

Causale Modellen met Diskrete, Latente Variabelen:
een variant op de lisrel-benadering

J.A.P. Hagenaars

A.G.J. Heinen

P.A.M. Hamers

1. Inleiding

In zijn vele uiteenzettingen van het loglineaire model gebruikt Goodman frekvent de termen "modified multiple regression approach" en "modified path analysis approach" (o.a. Goodman 1972, 1973). Hoewel deze terminologie hier en daar aanleiding heeft gegeven tot misverstanden, komt daardoor toch goed naar voren dat m.b.v. loglineaire modellen hypothesen over kausale relaties tussen diskrete (nominale) variabelen getoetst kunnen worden die op diverse punten duidelijk overeenkomst vertonen met het soort hypothesen dat getoetst wordt bij multiële regressie en padanalyse.

De (analoge) lijn van multiële regressie naar padanalyse doortrekkend, zouden we terecht komen bij zoiets als een "modified lisrel approach", waarvan het kenmerkende zou zijn dat in de kausale modellen latente, niet-rechtstreeks gemeten variabelen voor kunnen komen. (Voor de lisrel-benadering zie o.a. Jöreskog, 1973 en Saris, 1974.)

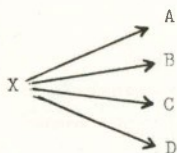
Op welke wijze loglineaire analyses met latente variabelen uitgevoerd kunnen worden, is in principe aangegeven door Goodman, 1974a, 1974b, 1975 en Haberman, 1979. Goodman sluit hierbij expliciet aan bij het door Lazarsfeld ontwikkelde latente-klasse-analyse-model (Lazarsfeld en Henry, 1968).

In de volgende paragraaf zal dit "klassieke" latente-klasse-analyse-model zeer kort uiteen gezet worden. In de daarop volgende paragraaf 3 wordt aangegeven hoe dit model waarin slechts één latente variabele voor komt, uitgebreid kan worden tot een model met meerdere latente variabelen. In de paragrafen 4 en 5 wordt dan aangegeven op welke wijze vrij complexe kausale modellen met diverse latente en manifeste variabelen geanalyseerd kunnen worden. Hoewel de werkwijze niet nieuw is, wordt in de literatuur nog weinig of geen aandacht besteedt aan de modellen die in deze beide paragrafen besproken worden. Tenslotte wordt in de laatste paragraaf kort ingegaan op de beschikbare algoritmes en op enkele moeilijkheden verbonden aan de uitvoering ervan.

2. Modellen met één latente variabele

Uitgangspunt zijn enkele manifeste, rechtstreeks gemeten variabelen van nominaal nivo (bv. A, B, C en D) die als indicatoren beschouwd worden van één onderliggende, niet-rechtstreeks gemeten, polytome variabele X. De samenhangen tussen de indicatoren worden geheel verklaard vanuit de effecten van de latente variabele op elke manifeste variabele, m.a.w. de manifeste variabelen zijn onafhankelijk van elkaar met konstanthouding van de latente variabele. Men kan dit als volgt in een pijlendiagram weergeven.

figuur 1.



Ter bepaling van de gedachten zou men zich X kunnen voorstellen als de variabele "politieke opvatting" met als categorieën socialistisch, liberaal en christen-democratisch en A, B, C en D als stemgedrag, mening over nationalisatie van bedrijven, over abortus en over persvrijheid.

Een "gewone" loglineaire analyse van het in figuur 1 weergegeven model is niet mogelijk omdat immers de frekwenties f_{ijkl}^{ABCDX} niet gegeven zijn vanwege het latente karakter van X. Men gaat dan als volgt te werk.

In het latente-klasse-analyse-model wordt er van uitgegaan dat de relatie tussen de latente variabele en een manifeste variabele probabilistisch is. De score op de latente variabele legt niet zonder meer eenduidig de score op een manifeste variabele vast, maar gegeven dat een onderzoekspersoon behoort tot categorie t van variabele X, heeft hij een bepaalde kans om op variabele A tot categorie i te behoren. Deze (konditionele) kans wordt aangeduid met P_{it}^{AX} . Er wordt verder van uitgegaan dat deze kans voor alle onderzoekspersonen gelijk is (homogeniteitsassumptie). Vanzelfsprekend geldt ook (indien $i=1 \dots I$):

$$\sum_{i=1}^I P_{it}^{AX} = 1 \quad (1)$$

Op analoge wijze kan de betekenis van P_{jt}^{BX} , P_{kt}^{CX} en P_{lt}^{DX} omschreven worden. Essentieel binnen het model is verder dat de konditionele kansen behorende bij verschillende manifeste variabelen onafhankelijk van elkaar zijn. Deze assumptie van lokale onafhankelijkheid houdt in dat binnen een bepaalde latente klasse, voor een bepaalde waarde t van X, de scores op de ene mani-

feste variabele niet samenhangen met die op een andere manifeste variabele. Op basis hiervan kan de kans dat een willekeurig onderzoekselement op de gecombineerde variabele ABCDX de score ijkl krijgt, als volgt bepaald worden:

$$P_{ijkl}^{ABCDX} = P_t^X P_{it}^{\bar{A}X} P_{jt}^{\bar{B}X} P_{kt}^{\bar{C}X} P_{lt}^{\bar{D}X} \quad (2)$$

De kans op score ijkl op de gecombineerde manifeste variabele ABCD wordt dan (indien $t = 1 \dots T$):

$$P_{ijkl}^{ABCD} = \sum_{t=1}^T P_{ijkl}^{ABCDX} \quad (3)$$

de kans op score t op X is gelijk aan:

$$P_t^X = \sum_{i,j,k,l} P_{ijkl}^{ABCDX} \quad (4)$$

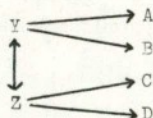
Goodman (1974a, 1974b) geeft een iteratieve procedure aan met behulp waarvan de maximum-likelihood-schattingen $\hat{P}_t^X$, $\hat{P}_{it}^{\bar{A}X}$ etc. gevonden kunnen worden op basis van de modelassumpties en de manifeste proporties P_{ijkl}^{ABCD} .

Via (2) kan dan $\hat{P}_{ijkl}^{ABCDX}$ verkregen worden. Vermenigvuldiging van deze laatste grootte met de steekproefomvang N geeft de geschatte verwachte frekwenties $\hat{F}_{ijkl}^{ABCDX}$ voor het model in figuur 1, zodat op de "gebruikelijke" wijze nu de effecten in het model berekend kunnen worden. Een toets op de adekwaatheid van het model is mogelijk door de vergelijking van $\hat{P}_{ijkl}^{ABCD}$, te berekenen via (3), met de manifeste proporties P_{ijkl}^{ABCD} (of liever van $\hat{F}_{ijkl}^{ABCD}$ met $\hat{F}_{ijkl}^{ABCD}$).

3. Modellen met twee of meer latente variabelen

Hieronder zal alleen ingegaan worden op een model met twee latente variabelen Y en Z die ieder geïndiceerd worden door twee manifeste variabelen A en B resp. C en D. Het model kan als volgt weergegeven worden:

Figuur 2.



Ook in dit model geldt weer dat alle samenhangen tussen de manifeste variabelen tot stand komen vanuit de effecten van de latente variabelen. Het model van figuur 2 is dan ook te herleiden tot een "klassiek" latente-klasse-analyse-model, door de $R \times S$ -categorieën van de gecombineerde variabele YZ te beschouwen als de T-kategorieën van de ene latente variabele X.

Wel gelden dan t.o.v. het model van figuur 1 extra restricties. Immers, het model in figuur 2 postuleert dat de variabelen X en A onafhankelijk van elkaar zijn bij konstanthouding van de variabele Y , m.a.w. de conditionele waarschijnlijkheid dat een willekeurig onderzoekselement op A i scoort, gegeven dat hij op de gecombineerde variabele YZ rs scoorde, $P_{irs}^{\bar{A}YZ}$, is alleen afhankelijk van de score r van Y , en niet van de score s van Z . Dit brengt de volgende gelijkheden met zich mee:

$$P_{ir1}^{\bar{A}YZ} = P_{ir2}^{\bar{A}YZ} = \dots = P_{irs}^{\bar{A}YZ} \quad (5)$$

Op basis van soortgelijke overwegingen geldt bovendien:

$$\begin{aligned} P_{jr1}^{\bar{B}YZ} &= P_{jr2}^{\bar{B}YZ} = \dots = P_{jrS}^{\bar{B}YZ} \\ P_{k1s}^{\bar{C}YZ} &= P_{k2s}^{\bar{C}YZ} = \dots = P_{kRs}^{\bar{C}YZ} \end{aligned} \quad (6)$$

$$P_{11s}^{\bar{D}YZ} = P_{12s}^{\bar{D}YZ} = \dots = P_{1Rs}^{\bar{D}YZ}$$

De maximum-likelihood-schattingen $\hat{P}_{ijklrs}^{ABCDYZ}$ kunnen gevonden worden door de eerder vermelde iteratieve procedure van Goodman toe te passen, waarbij de variabele X gelijk is aan de gecombineerde variabele YZ , en t.a.v. de conditionele waarschijnlijkheden de restricties (5) en (6) gelden. Goodman, 1974a, 1974b laat zien hoe deze restricties gemakkelijk in zijn procedure in te bouwen zijn.

Op deze wijze kunnen dus de schattingen $\hat{P}_{ijklrs}^{ABCDYZ}$ verkregen worden, waarna de diverse effecten in het model eenvoudig te berekenen zijn. Met name is het dus zo mogelijk de relatie tussen de "theoretische" variabelen Y en Z te bepalen, en wel met behulp van $\hat{P}_{rs}^{YZ}$.

Uiteraard maakt de hier geschetste gedachtengang het ook mogelijk ingewikkelder modellen met meer latente variabelen en/of meer indicatoren op te stellen.

4. Causale modellen zonder restricties m.b.t. de verdeling van X

Latente klasse analyse is vooral gebruikt om na te gaan of de relaties tussen een aantal indicatoren verklaard kunnen worden vanuit de effecten van één of meer onderliggende, theoretische variabelen. Maar eenmaal dat deze analyse uitgevoerd is, komt de vraag aan de orde op welke wijze de latente variabelen gerelateerd zijn aan andere onderzoeksvariabelen. Bv. in het geval Y en Z in figuur 2 houdingen weergegeven: op welke wijze zijn achtergrondfactoren zoals opleiding van invloed op deze houdingen (fig. 3a), dan

wel: op welke wijze beïnvloeden deze houdingen bepaalde gedragsvariabelen (fig. 3b).

figuur 3.



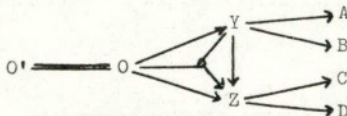
Omdat de onderzoeker niet beschikt over de individuele scores op de latente variabelen, zou hij als volgt te werk kunnen gaan. Eerst worden de kansen berekend dat een onderzoekselement een bepaalde score t op X heeft, gegeven zijn antwoordpatroon op de manifeste variabelen A, B, C en D :

$$\frac{\hat{P}_{ijklt}^{ABCDX}}{\hat{P}_{ijkl}^{ABCD}} = \frac{\hat{P}_{ijklt}^{ABCDX}}{\hat{P}_{ijkl}^{ABCD}} \quad (7)$$

Men kan een onderzoekselement nu die score op X toekennen waarvoor $\frac{\hat{P}_{ijklt}^{ABCDX}}{\hat{P}_{ijkl}^{ABCD}}$ het grootst is. Een nadeel van deze werkwijze is dat er een zekere mate van onbetrouwbaarheid in de scores van X optreedt. Een aantal personen zullen immers "verkeerd" geplaatst worden. De samenhangen van de latente variabele(n) met de achtergrondfactoren en/of gedragsvariabelen zullen uiteraard hierdoor beïnvloed worden.

Beter is het volgende alternatief, geïllustreerd aan de hand van figuur 3a. O (opleiding) wordt "geherdefinieerd" als een latente variabele O , die "perfekt" verbonden wordt met de manifeste variabele O' , de variabele opleiding zoals waargenomen. In een pijlendiagram:

figuur 4.



O, Y en Z in figuur 4 zijn latente variabelen. Over de verdeling van deze variabelen worden geen assumpties gemaakt, met name worden drie-variabelen interacties dus niet uitgesloten. Door de $O \times Y \times Z$ -categorieën van de gekombineerde variabele OYZ te beschouwen als de T -categorieën van de ene latente variabele X , kan het model in figuur 4 herleid worden tot een klassiek

latente-klasse-analyse-model met één latente variabele, zij het dat wederom bepaalde restricties aan de conditionele kansen opgelegd worden.

In figuur 4 is de variabele A onafhankelijk van O en Z met konstanthouding van Y. $P_{ir}^{\bar{A}Y}$ is dus alleen afhankelijk van de score op Y en niet van de scores op O en Z. De gelijkheden vermeld in (5) gelden ook voor het model in figuur 4, met dien verstande dat in (5) Z vervangen wordt door de gekombineerde variabele OZ. Ook de gelijkheden in (6) gelden mutatis mutandis voor het model in figuur 4.

De relatie tussen O en O' is in figuur 4 weergegeven m.b.v. een dubbele lijn om aan te geven dat O geen "echte" latente variabele is, maar in feite identiek aan O'. In termen van de konditionele kansen $P_{m q}^{\bar{O}'O}$ betekent dit dat voor elke categorie rs van de gekombineerde variabele YZ geldt:

$$\begin{aligned} P_{m q}^{\bar{O}'O} &= 1 \quad \text{als } m = q \\ &= 0 \quad \text{als } m \neq q \end{aligned} \quad (8)$$

Ook gelijkheden zoals (8) zijn, zoals Goodman, 1974a, 1974b aangeeft, gemakkelijk in de eerder vermelde iteratieve procedure in te bouwen.

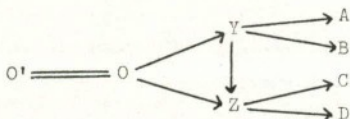
Op deze wijze kunnen de schattingen $\hat{P}_{ijklm, qrs}^{ABCD'O'YZ}$ verkregen worden en dus ook $\hat{P}_{ijklm, qrs}^{ABCD'O'YZ}$, die vanwege de identiteit tussen O en O' samengevoegd kunnen worden tot $\hat{P}_{ijklqrs}^{ABCD'OYZ}$ met behulp waarvan op de gebruikelijke wijze de de effecten in het model van figuur 4 berekend kunnen worden.

Zo kunnen dus nu de relaties tussen de latente variabelen en de overige onderzoeksvariabelen op betrouwbare wijze vastgesteld worden. Om echter een "volwaardig" analyseinstrument te zijn voor de evaluatie van complexe causale modellen zou de tot nu toe beschreven procedure nog aangevuld moeten worden met de mogelijkheid restricties in te bouwen ten aanzien van de verdeling van de latente variabele X, d.w.z. ten aanzien van de verdeling van de gekombineerde variabele OYZ.

5. Causale modellen met restricties m.b.t. de verdeling van X

Een mogelijke restrictie t.a.v. de verdeling van de latente variabelen in figuur 4 is het afwezig zijn van de drie-variabelen interactie OYZ, zoals weergegeven in figuur 5.

figuur 5.



Voor het model in figuur 5 geldt in eerste instantie al hetgeen in par.4 opgemerkt is over het model in figuur 4. De schattingsprocedure verloopt dus hetzelfde, met dien verstande dat in figuur 5 extra restricties van toepassing zijn t.a.v. P_t^X , t.a.v. P_{qrs}^{OYZ} dus. In de notatie van Goodman: de modelfrekwenties F_{qrs}^{OYZ} hebben betrekking op het model $\{OY, OZ, YZ\}$. De voorlopige, tussentijdse schattingen $\hat{F}_{qrs}^{OYZ}$ moeten dus telkens omgezet worden tot schattingen $\hat{P}_{qrs}^{OYZ}$ die voldoen aan hetgeen in het model gespecificeerd is. Dit kan via de "gebruikelijke" iteratieve "proportional fitting" procedure, waarbij de nieuwe schattingen $\hat{P}_{qrs}^{OYZ}$ voldoen aan:

$$\hat{P}_{qr+}^{OYZ} = P_{qr+}^{OYZ}$$

$$\hat{P}_{q+s}^{OYZ} = P_{q+s}^{OYZ}$$

(9)

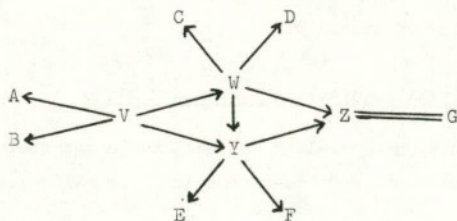
$$\hat{P}_{+rs}^{OYZ} = P_{+rs}^{OYZ}$$

Uiteindelijk zijn zo de geschatte, verwachte frekwenties $\hat{F}_{ijklqrs}^{ABCDVYZ}$ te bepalen en daarmee ook de effecten in figuur 5.

In de volgende paragraaf zal nog nader ingegaan worden op het algoritme. In de rest van deze paragraaf zullen enkele konkrete toepassingen besproken worden. Dat al deze voorbeelden betrekking hebben op dichotome variabelen heeft alleen te maken met de op dat moment beschikbare programmatuur en verwijst geenszins naar een principiële inperking van de modellen.*

Het eerste model (figuur 6) is toegepast op gefingeerde data en is alleen gebruikt om de procedure uit te testen. Uitkomsten van de analyse zullen dan ook niet vermeld worden.

figuur 6.



* De programma's zijn geschreven door C.J.M. van Brunshot en M.A.G. van der Walle.

V, W en Y zijn latente variabelen die aangeven de politieke interesse van de onderzoekspersonen in resp. januari, februari en maart, met als categorieën: wel interesse, geen interesse. A, C en E geven weer of de onderzoekspersoon al dan niet in de krant over politiek leest in deze drie maanden en B, D en F zijn de antwoorden op de vraag of men in de drie maanden al dan niet naar tv-programma's over politiek kijkt. Z tenslotte is de "latente" variabele stemgedrag in mei, waarbij G de perfecte indikator: gaat wel vs. gaat niet stemmen is.

In figuur 6 wordt er van uitgegaan dat "kijken" en "lezen" beide indicatoren zijn van de theoretische variabele politieke interesse, die in de betreffende onderzoeksperiode niet konstant hoeft te zijn.

(Zou men na willen gaan of A t/m F indicatoren zijn van één stabiele, theoretische variabele politieke interesse, dan had men voor de zes indicatoren één latente variabele U moeten postuleren.) Hoewel het mogelijk is de veranderingen in de tijd aan extra voorwaarden te onderwerpen, bv. een markov-keten te postuleren (Bishop e.a., 1975), wordt dit in figuur 6 niet verondersteld. Wel wordt de afwezigheid van de drie-variabelen interactie VWY aangenomen.

De politieke interesse in januari bepaalt niet, die van februari en maart bepalen wel rechtstreeks het stemgedrag in mei. Daarbij worden alle drie- en vier-variabelen interacties tussen de latente variabelen uitgesloten. De adequaatheid van het opgestelde model kan getoetst worden via de likelihood ratio L, dan wel Pearson's χ^2 . De geobserveerde frekwenties zijn in dit geval: $f_{ijklmno}^{ABCDEFG}$; de verwachte frekwenties: $\hat{f}_{ijklmno}^{ABCDEFG}$. Het aantal vrijheidsgraden bedraagt in het algemeen:

$$df. = (\text{aantal geobserveerde frekwenties min één}) \text{ minus } (\text{het aantal onafhankelijk te schatten parameters}) \quad (10)$$

Het aantal geobserveerde frekwenties is in dit geval:

$IJKLMNO - 1 = 2^7 - 1 = 127$. X, de gekombineerde latente variabele VWYZ, heeft $2^4 = 16$ categorieën. Aanreziën de latente waarschijnlijkheden P_t^X tot één sommeren, zijn er 15 te schatten parameters. T.a.v. deze parameters gelden echter nog verdere restricties: de afwezigheid van bepaalde interacties, en wel VZ, VWZ, VWY, VYZ, WYZ, VWYZ. Omdat alle latente variabelen dichotomieën zijn, levert dit zes extra vrijheidsgraden op. Daarnaast moeten er nog een aantal konditionele waarschijnlijkheden geschat worden, zoals $P_{ip}^{\bar{AV}}$. Omdat $P_{ip}^{\bar{AV}}$ alleen afhangt van V en omdat deze konditionele waarschijnlijkheden binnen één latente klasse (categorie) tot één sommeren, geeft dit twee te schatten parameters, bv. $P_{11}^{\bar{AV}}$ en $P_{22}^{\bar{AV}}$.

Voor de indicatoren A t/m F tesamen zijn er dus 12 te schatten parameters. p_{os}^{GZ} hoeft uiteraard niet geschat te worden, maar wordt a priori op 0 of 1 gesteld. Al bij al genomen levert dit dan: $127 - 15 + 6 - 12 = 106$ vrijheidsgraden op ter toetsing van het model.

Het volgende model heeft betrekking op een bekende data-set van Lazarsfeld (1972).

Tabel 1. Lazarsfeld data

A = verwachte verkiezingswinnaar t_1

1 = republikeinen

B = stemintentie t₁

2 = democraten

C = verwachte verkiezingswinnaar t_2

D = stemintentie t_2

A	B	C	D	frekw.	A	B	C	D	frekw.	A	B	C	D	frekw.	A	B	C	D	frekw.
1	1	1	1	59	1	2	1	1	3	2	1	1	1	13	2	2	1	1	6
1	1	1	2	1	1	2	1	2	1	2	1	1	2	0	2	2	1	2	7
1	1	2	1	10	1	2	2	1	0	2	1	2	1	14	2	2	2	1	3
1	1	2	2	0	1	2	2	2	3	2	1	2	2	4	2	2	2	2	69

total 193

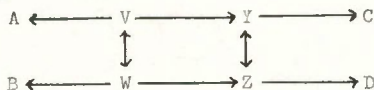
Dit is één van de beroemde "16-fold tables" aan de hand waarvan maten ontwikkeld zijn om de richting van de kausale beïnvloeding vast te stellen. In dit geval: bepaalt de stemintentie, wie men als winnaar verwacht (een soort "wishfull thinking"-effect) dan wel bepaalt de verwachting over wie wint, de stemintentie (een "bandwagon"-effect). Lazarsfeld concludeerde dat het bandwagon-effect sterker was dan het wishfull thinking-effect.

Via loglineaire analyse werd nagegaan in hoeverre de samenhang tussen verwachting en intentie op t_2 verklaard zou kunnen worden vanuit de stabiliteit van deze variabelen in de loop der tijd, in combinatie met het bandwagon- en wishfull thinking-effekt. Uit de analyse zou dan tevens moeten blijken welke van deze laatste twee effecten het sterkste is. Het model {AB, AC, BD, AD, BC} bleek niet houdbaar in het licht van de data. Verdere analyse toonde aan dat met name het direkte (bandwagon) effect AD en het direkte (wishfull thinking) effect BC niet significant waren, terwijl het direkte effect CD, dat geen uitsluitsel geeft over de richting van de kausale beïnvloeding, niet gemist kon worden.

Via latente klasse analyse werd nagegaan of de afwezigheid van de veronderstelde effecten wellicht te wijten was aan onbetrouwbare meting der variabelen. Hoewel de uitkomsten van deze analyse evenmin Lazarsfelds stelling m.b.t. het bestaan van een bandwagon-effect ondersteunen, zullen we één

(passend) model wat nader bespreken en omdat het een wat extreme situatie weergeeft, nl. een model waarin onbetrouwbaarheid in alle manifeste variabelen wordt verondersteld, en omdat van dit model ook de uitkomsten van een "gewone" loglineaire analyse bekend zijn zodat een vergelijking mogelijk is.

figuur 7.



V, W, Y en Z zijn de latente variabelen die overeenkomen met resp. A, B, C en D. Behalve de restricties die onmiddellijk uit het model voortvloeien, werd er ook nog van uitgegaan dat:

$$P_{11}^{\bar{A}V} = P_{22}^{\bar{A}V}, P_{11}^{\bar{B}W} = P_{22}^{\bar{B}W}, P_{11}^{\bar{C}Y} = P_{22}^{\bar{C}Y}, P_{11}^{\bar{D}Z} = P_{22}^{\bar{D}Z}. \quad (11)$$

Deze restrictie is met name ook noodzakelijk om het aantal onbekenden het aantal bekenden niet te laten overtreffen.

Toepassing van het model op tabel 1 gaf als toetsingsuitkomst:

$$L = 4,59 \quad df. = 3 \quad p = 0,20.$$

De konditionele waarschijnlijkheden vermeld in (11) waren als volgt:

$$\hat{P}_{11}^{\bar{A}V} = 0,98; \hat{P}_{11}^{\bar{B}W} = 0,96; \hat{P}_{11}^{\bar{C}Y} = 0,94; \hat{P}_{11}^{\bar{D}Z} = 0,99.$$

Alle indicatoren vormen dus binnen dit model een zeer goede weergave van de onderliggende theoretische variabelen.

De uitkomsten van de loglineaire analyse op de verwachte frekwenties zijn weergegeven in onderstaande tabel, waarin tevens de resultaten staan van de "gewone" loglineaire analyse voor het model $\{AB, AC, BD, CD\}$.

Dit model heeft 7 vrijheidsgraden; de L-waarde bedraagt 7,75; $p = 0,35$.

Indien we de beide modellen als hiërarchisch geordend mogen beschouwen, kan het verschil tussen beide getoetst worden:

$$L_1 - L_2 = 7,75 - 4,59 = 3,16 \quad df. = 7 - 3 = 4 \quad p = 0,50.$$

De konklusie zou dan luiden dat invoering van de latente variabelen, in feite van de vier onafhankelijke parameters in (11), geen significante verbetering van het "gewone" model geeft.

Tabel 2. Analyse van de Lazarsfeld data *

Model in figuur 7-latente klasse analyse Model {AB, AC, BD, CD}

$$\tau_{11}^{VA} = 8,10 \quad \tau_{11}^{WB} = 4,81 \quad \tau_{11}^{YC} = 4,07 \quad \tau_{11}^{ZD} = 9,73$$

$$\tau_1^V = 0,57$$

$$\tau_1^W = 0,58$$

$$\tau_1^Y = 0,79$$

$$\tau_1^Z = 3,51$$

$$\tau_{11}^{VW} = 2,11$$

$$\tau_{11}^{VY} = 1,78$$

$$\tau_{11}^{WZ} = 7,00$$

$$\tau_{11}^{YZ} = 2,32$$

$$\tau_1^A = 0,66$$

$$\tau_1^B = 1,06$$

$$\tau_1^C = 0,89$$

$$\tau_1^D = 1,48$$

$$\tau_{11}^{AB} = 1,89$$

$$\tau_{11}^{AC} = 1,62$$

$$\tau_{11}^{BD} = 2,98$$

$$\tau_{11}^{CD} = 1,87$$

* de effecten zijn "stapsgewijs" berekend volgens Goodman, 1973.

Door het invoeren van de latente variabelen hoop men in feite voor de onbetrouwbaarheid in de manifeste variabelen te corrigeren. Onbetrouwbaarheid van de metingen heeft met name afzwakking van de sterkte van de verbanden tot gevolg. Het is dan ook niet zo verwonderlijk dat de relaties tussen de latente variabelen beduidend sterker zijn dan tussen de overeenkomstige manifeste variabelen.

Verder blijken de latente variabelen V, W, Y en Z in het latente model schever verdeeld te zijn dan A, B, C en D in het "gewone" model. Omdat alle variabelen dichotomieën zijn, geldt dus dat de variantie van de latente variabelen kleiner is dan van de manifeste. Ook dit viel te verwachten aangezien de variantie van de manifeste variabelen opgebouwd is uit de variantie van de "ware" score plus de variantie van de "toevalsterm".

Het laatste voorbeeld van deze paragraaf is wellicht het meest realistische. Het is niet gebaseerd op gefingeerde data en er wordt niet van uitgegaan dat alle variabelen onbetrouwbaar zijn gemeten.

De gegevens in tabel 3 zijn ontleend aan een onderzoek van het CBS (1979) onder leerlingen die in 1977 voor de eerste maal tot het eerste leerjaar van een school voor VWO, HAVO, MAVO, of LBO werden toegelaten, en zijn ter

beschikking gesteld door het Steinmetzarchief. De analyse van deze data is een zeer voorlopige, met name door de 'noodzaak' de variabelen te dichotomiseren.

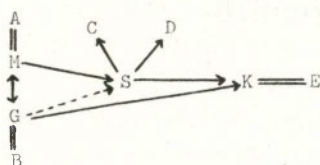
Tabel 3. CBS-data schoolkeuze

A = sociaal milieu 1 = hoog 2 = laag
 B = geslacht 1 = jongen 2 = meisje
 C = CITO-toets prestatiescore 1 = hoog 2 = laag
 D = advies schoolhoofd 1 = VWO, HAVO, MAVO 2 = overige
 E = schoolkeuze 1 = VWO, HAVO, MAVO 2 = overige

A	B	C	D	E	frekw.	A	B	C	D	E	frekw.	A	B	C	D	E	frekw.	A	B	C	D	E	frekw.
1	1	1	1	1	3698	1	2	1	1	1	2766	2	1	1	1	1	3690	2	2	1	1	1	4201
1	1	1	1	2	36	1	2	1	1	2	26	2	1	1	1	2	19	2	2	1	1	2	19
1	1	1	2	1	50	1	2	1	2	1	46	2	1	1	2	1	99	2	2	1	2	1	91
1	1	1	2	2	220	1	2	1	2	2	123	2	1	1	2	2	397	2	2	1	2	2	377
1	1	2	1	1	571	1	2	2	1	1	578	2	1	2	1	1	795	2	2	2	1	1	987
1	1	2	1	2	13	1	2	2	1	2	13	2	1	2	1	2	14	2	2	2	1	2	25
1	1	2	2	1	65	1	2	2	2	1	65	2	1	2	2	1	185	2	2	2	2	1	119
1	1	2	2	2	414	1	2	2	2	2	377	2	1	2	2	2	1179	2	2	2	2	2	1395
																		totaal					22653

Voor de data in tabel 3 werden diverse modellen uitgetoetst. Daarbij werd steeds de variabele sociaal milieu (A) perfect verbonden met de latente variabele M, werd B beschouwd als een perfecte indikator van de latente variabele G(eslacht) en de variabele E-schoolkeuze als identiek met de latente variabele K. De uitslag op de Cito-toets (C) en het (zonder kennis van de uitslag van de toets gegeven) advies van het schoolhoofd (D) werden als indicatief beschouwd voor de latente variabele schoolgeschiktheid S. De resultaten van een relatief goed passend model zijn in schema 1 bijeen gezet.

Schema 1. Analyse CBS-data*



$$L = 145,50 \quad df. = 20 \quad p < 0,001$$

$$\hat{P}_{11}^{CS} = 0,83 \quad \hat{P}_{22}^{CS} = 0,75 \quad \tau_{11}^{SC} = 1,96$$

$$\hat{P}_{11}^{DS} = 0,99 \quad \hat{P}_{22}^{DS} = 0,99 \quad \tau_{11}^{SD} = 9,37$$

$$\tau_{11}^{HG} = 1,09 \quad \tau_{11}^{HS} = 1,23 \quad \tau_{11}^{GS} = 0,99$$

$$\tau_{11}^{GK} = 1,07 \quad \tau_{11}^{SK} = 6,03$$

* De effecten zijn stapsgewijs berekend volgens Goodman, 1973.

Op basis van de L-toets zou het model in schema 1 verworpen moeten worden. Om diverse redenen, die we hier slechts zeer kort willen vermelden, wordt het model toch acceptabel geacht. Dit model past in ieder geval t.o.v. andere modellen relatief goed bij de data. Verder zijn de residuen ($f - \hat{F}$) over het algemeen zeer klein. Bovendien is bij zeer grote aantallen, zoals hier $N = 22.653$, elke kleine afwijking significant en dreigt het gevaar van "overfitting" (zie ook Zahn en Fein, 1979). Een vergelijking tenslotte van de (cruciale) frekwenties f_{ij}^{CD} met $\hat{F}_{ij}^{CD}$ liet een vrijwel perfecte overeenkomst zien: $L = 0,0002$.

Kijken we nu naar de overige resultaten. Variabele D, het advies van het hoofd, is een vrijwel perfecte indikator van de latente variabele schoolgeschiktheid i.t.t. de Cito-toets, die in veel geringere mate met de latente variabele samenhangt. Bij de interpretatie van de latente variabele S moet dit uiteraard voor ogen gehouden worden.

De latente variabele schoolgeschiktheid blijkt een zéér grote directe invloed uit te oefenen op de schoolkeuze. Interessanter wellicht zijn de invloeden van geslacht en milieu op de keuze. Geslacht oefent een niet al te sterke directe invloed uit op de schoolkeuze, maar (vrijwel) geen indirecte invloed via schoolgeschiktheid. Jongens hebben een wat "hogere" schoolkeuze dan meisjes, maar dit verband komt niet tot stand via een verschillende schoolgeschiktheid van jongens en meisjes.

Bij milieu is de situatie in zekere zin precies omgekeerd. Het afkomstig zijn uit een hoger sociaal milieu, leidt tot een grotere schoolgeschiktheid en dit laatste leidt weer tot een "hogere" schoolkeuze. Het verband tussen milieu en schoolgeschiktheid is daarbij vrij sterk. Buiten de faktor schoolgeschiktheid om, heeft milieu echter geen verdere directe invloeden op de schoolkeuze.

Het lijkt alleszins de moeite waard deze conclusies ook nader te bekijken voor de oorspronkelijke, niet gedichotomiseerde variabelen.

6. Evaluatie

Haberman, 1979, blz. 558 geeft een algemene versie van de iteratieve proportional fitting procedure voor de oplossing van elk model met latente variabelen dat geformuleerd kan worden als een hiërarchisch loglineair model. In het algemeen convergeert dit algoritme voor modellen met latente

variabelen aanzienlijk langzamer dan voor modellen met alleen manifeste variabelen.

De "specifieke" variant die Goodman van dit algoritme ontwikkelde, heeft als belangrijkste voordeel dat op zeer eenvoudige wijze allerlei konditionele waarschijnlijkheden aan elkaar gelijk kunnen worden gesteld, met name dan ook gelijkheden die niet onmiddellijk uit het model volgen maar om additionele redenen aangebracht worden (zoals in (11)).

Haberman prefereert een alternatieve oplossingsmethode, een variant van het Newton-Raphson algoritme, het "scoring"-algoritme genoemd. De convergentie van deze procedure is veel sneller; restricties ten aanzien van de loglineaire effecten zijn eenvoudig aan te brengen; de variantie-covariantiematrix van de geschatte loglineaire effecten wordt "automatisch" berekend. Kortom, deze procedure zal meestal de voorkeur verdienen.

Toch zal hieronder nog kort ingegaan worden op de variant van Goodman, omdat deze procedure meer algemeen bekend en toegepast is en omdat de in dit artikel gepresenteerde modellen via de door Goodman aangegeven lijnen ontwikkeld zijn. (Een uitvoerigere beschrijving van zijn algoritme is te vinden in Goodman, 1974a en 1974b en Hagenaars, 1976.)

De procedure begint met startwaarden voor de modelparameters P_t^X , $P_{it}^{\bar{X}}$, $P_{jt}^{\bar{X}}$, $P_{kt}^{\bar{X}}$, $P_{lt}^{\bar{X}}$. Deze startwaarden moeten uiteraard tussen 0 en 1 liggen en waar nodig tot één sommeren. Met behulp van deze startwaarden berekend men $\hat{P}_{ijkl}^{ABCDX}$ op grond van vergelijking (2); daarna $\hat{P}_{ijkl}^{ABCD}$ (zie (3)); vervolgens $\hat{P}_{ijkl}^{ABCDX}$ m.b.v. (7). Via deze grootheden en de manifeste proporties P_{ijkl}^{ABCD} berekent men nieuwe schattingen voor de modelparameters:

$$\begin{aligned} \hat{P}_t^X &= \sum_{i,j,k,l} P_{ijkl}^{ABCD} \hat{P}_{ijkl}^{ABCDX} \\ \hat{P}_{it}^{\bar{X}} &= \left(\sum_{i,j,k,l} P_{ijkl}^{ABCD} \hat{P}_{ijkl}^{ABCDX} \right) / \hat{P}_t^X \\ \hat{P}_{jt}^{\bar{X}} &= \left(\sum_{i,j,k,l} P_{ijkl}^{ABCD} \hat{P}_{ijkl}^{ABCDX} \right) / \hat{P}_t^X \\ \text{en analoge formules voor } \hat{P}_{kt}^{\bar{X}} \text{ en } \hat{P}_{lt}^{\bar{X}}. \end{aligned} \quad (12)$$

Indien er verder geen restricties aan de parameters aangebracht hoeven te worden, substitueert men de nieuwe schattingen voor de oude en herhaalt de procedure tot de schattingen niet meer (noemenswaard) veranderen. Indien er wel restricties aangebracht moeten worden, hetgeen voor de meeste modellen het geval zal zijn, dan dient dit eerst te gebeuren.

Bij konditionele kansen die aan elkaar gelijk moeten zijn, wordt als nieuwe

startwaarde het gewogen gemiddelde van de gevonden waarden gebruikt, waarbij de weging geschiedt via de in de noemer voorkomende waarde $\hat{P}_t^X$. Zo ook kunnen de konditionele kansen aan een bepaalde konstante, bv. 0,50 gelijk worden gesteld door na afloop van de berekeningen in (12) als nieuwe startwaarde steeds deze konstante in te voeren. De procedure is verder dusdanig dat het hanteren van 0 of 1 als startwaarde tot gevolg heeft dat deze waarden tijdens de hele procedure behouden blijven. Hoewel Goodman deze mogelijkheid niet noemt, zou het zo ook mogelijk moeten zijn bepaalde ordinale restrikties aan de parameters aan te brengen, zoals bv. bij een Guttman-schaal vereist is.

De toevoeging die met name in de vorige paragraaf gehanteerd is, heeft betrekking op het aanbrengen van restrikties t.a.v. $\hat{P}_t^X$. Na het berekenen van (12) en het aanbrengen van de eerder genoemde restrikties, worden de schattingen $\hat{P}_t^X$ dusdanig gewijzigd dat de interacties die in het model afwezig worden verondersteld, ook niet aanwezig zijn in de verdeling van de (gecombineerde) latente variabele X. Dit kan bereikt worden door een (additionele) proportional fitting procedure waarbij de nieuwe schattingen moeten voldoen aan bepaalde "marginalen" van de oude schattingen (zie (9)). Daarbij kunnen zowel de variant van de multiële regressie als van de padanalyse gebruikt worden (Goodman, 1972 en 1973). De rechtvaardiging van het feit dat deze restrikties aangebracht worden voor de frekwenties $\hat{F}_t^X$ en niet bij $\hat{F}_{ijklt}^{ABCDX}$ is gelegen in het zgn. "collapsibility"-theorema (Bishop e.a., 1975, Hagenaars en Heinen, 1980).

Welk algoritme uiteindelijk de voorkeur verdient, heeft vooral te maken met de snelheid van convergeren, maar ook met het gemak waarmee bepaalde restrikties in het model aangebracht kunnen worden. (Voor andere, alternatieve algoritmen, zie verder ook nog Goodman, 1979)

Een belangrijk probleem bij de analyse van modellen met latente variabelen, welk oplossingsalgoritme men ook kiest, is de identificeerbaarheid. Goodman, 1974b geeft de voldoende en noodzakelijke voorwaarden aan. De noodzakelijke voorwaarde is uiteraard dat het aantal bekenden, de manifeste frekwenties, niet geringer is dan het aantal onbekenden, de te schatten parameters. De voldoende voorwaarde is minder eenvoudig weer te geven. Een praktisch uitvoerbaar alternatief is het uitvoeren van de iteratieve procedure met verschillende beginschattingen. In het geval van een unieke oplossing zullen deze doorgaans dezelfde oplossing geven; in het geval van onderidentifikatie zullen doorgaans verschillende oplossingen gevonden worden die alle even goed

bij de data passen (dezelfde L-waarde geven).

Een verder probleem, dat echter inherent is aan elke vorm van "tabelanalyse", is het grote aantal cellen voor zelfs een model met niet al te veel variabelen. De moeilijkheid is dan vooral gelegen in de hantering van L of χ^2 ter toetsing van de adekwaatheid van het model. Een eenduidige oplossing voor dit probleem lijkt niet zo voor de hand te liggen.

Ondanks deze problemen biedt loglineaire analyse van modellen met latente variabelen interessante perspectieven. Hoewel diverse oplossingsalgoritmen al vele jaren bekend zijn, is het aantal toepassingen nog zeer gering; uiteraard hangt de definitieve beoordeling van de waarde van deze werkwijze af van zijn toepasbaarheid in konkreet empirisch onderzoek.

Literatuur

- Fielding A. (1977), Latent Structure Models, in: C.A. O'Muircheartaigh en C.Payne (eds), The Analysis of Survey Data; Volume I Exploring Data Structures, London.
- Goodman L.A. (1972), A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables, in: American Sociological Review, jrg. 37, blz. 28-46.
- Goodman L.A. (1973), The Analysis of Multidimensional Contingency Tables when some Variables are Posterior to Others; a Modified Path Analysis Approach, in: Biometrika, jrg. 59, blz. 579-596.
- Goodman L.A. (1974a), The Analysis of Systems of Qualitative Variables when some of the Variables are Unobservable; Part I: a Modified Latent Structure Approach, in: American Journal of Sociology, jrg. 79, blz. 1179-1259.
- Goodman L.A. (1974b), Exploratory Latent Structure Analysis Using both Identifiable and Unidentifiable Models, in: Biometrika, jrg.61, blz. 215-231.
- Goodman L.A. (1975), A New Model for Scaling Response Patterns: an Application of the Quasi-Independence Concept, in: Journal of the American Statistical Association, jrg. 70, blz. 755-768.
- Goodman L.A. (1979), Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories, in: Journal of the American Statistical Association, jrg. 74, blz. 537-552.
- Haberman S.J. (1979), Analysis of Qualitative Data; volume 2 New Developments, New York.
- Hagenaars J.A.P. (1976), Latente Variabelen, in: Sociale Wetenschappen, jrg. 19, blz.250-278.
- Hagenaars J.A.P. (1978), Latent Probability Models with Direct Effects Between Indicators, in Quality and Quantity, jrg.12, blz.205-222.

- Hagenaars J.A.P. en Heinen A.G.J. (1980), Analyse van Contingentietabellen met behulp van het Loglineaire Model, in: Segers J.G. en Hagenaars J.A., Sociologische Onderzoeksmethoden; deel 2 Technieken van Kausale Analyse, Assen.
- Jöreskog K.G. (1973), A general Method for Estimating a Linear Equation System, in: A.S. Goldberger en O.D. Duncan, Structural Equation Models in the Social Sciences, blz.85-113.
- Lazarsfeld P.F. en Henry N.W. (1968), Latent Structure Analysis, New York
- Saris W.E. (1974), De LISREL-Procedure voor het Analyseren van Lineaire Structurele Modellen met Ongemeten Variabelen, in: Sociale Wetenschappen, jrg. 17, blz. 124-146.
- Saris W.E. (1979), Methods and Techniques for the Analysis of Survey Data with Measurement Error, Amsterdam.
- Zahn D.A. en Fein S.B. (1979), Large Contingency Tables with Large Cell Frequencies: a Model Search Algorithm and Alternative Measures of Fit, in: Psychological Bulletin, jrg.86, blz. 1189-1200.