.09	.08	.08	.00	.21	.03	.00	.02	.02	.07														
12	DLO	.10	.13	.08	.09	.05	.14	.02	.06	.02	.14	.01													
13	LL6	.07	.04	.03	.01	.10	.00	.03	.03	.04	.03	.06	.02												
14	ADV	.37	.40	.30	.15	.07	.24	.11	.10	.06	.24	.07	.33	.08											
15	KGS	.34	.33	.26	.08	.18	.14	.11	.14	.08	.22	.12	.18	.11	.46										
16	BIL	.12	.14	.06	.07	.03	.04	.05	.08	.02	.11	.02	.06	.02	.18	.09									
17	BIM	.22	.23	.16	.11	.08	.12	.04	.08	.03	.16	.03	.12	.03	.33	.20	.40								
18	PRE	.37	.36	.26	.14	.11	.23	.12	.13	.08	.23	.11	.31	.11	.72	.61	.15	.30							
19	TON	.46	.47	.35	.19	.09	.27	.15	.12	.04	.28	.09	.34	.09	.80	.52	.17	.34	.71						
20	AOS	.21	.23	.18	.05	.04	.09	.07	.08	.01	.11	.04	.08	.04	.34	.25	.08	.14	.31	.42					
21	LLS	.14	.18	.14	.07	.13	.04	.06	.06	.05	.13	.08	.01	.03	.25	.21	.01	.06	.26	.33	.36				
22	EXT	.26	.26	.22	.12	.04	.17	.09	.11	.02	.17	.05	.13	.06	.47	.33	.06	.21	.42	.57	.34	.42			
23	ASL	.14	.15	.10	.00	.00	.09	.23	.08	.01	.16	.03	.06	.01	.16	.16	.05	.10	.18	.17	.11	.04	.09		
24	INS	.00	.01	.01	.03	.04	.08	.05	.03	.00	.03	.00	.00	.01	.04	.03	.02	.01	.02	.05	.02	.00	.05	.02	
25	EIN	.45	.45	.35	.18	.07	.28	.14	.14	.04	.27	.09	.38	.10	.74	.49	.17	.33	.69	.81	.35	.25	.50	.18	.13

Tabel 3b: correlaties optimale scores uit SPSS.

01 BVA	.07	.08
02 OPV	.04	.03
03 OPM	.05	.04
04 AKG	.01	.02
05 URB	-.01	-.02
06 INT	.05	.03
07 ASO	.00	-.00
08 OOA	.03	.03
09 DWO	-.04	.02
10 BMB	.03	.03
11 KLS	.01	.01
12 DLO	.12	.13
13 LL6	.01	.01
14 ADV	.13	.12
15 KGS	.03	.02
16 BIL	.00	.02
17 BIM	.02	.01
18 PRE	.15	.14
19 TON	.41	.35
20 AOS	-.05	.02
21 LLS	.01	-.00
22 EXT	-.02	.09
23 ASL	.01	.02
24 INS	.13	.16
MCOR	.67	.64

Tabel 6: regressiegewichten uit lineaire wegingen

01 BVA	.06	.06	.08
02 OPV	.02	.01	.06
03 OPM	.03	.02	.04
04 AKG	.02	.01	.05
05 URB	-.01	.01	.02
06 INT	.03	.09	.08
07 ASO	.01	.05	.02
08 OOA	.02	.04	.02
09 DWO	-.01	.07	.01
10 BMB	.01	.00	.01
11 KLS	.01	.00	.01
12 DLO	.09	.02	.07
13 LL6	.02	.02	.03
14 ADV	.16	.38	.19
15 KGS	.01	.03	.07
16 BIL	.00	.06	.04
17 BIM	.02	.08	.04
18 PRE	.14	.02	.10
19 TON	.43	.07	.38
20 AOS	.01	.26	.03
21 LLS	-.02	.30	.18
22 EXT	.06	.19	.08
23 ASL	.02	.02	.04
24 INS	.09	.21	.12
MCOR	.73	.97	.80

Tabel 7: regressiegewichten uit niet-lineaire analyses

01 BVA	.45	.19	.32
02 OPV	.45	.10	.42
03 OPM	.35	.08	.34
04 AKG	.18	.07	.16
05 URB	.07	.00	.06
06 INT	.28	.19	.18
07 ASO	.14	.21	.19
08 OOA	.14	.18	.21
09 DWO	.04	.12	.14
10 BMB	.27	.01	.16
11 KLS	.09	.07	.10
12 DLO	.38	.28	.44
13 LL6	.10	.03	.06
14 ADV	.74	.47	.78
15 KGS	.49	.16	.29
16 BIL	.17	.23	.11
17 BIM	.33	.23	.25
18 PRE	.69	.19	.70
19 TON	.81	.71	.90
20 AOS	.35	.77	.47
21 LLS	.25	.82	.42
22 EXT	.50	.86	.62
23 ASL	.18	.07	.20
24 INS	.18	.24	.21

Tabel 8: correlaties variabelen met niet-lineaire schalen.

Tabel 11a en 11b: Stapsgewijze regressie op Dronkers en HOMALS correlaties

01	BVA	.24	.23	.19	.20	.19	.12	.08	.08
02	OPV	.20	.18	.12	.11	.10	.05	.04	.03
03	OPM	.12	.11	.09	.09	.08	.05	.04	.04
04	AKG	.12	.11	.09	.09	.07	.06	.02	.02
05	URB	-.02	-.04	-.01	-.02	-.00	-.03	-.02	-.02
06	ASO		.02	.01	.01	.01	.01	.00	-.00
07	ASL		.05	.05	.05	.06	.02	.01	.02
08	BIL		.01	.01	.01	.01	.01	.01	.02
09	BIM		.16	.13	.14	.12	.03	.01	.01
10	INT			.17	.17	.14	.06	.04	.03
11	OOA			.05	.05	.04	.03	.03	.03
12	DWO			.08	.08	.07	.04	.02	.02
13	BMB			.11	.10	.08	.04	.03	.03
14	INS			.23	.23	.20	.18	.16	.16
15	KLS				.04	.04	.01	.01	.00
16	DLO					.29	.15	.12	.13
17	LL6					.04	.02	.01	.01
18	ADV						.30	.13	.12
19	KGS						.06	.03	.02
20	PRE						.20	.14	.14
21	TON							.38	.35
22	AOS								.02
23	LLS								-.01
24	EXT								.08
	MCOR	.23	.26	.34	.34	.42	.58	.63	.64
01	BVA	.23	.20	.19	.19	.19	.10	.06	.06
02	OPV	.23	.19	.14	.14	.13	.04	.02	.02
03	OPM	.12	.11	.10	.10	.09	.05	.03	.03
04	AKG	.12	.11	.08	.09	.07	.04	.02	.02
05	URB	-.03	-.04	-.01	-.02	-.00	-.02	-.02	-.01
06	ASO		.04	.03	.03	.04	.02	.01	.01
07	ASL		.09	.08	.07	.07	.02	.02	.02
08	BIL		.02	.02	.02	.02	.00	.00	.00
09	BIM		.19	.18	.18	.16	.05	.03	.02
10	INT			.13	.13	.11	.05	.03	.03
11	OOA			.02	.03	.02	.01	.02	.02
12	DWO			-.00	-.00	-.01	-.02	-.02	-.01
13	BMB			.07	.07	.05	.02	.01	.01
14	INS			.12	.12	.10	.09	.09	.09
15	KLS				.04	.03	.01	.01	.01
16	DLO					.27	.12	.09	.09
17	LL6					.06	.02	.02	.02
18	ADV						.41	.17	.16
19	KGS						.06	.02	.01
20	PRE						.21	.14	.14
21	TON							.46	.43
22	AOS								.01
23	LLS								-.02
24	EXT								.06
	MCOR	.27	.32	.36	.36	.43	.66	.72	.73

01 BVA	.28	.22	.06
02 OPV	.24	.14	.01
03 OPM	.15	.10	.02
04 AKG	.13	.09	.01
05 URB	.02	.03	.01
06 ASO		.03	.05
07 ASL		.08	.02
08 BIL		.05	.06
09 BIM		.16	.08
10 INT		.20	.09
11 OOA		.05	.04
12 DWO		.11	.07
13 BMB		.08	.00
14 INS		.20	.21
15 KLS		.04	.00
16 DLO		.32	.03
17 LL6		.07	.02
18 ADV			.38
19 KGS			.03
20 PRE			.06
21 TON			.07
22 AOS			.26
23 LLS			.30
24 EXT			.19
MCOR	.31	.55	.97

Tabel 12: Stapsgewijze niet-lineaire regressie

01 BVA	.08	.08	01 BVA	.14	.02
02 OPV	.10	.03	02 OPV	.34	.24
03 OPM	.04	.04	03 OPM	.20	.26
04 AKG	.05	.05	04 AKG	.16	.10
05 URB	.03	.01	05 URB	.01	.00
06 INT	.10	.07	06 INT	.26	.08
07 ASO	.03	.04	07 ASO	.17	.06
08 OOA	.03	.02	08 OOA	.14	.02
09 DWO	.05	.06	09 DWO	.13	.21
10 BMB	.00	.01	10 BMB	.09	.10
11 KLS	.00	.01	11 KLS	.10	.09
12 DLO	.08	.04	12 DLO	.46	.37
13 LL6	.02	.03	13 LL6	.08	.02
14 ADV	.19	.25	14 ADV	.65	.62
15 KGS	.05	.05	15 KGS	.30	.13
16 BIL	.07	.05	16 BIL	.14	.17
17 BIM	.03	.08	17 BIM	.21	.28
18 PRE	.09	.09	18 PRE	.48	.40
19 TON	.38	.28	19 TON	.89	.89
20 AOS	.02	.15	20 AOS	.57	.77
21 LLS	.25	.20	21 LLS	.68	.74
22 EXT	.10	.14	22 EXT	.77	.82
23 ASL	.05	.04	23 ASL	.09	.12
24 INS	.11	.16	24 INS	.17	.25
MCOR	.96	.96			

Tabel 13a: regressiegewichten en 13b: correlaties uit CANALS voor meisjes (eerste kolom) en jongens (tweede kolom.

		ADV		
BVA	PRE	H	M	L
H	H	44	13	1
H	M	10	10	15
H	L	0	1	6
M	H	166	138	22
M	M	49	253	379
M	L	1	19	246
L	H	11	23	9
L	M	6	56	135
L	L	0	4	152

Tabel 15a: gegevens voor interactie-analyse.

BVA:	.024	-.055	.080	PRE:	.024	-.034	.025
	.024	-.008	-.013		.024	-.000	-.023
	.024	.039	.020		.024	.034	.024
ADV:	.024	-.044	.031				
	.024	-.012	-.035				
	.024	.020	.009				

Tabel 15b: orthogonale polynomen voor interactie-analyse.

B1P1:	.00057	.00057	.00057	.00057	.00057	.00057	.00057	.00057	.00057
B1P2:	-.00081	.00000	.00081	.00081	.00000	.00081	.00081	.00000	.00081
B1P3:	.00059	-.00055	.00058	.00059	-.00055	.00058	.00059	-.00055	.00058
B2P1:	-.00131	-.00131	-.00131	-.00019	-.00019	-.00019	.00094	.00094	.00094
B2P2:	.00189	.00000	-.00189	.00027	.00000	-.00027	-.00135	.00000	.00135
B2P3:	-.00136	.00127	-.00135	-.00019	.00018	-.00019	.00097	-.00091	.00097
B3P1:	.00190	.00190	.00190	-.00030	-.00030	-.00030	.00048	.00048	.00048
B3P2	-.00273	.00000	.00273	.00043	.00000	-.00043	-.00069	.00000	.00069
B3P3	.00197	-.00184	.00196	-.00031	.00029	-.00031	.00050	-.00046	.00049

Tabel 15c: tensor basis voor functies op BVA x PRE.

HH	-.00196	-.00147	-.00119	-.00120
HM	-.00023	-.00044	-.00043	-.00043
HL	.00026	.00001	.00034	.00033
MH	-.00095	-.00092	-.00083	-.00083
MM	.00011	.00010	-.00006	-.00006
ML	.00054	.00056	.00070	.00070
LH	-.00033	-.00063	-.00046	-.00046
LM	.00029	.00039	.00030	.00031
LL	.00082	.00084	.00107	.00108
AH	-.0446	-.0442	-.0414	-.0439
AM	-.0114	-.0118	-.0148	-.0121
AL	.0194	.0195	.0202	.0196
MC	.4632	.4592	.4315	.4289

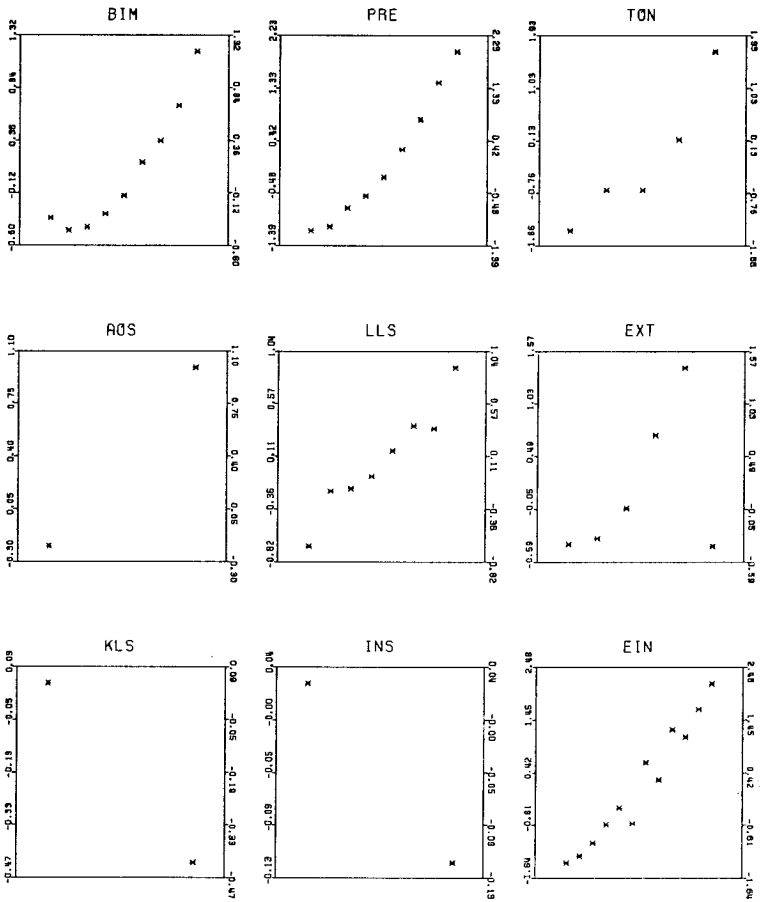
Tabel 15d: vier analyses: scores voor BVA x PRE, scores voor ADV, en gekwadrateerde multipele correlaties

-.00040	.00020	.00020
.00007	-.00001	-.00006
.00033	-.00019	-.00014
-.00091	.00021	.00070

-.00048
.00006
.00042
-.00016

Tabel 15e: functie op BVA x PRE ontbonden in hoofdeffecten en interacties.

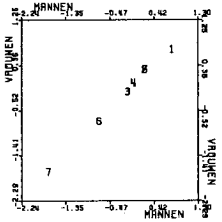
BVA
OPV
OPM
AKG
URB
OOA
DWO
INT
ASO
BMB
ASL
DLO
LLG
ADV
KGS
BIL



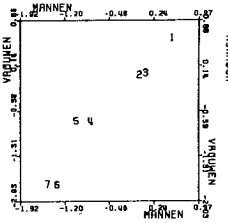
FIGUUR 4.1 t/m 4.25

Optimale niet-lineaire transformatie uit HOMALS (Y-as) geplot tegen de kategorienummers (X-as)

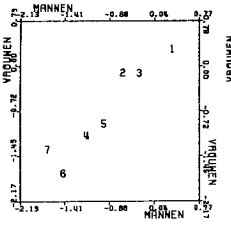
HOM-BVA



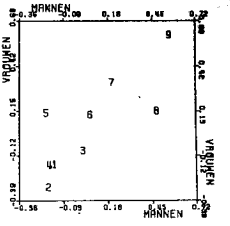
HOM-OPV



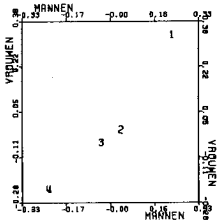
HOM-OPM



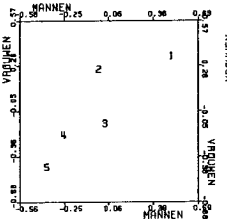
HOM-AKG



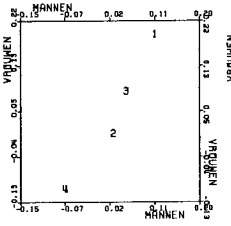
HOM-URB



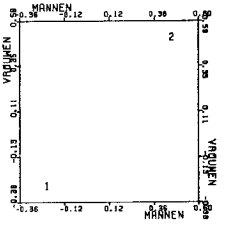
HOM-OOA



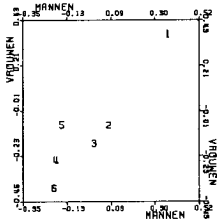
HOM-DWO



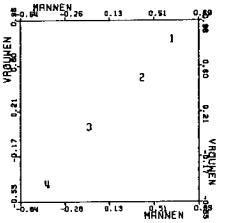
HOM-INT



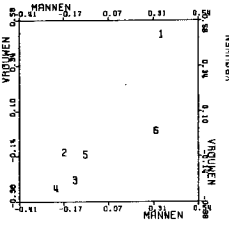
HOM-ASO



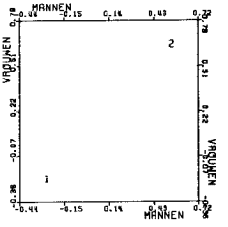
HOM-BMB



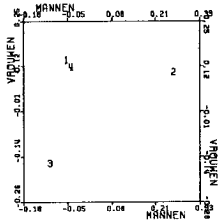
HOM-ASL



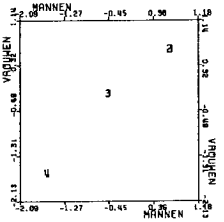
HOM-DLO



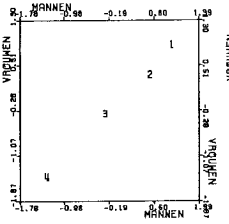
HOM-LL6



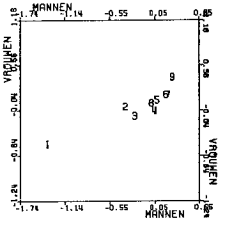
HOM-ADV

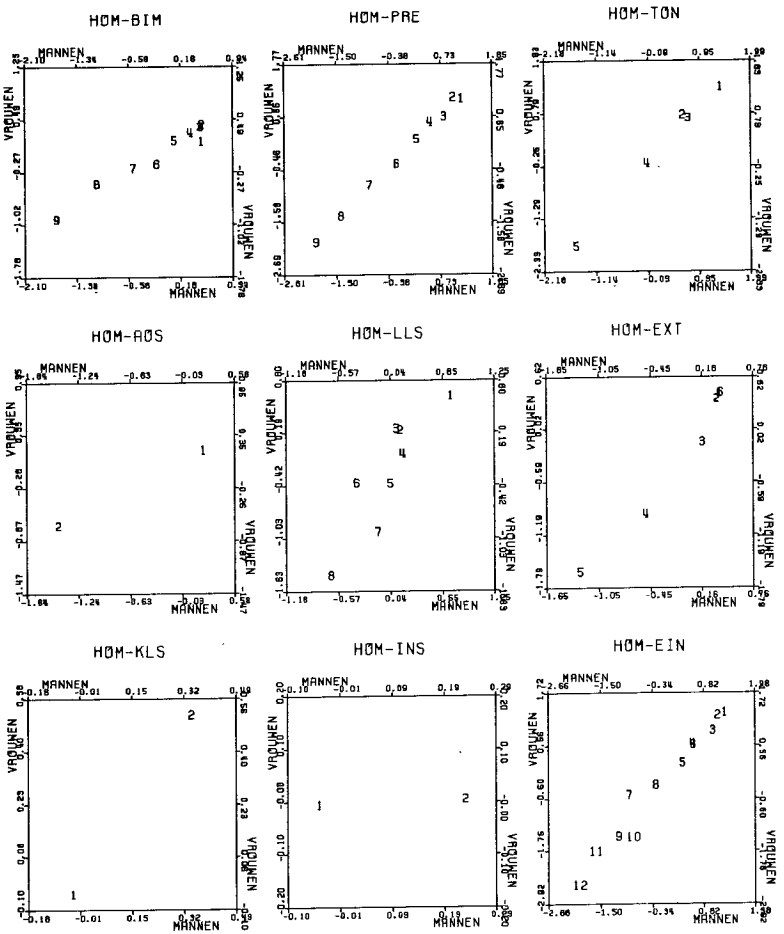


HOM-KOS



HOM-BIL

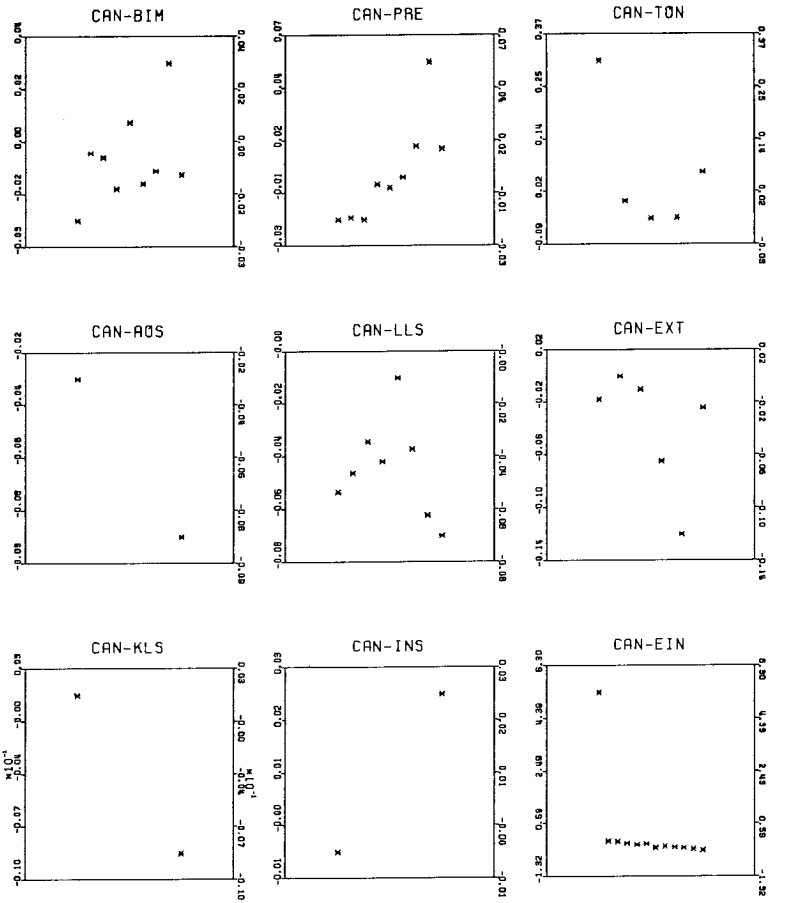




FIGUUR 5.1 t/m 5.25

Optimale niet-lineaire transformatie uit HOMALS voor vrouwen (Y-as) geplot tegen idem voor mannen (X-as).

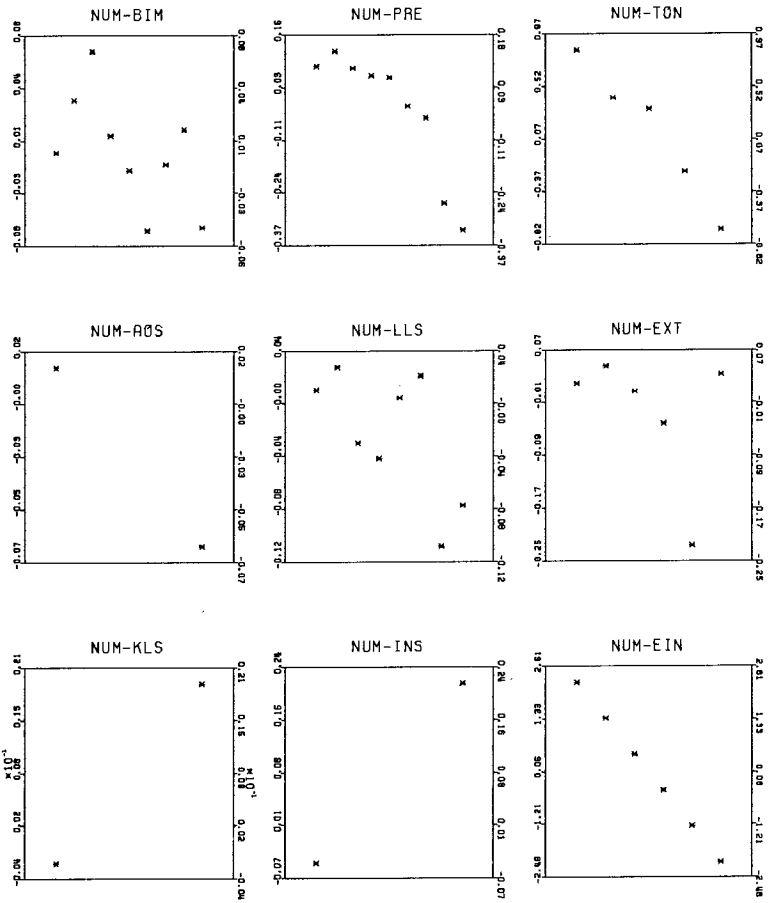
CAN-BVA
CAN-OPV
CAN-OPM
CAN-AKG
CAN-URB
CAN-OOA
CAN-DWO
CAN-INT
CAN-ASO
CAN-BMB
CAN-ASL
CAN-DLO
CAN-LL6
CAN-ADV
CAN-KGS
CAN-BIL



FIGUUR 9.1 t/m 9.25

Optimale niet-lineaire transformatie uit CANALS (**Y**-as) geplot tegen kategorienummers (X-as)

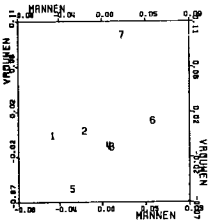
NUM-BVA
NUM-OPV
NUM-OPM
NUM-AKG
NUM-URB
NUM-OOA
NUM-DWO
NUM-INT
NUM-ASO
NUM-BMB
NUM-ASL
NUM-DLO
NUM-LL6
NUM-ADV
NUM-KGS
NUM-BIL



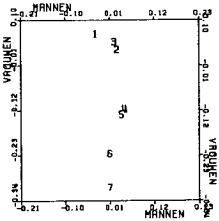
FIGUUR 10.1 t/m 10.25

Optimale niet-lineaire transformatie uit CANALS (Y-as) geplot tegen kategorienummers. De transformatie van 25:EIN is numeriek gehouden.

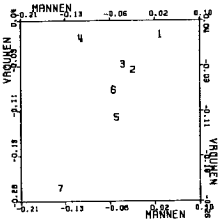
CAN-BVA



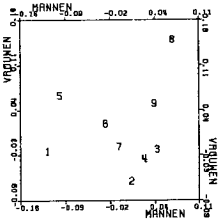
CAN-OPV



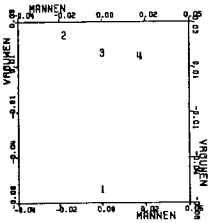
CAN-OPM



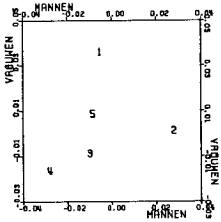
CAN-AKG



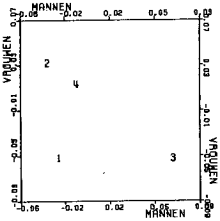
CAN-URB



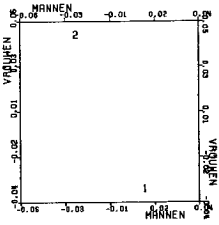
CAN-OOA



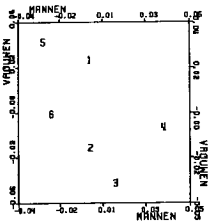
CAN-DWO



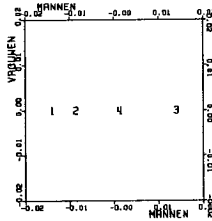
CAN-INT



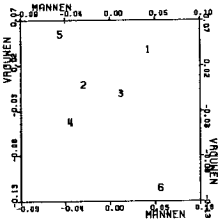
CAN-ASO



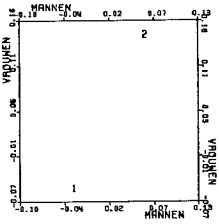
CAN-BMB



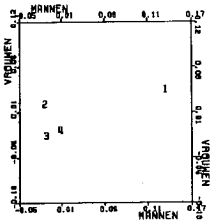
CAN-ASL



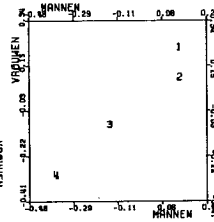
CAN-OLO



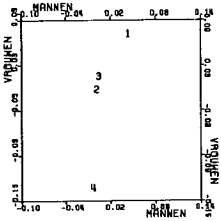
CAN-LL6



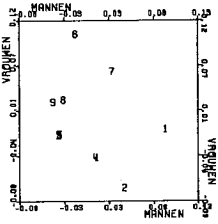
CAN-ADV

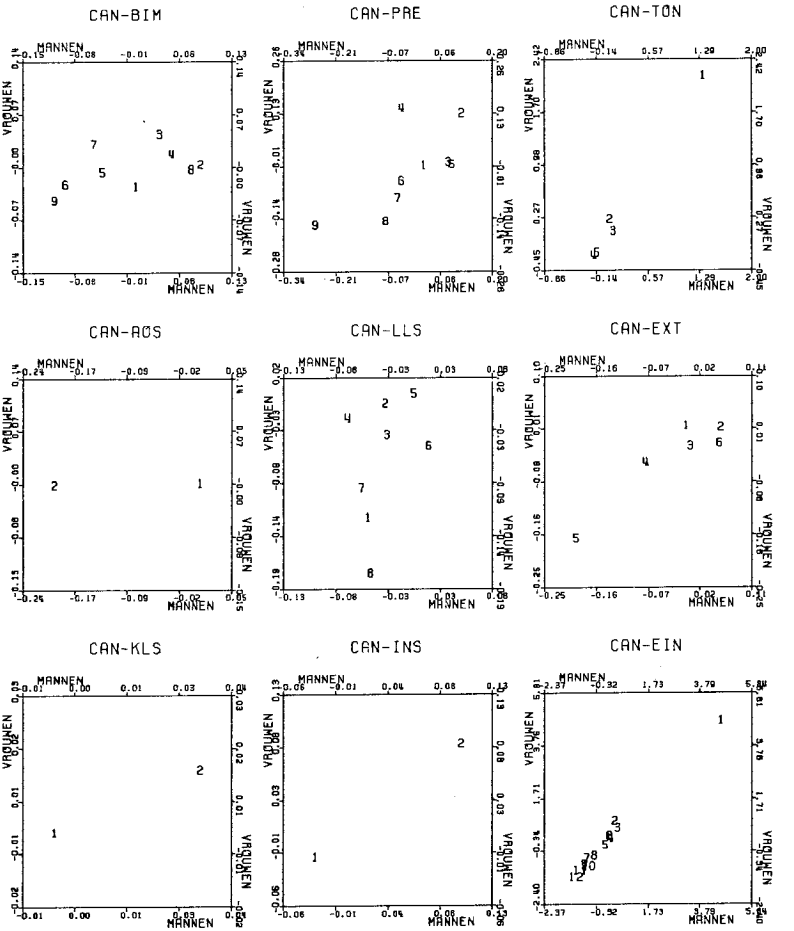


CAN-KGS



CAN-BIL





FIGUUR 14.1 t/m 14.25

Optimale niet-lineaire transformatie voor vrouwen uit CANALS (Y-as) geplot tegen idem voor mannen (X-as).

Appendix: Lijst van gebruikte variabelen met afkortingen

01	BVA	Beroep vader
02	OPV	Opleiding vader
03	OPM	Opleiding moeder
04	AKG	Aantal kinderen gezin
05	URB	Urbanisatiegraad
06	INT	Interesse ouders
07	ASO	Aspiratieniveau ouders
08	OOA	Ontvankelijk opvatting anderen
09	DWO	Dwang ouders
10	BMB	Belang beroepskeuze meisje
11	KLS	Kleuterschool
12	DLO	Zitten blijven
13	LL6	Aantal leerlingen 6e klas
14	ADV	Advies onderwijzer
15	KGS	Klassegemiddelde toets
16	BIL	Beroepsinteresse L
17	BIM	Beroepsinteresse M
18	PRE	Prestatiescore
19	TON	Eerste keuze vervolgonderwijs
20	AOS	Andere opleidingen aan school
21	LLS	Aantal leerlingen secundair
22	EXT	Aantal extra-curriculaire activiteiten
23	ASL	Aspiratieniveau leerling
24	INS	Instemming ouders schoolkeuze
25	EIN	Bereikt eindniveau secundair onderwijs

Voor nadere informatie over de variabelen en hun categorieën verwijzen we naar de oorspronkelijke ITS publicatie, en naar Dronkers (1979).

Literatuur

Broman, S.A.; P.L. Nichols, W.A. Kennedy, *Preschool IQ*, Hillsdale N.J., Earlbaum Associates, 1975.

Buikhuisen, W.; *An alternative approach to the aetiology of crime*. WODC rapport, Ministerie van Justitie, februari 1978.

Dessens, J.; W. Janssen, *Van Jaar tot Jaar: een commentaar op Dronkers en een analyse op interacties*. Mens en Maatschappij, 54, 1979, 87-98.

Dronkers, J., *Manipuleerbare variabelen in de schoolloopbaan.* In: J.L. Peschar, W.C. Ultee (eds): Sociale Stratificatie. Deventer, Van Loghum Slaterus, 1978.

Dronkers, J.; M. Jungbluth, *Schoolloopbaan en geslacht.* Paper gepresenteerd op de Onderwijsresearchdagen, Nijmegen, 1979.

Eysenck, H.J., *Crime and Personality.* London, Routlegde, Kegan, and Paul, 1964.

Jencks, C. e.a., *Inequality.* New York, Basic Books, 1972.

Jensen, A., *How much can we boost IQ and educational achievement.* Harvard Educational Review, 39, 1969, 1-123.

Leeuw, J. de, *De politieke relevantie van correlaties.* Sociologische Gids, 25, 1978, 31-39.

Mednick, S.A.; Cristiansen, K.O., *Biosocial basis of criminal behaviour.* New York, Gardner, 1977.

Pearson, K., *The grammar of science.* London, Scott, 1892 (2^e ed 1900, 3^e ed 1911).

Rutter, M., N. Magde, *Cycles of disadvantage.* London, Heinemann, 1976.

8. REACTIE OP DE LEEUW EN STOOP

Albert Verbeek

Het is een genoegen om kennis te nemen van deze analyse en van de daarin gepresenteerde inzichten zowel op inhoudelijk als op methodologisch-technisch terrein. Dit is des te meer een verademing omdat anno 1979 nog steeds door velen 'invloed' 'regressiecoëfficiënt' en 'associatie' als synoniemen gebruikt worden. Een typerend voorbeeld is de geparafraseerde zin 'het verschil tussen sexen van de invloed van X_1 op Y is 0,01 en dus opmerkelijk veel kleiner dan het verschil in invloed van X_2 op Y dat 0,05 bedraagt'. Bedoeld maar onvermeld is dat het hier om een verschil in gestandaardiseerde regressiecoëfficiënten gaat, terwijl de veel interessanter verschillen in ligging (= gemiddelde), in standaard deviaties en in ongestandaardiseerde regressiecoëfficiënten buiten beeld blijft, evenals het kleine aantal categorieën van de variabele Y, waardoor regressie een nogal lomp instrument wordt.
De multivariate analyse van samenhang blijft heel vaak beperkt tot de berekening van enkele of zeer vele correlaties, regressiecoëfficiënten, padcoëfficiënten of factorscores, doorgaans opgesierd door een haast obscene belangstelling voor significantie. Echter, de constructie van een model blijft doorgaans achterwege, en dus ook een analyse van de residuen, onderzoek naar niet-lineariteit van verbanden (= locale onzuiverheid), naar homoscedastisiteit en zelfs naar normaliteit in die gevallen waar dit voor de conclusies (bijvoorbeeld bij betrouwbaarheidsintervallen) wel gebruikt wordt.

De analyse van De Leeuw en Stoop is een fraai voorbeeld tot welke detaillering en daarmee precisie een deskundig uitgevoerde analyse kan gaan. Zelfs als de conclusie luidt dat voor een aantal verbanden een eerste orde, lineair regressiemodel een uitstekende benadering biedt dan is de winst van deze vorm van modelbouw nog groot en wel vooral om de volgende twee redenen. Ten eerste weten we nu dat het regressiemodel een goed model biedt. Ten tweede hebben we, indien gewenst, direct een verfijning van het regressiemodel achter de hand. We weten, of vermoeden in welke richtingen er (kleine) afwijkingen zijn van het lineaire verband.

Een mogelijk bezwaar van hun gedetailleerde rapportage is dat het, althans in deze vorm, voor beleidsvoerders nog slecht te lezen is; het is wellicht nog te gedetailleerd en duidelijke, versimpelde beleidsondersteunende conclusies ontbreken. De korte tijd waarin dit rapport tot stand gebracht werd, de nadruk op analyse en het feit dat er met dit materiaal bitter weinig beleidsondersteuning te geven

is (zie 2.2.1.), zijn hier echter duidelijke redenen voor.

De aanpak maakt het verleidelijk om in te gaan op de methodologische grondslagen van de gebruikte schaalmethoden. Maar ik wil me hier verder beperken tot een drietal niet al te technische opmerkingen.

In 2.1.3. merken De Leeuw en Stoop op dat zij 'globale statistische informatie' zoals tests van hypothesen en schattingen van parameters over het algemeen niet erg informatief vinden. In 0.3. bepleiten zij, en mijns inziens terecht, meer waardering voor (goede) descriptieve en exploratieve analyses. Zoals ook in de discussie blijkt gaat hun interesse vooral uit naar modellen met veel parameters die daarmee een grote mate van flexibiliteit en detaillering mogelijk maken tegen de prijs van een groter risico van kanskapitalisatie. In de meest gebruikte (of beter misbruikte) statistische toetsings- en schattingsmodellen zijn veelal erg weinig of erg onverstandig parameters ingebouwd, en in veel analyses wordt weinig gedaan aan het checken van het model zelf. Toch is er voor een analyse van een model met veel parameters veel inzicht nodig, en blijft de vraag overeind naar de nauwkeurigheid van de verkregen schatters. Deze vraag zou in elk geval met kruisvalidering (jack-knifing, bootstrapping), gedeeltelijk beantwoord kunnen worden, en wellicht ook door middel van berekeningen van asymptotische eigenschappen. In de discussie blijkt dat hieraan gewerkt wordt, en verwijst De Leeuw ook naar het uitstekende artikel van B. Efron (1).

In de conclusies wordt op verschillende plaatsen beoordeeld of een samenhang 'redelijk additief' of 'redelijk lineair' is en of een verdeling 'redelijk normaal' is. Redelijkheid moet hierbij onder andere afgemeten worden aan de vraag in hoeverre de verdere analyse gebruik maakt van deze assumpties en in hoeverre de verdere analyse gevoelig is voor afwijkingen. Omdat daarover slechts verspreide resultaten bekend zijn blijft deze noodzakelijke beoordeling een ietwat riskante combinatie van kennis, ervaring en behendigheid in het doorrekenen van de consequenties van afwijkingen.

Een interessant detail uit het betoog is nog hun klacht over het suggestief taalgebruik van statistici (en van wiskundigen). Een naam is voor een wiskundige meestal een gekozen label voor een goed begrepen abstractie. Hoe suggestiever de naam, des te gemakkelijker is deze te onthouden. Voor gebruikers van wiskunde en met name statistiek is de onderliggende abstractie vaak minder paraat en gaat de naam mede als verklaring dienen. Dan blijken veel termen voor deze categorie gebruikers té suggestief, en zou een kleurlozer taalgebruik hen wellicht tot een grondiger begrip dwingen. Men denke aan woorden als 'normaal', 'regulier','signi-

ficant', 'onderliggende factoren', 'associatiemaat', 'cluster', 'interactie' en 'effect'. Nauw verwant hiermee is dat eenzelfde formule of analysetechniek vaak in wezenlijk verschillende situaties met verschillende interpretaties toegepast kan worden (effect versus samenhang). Een te sterke koppeling tussen formules en taalgebruik leidt dan bijvoorbeeld tot het misverstand dat regressie-analyse van surveydata iets zegt over causaliteit.

Noten

De reactie is eerst na afloop van de ORD op papier gezet waardoor ik ook dankbaar gebruik heb kunnen maken van de daar gevoerde discussie.

Literatuur

Efron, B., *Bootstrap methods: another look at the jack-knife*, Annals of Statistics, 7 (1979) p. 1-26.