Padanalyse: uitgangspunten en basisbegrippen

J. Dessens
W. Jansen
P. G. Swanborn*

Object van onderzoek in de sociologie vormen mensen, groepen, instituties, systemen; ook wel eenheden genoemd. Proposities of hypothesen worden geformuleerd in termen van kenmerken van die eenheden: proposities relateren kenmerken van eenheden aan elkaar (1). Als voorbeeld de propositie: 'In groep A gaat normloosheid X (gemeten via een anomie-schaal) samen met anti-semitisme Y (gemeten via een anti-semitisme-schaal)'. Een dergelijke propositie kan gefalsifieerd worden door de afwezigheid van een empirisch vastgesteld statistisch verband. Veelal worden proposities in causale termen geformuleerd: 'normloosheid leidt tot anti-semitisme', of 'anti-semitisme leidt tot normloosheid', of 'anti-semitisme en normloosheid worden veroorzaakt door een autoritaire instelling'. Proposities op hun beurt vormen de bouwstenen voor een theorie. Een theorie kan dan ook omschreven worden als een stelsel samenhangende causale proposities, waarvan er enkele empirisch toetsbaar zijn.

Padanalyse stelt ons in staat proposities in hun samenhang (theorie(fragmenten)) middels een formeel model te representeren, waarbij in het algemeen toetsing van verschillende aspecten van het model aan de empirie mogelijk is. Anderzijds is het ontoetsbare, geassumeerde gedeelte van een padmodel in het algemeen vrij omvangrijk; de empirie is m.a.w. maar in zeer beperkte mate in staat de assumpties van de onderzoeker te verifiëren of

* De auteurs studeerden allen af aan de Rijksuniversiteit Utrecht. Thans zijn zij werkzaam binnen de vakgroep Theorie en Methodologie van het Sociologisch Instituut Utrecht: J. Dessens en W. Jansen als wetenschappelijk medewerker en P. G. Swanborn als lector.

te falsifiëren. Enig wantrouwen wordt ook gevoed door de zeer geringe percentages verklaarde variantie in de gepubliceerde toepassingen. Een evaluatie van de mogelijkheden van de padanalyse - één van de momenteel meest populaire technieken van multivariate analyse - is echter eerst mogelijk na een grondige studie van de achtergronden en procedures. In dit artikel willen wij hiermee een begin maken.

De padanalyse-techniek is ontwikkeld in de genetica door Sewall Wright (1921, 1934, 1960) en Li (1956), in de sociale wetenschappen geïntroduceerd door Boudon (1965) en Blalock (1967), waarbij met name wat betreft het identificatieprobleem door deze laatste auteurs teruggegrepen wordt op de econometrische literatuur. Vanaf het einde van de zestiger jaren hebben hierop voortgebouwd o.a. Duncan (1966), Land (1969) en Heise (1969).

De assumpties, die ten grondslag liggen aan de padanalyse, zijn, dat de variabelen alle op interval- of rationiveau zijn gemeten; dat de verbanden lineair zijn en monocausaal; dat er geen sprake is van interactie-effecten, en dat de restvariabelen van de endogene variabelen ongecorreleerd zijn met de exogene variabelen (2).

In dit artikel willen wij de theorie en de techniek van de padanalyse met gebruikmaking van deze assumpties aan de orde stellen. Daarbij voeren we ook nog de vereenvoudiging in, dat we geen onderscheid maken tussen steekproefgegevens en populatieparameters. In feite werken we immers meestal met steekproefgegevens (wat in een meer exacte verhandeling in de notatie tot uitdrukking zou moeten komen), waarbij de schatting van de populatieparameters een apart probleem vormt.

In een volgend artikel willen we op enkele complicaties, verband houdende met tweezijdige causaliteit, meermomentopnamen en de problematiek van de ongemeten variabelen, ingaan.

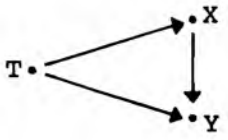
De specificatie van een theorie

Zeer vereenvoudigd zou men specificatie of opstelling van een theorie kunnen definiëren als het expliciteren van de kenmerken of variabelen binnen de theorie en, voor elk paar van variabelen, het aangeven of er binnen dat paar een directe causale beïnvloeding is, en zo ja, in welke richting. Ter illustratie gaan we uit van de volgende theorie, bestaande uit drie proposities:

propositie 1: T veroorzaakt X
propositie 2: T veroorzaakt Y
propositie 3: X veroorzaakt Y

Met deze drie proposities hebben we een theorie gespecificeerd. Een dergelijke theorie kunnen we grafisch weergeven, waarbij we een causale invloed aangeven met een pijl, met een pijlpunt in de richting van de veroorzaakte variabele, zoals in diagram 1:

Diagram 1



Een diagram is een netwerk, waarbij de variabelen worden voorgesteld door punten (in deze tekst staan veelal i.p.v. punten letters, die de variabelen aanduiden) en de proposities door rechte en gebogen pijlen voor resp. causale en niet-causale relaties. Gewoonlijk wordt bij de gebogen pijlen de correlatiecoëfficiënt vermeld, en bij de rechte pijlen een vooralsnog ongedefinieerde coëfficiënt, die we later padcoëfficiënt zullen noemen.

Bovenstaande theorie impliceert een reductie t.o.v. de logische mogelijkheden, omdat binnen elk paar variabelen twee pijlrichtingen denkbaar zijn, alsmede de afwezigheid van enigerlei causale invloed (geen pijl). Dit betekent bij drie variabelen dus reeds $3^3 = 27$ mogelijke diagrammen, en bij n variabelen $3^{\binom{n}{2}}$ logische mogelijkheden, aangezien $\binom{n}{2}$ mogelijke paren te vormen zijn.

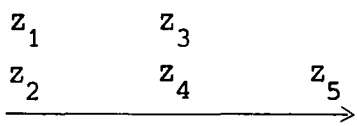
Hiermee is tevens geïllustreerd, hoe cruciaal de specificatie, vooraf, van een theorie is. Bij een theorie als door diagram 1 gerepresenteerd is het niet mogelijk om op basis van de empirisch gevonden correlaties te onderscheiden tussen dit diagram en een diagram waarin een of meer pijlen van richting zijn veranderd. Toetsing van de causaliteitsrichting is niet mogelijk; hooguit kan men aantonen, dat causaliteit afwezig is. Om deze reden wordt het falsificatieprincipe gehanteerd, d.w.z. we trachten zo veel mogelijk alternatieve theorieën te verwerpen, en die theorie als voorlopig juist te aanvaarden, die falsificatie zo goed mogelijk doorstaan heeft. Om een theorie te kunnen verwerpen is confrontatie met de empirie noodzakelijk. Is dit niet mogelijk, dan is falsificatie eveneens onmogelijk, en kan niets met zekerheid omtrent de realiteitswaarde van de theorie worden gezegd. Er zou wel een toetsingsmogelijkheid aanwezig zijn, indien we uitgaan van de veronderstelling, dat de relatie tussen X en Y niet d i r e c t causaal is. We stellen ons in dat geval 'kwetsbaarder'

op dan in de situatie, waarin alle pijlen zijn getrokken. Deze reductie van de complexiteit van de werkelijkheid in de theorie, die in een diagram tot uitdrukking komt middels de afwezigheid van één pijl, levert één voorspelling op, die, zoals bekend, neerkomt op een verwachte nulwaarde van de partiële correlatie tussen X en Y. Ook nu hebben we de gereduceerde theorie niet 'bewezen'; b.v. ook het diagram X $\longrightarrow$ T $\longrightarrow$ Y correspondeert met de voorspelling. Wel is het zo, dat een aantal mogelijkheden is uitgeschakeld.

We geven vervolgens een iets gecompliceerder voorbeeld. Hierin zijn vijf variabelen, alle in standaardscores gemeten (3), betrokken:

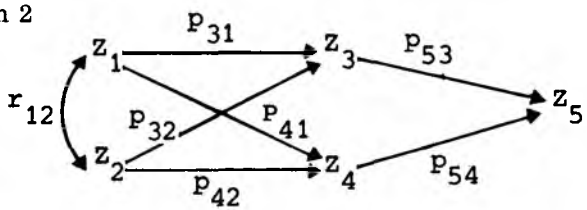
Z_1: opleiding
Z_2: leeftijd
Z_3: score op een F-schaal
Z_4: score op een anomie-schaal
Z_5: score op een anti-semitisme-schaal

Min of meer arbitrair kunnen we deze variabelen qua volgorde in de tijd als volgt ordenen:



Verder veronderstellen we, dat de variabelen Z_1 en Z_2 een directe invloed uitoefenen op Z_3 en Z_4, doch niet op Z_5; de variabelen Z_3 en Z_4 oefenen op hun beurt een directe invloed uit op Z_5, doch niet op elkaar. Weergegeven in een diagram:

Diagram 2



Stel nu, dat we geïnteresseerd zijn in de mate, waarin we de score op een F-schaal (Z_3) kunnen schatten vanuit onze kennis van de scores op de variabelen 'opleiding' (Z_2) en 'leeftijd' (Z_1).

We kunnen ons dan afvragen, of het mogelijk is de coëfficiënten p_{31} en p_{32} zodanig te bepalen, dat voor alle respondenten tegelijkertijd geldt, dat p_{31} maal de score op Z_1 plus p_{32} maal de score op Z_2 een zo goed mogelijke benadering is van de score op Z_3 (4).

Voorlopig in het midden gelaten op welke wijze de coëfficiënten p_{31} en p_{32} kunnen worden bepaald, is de factor $(p_{31} Z_1 + p_{32} Z_2) = \hat{Z}_3$ op te vatten als een schatting van de score op Z_3.

De gebruikelijke 'least-squares' uitdrukking $\Sigma (Z_3 - \hat{Z}_3)^2$ is op te vatten als de mate waarin de schatting van de scores op Z_3 vanuit de scores op Z_1 en Z_2 afwijkt van de werkelijke score Z_3. Zo kunnen we ook $\Sigma (Z_4 - \hat{Z}_4)^2$ en $\Sigma (Z_5 - \hat{Z}_5)^2$ opvatten als de mate waarin Z_4 niet vanuit Z_1 en Z_2 resp. Z_5 niet uit Z_3 en Z_4 is te schatten.

Hoe kunnen we deze afwijkingen in de schattingen verklaren? Ten eerste is het op grond van de multipele bepaaldheid van sociaal gedrag niet plausibel, dat een tweetal variabelen een derde, afhankelijke variabele, exact bepaalt. M.a.w. er zijn andere, al dan niet conceptualiseerbare variabelen, die tesamen met deze twee variabelen wèl de afhankelijke variabele exact bepalen. Deze, niet in de theorie opgenomen, variabelen noemen we restvariabelen. In de geformaliseerde weergave van de theorie, het model, introduceren we daarom een restvariabele e_i, die betrekking heeft op de variabele Z_i. In het algemeen trachten we zoveel mogelijk de belangrijkste variabelen in de theorie op te nemen, d.w.z. we proberen de invloed van de restvariabelen zo gering mogelijk te doen zijn. De totale invloed van de restvariabele e_i op de variabele Z_i drukken we uit in de coëfficiënt p_{ie_i}. Hoe groter p_{ie_i}, hoe groter de gemiddelde schattingsfout.

Ten tweede kunnen afwijkingen tussen de score Z_i en de schatting van de score Z_i het gevolg zijn van meetfouten. Dit brengt geen extra complicaties met zich mee, omdat deze meetfouten ook ondergebracht kunnen worden in e_i.

We kunnen nu de score op Z_3 als volgt opgebouwd denken:

$$Z_3 = p_{31} Z_1 + p_{32} Z_2 + p_{3e_3} e_3$$

Ook zien we met behulp van de eerder gegeven schatting:

$$\hat{Z}_3 = p_{31} Z_1 + p_{32} Z_2, \text{ dat } (Z_3 - \hat{Z}_3) = p_{3e_3} e_3$$

Voor de afhankelijke variabelen Z_4 en Z_5 volgt op analoge wijze:

$$Z_4 = p_{41} Z_1 + p_{42} Z_2 + p_{4e_4} e_4$$

$$Z_5 = p_{53} Z_3 + p_{54} Z_4 + p_{5e_5} e_5$$

Hiermee is geïllustreerd, dat de formele gedaante van een theorie niet alleen grafisch via een diagram (paddiagram genoemd), maar ook mathematisch via een stelsel van zgn. structuurvergelijkingen (padmodel genoemd) gerepresenteerd kan worden.

Wanneer er in de theorie alleen sprake is van monocausale relaties (m.a.w. geen wederzijdse causaliteit) gebruiken we ter representatie een zgn. recursief stelsel van structuurvergelijkingen. Een dergelijk stelsel heeft de volgende gedaante:

$$Z_1 = p_{1e_1} e_1$$

$$Z_2 = p_{21} Z_1 + p_{2e_2} e_2$$

$$Z_3 = p_{31} Z_1 + p_{32} Z_2 + p_{3e_3} e_3$$

$$Z_4 = p_{41} Z_1 + p_{42} Z_2 + p_{43} Z_3 + p_{4e_4} e_4$$

. . . . . .

$$Z_n = p_{n1} Z_1 + p_{n2} Z_2 + \ldots + p_{n(n-1)} Z_{n-1} + p_{ne_n} e_n$$

Hierin zijn:

Z_1 tot en met Z_n: naar standaardscores getransformeerde inhoudelijke variabelen, die expliciet in de theorie zijn opgenomen.

p_{21} tot en met $p_{n(n-1)}$ en p_{1e_1} tot en met p_{ne_n}: padcoëfficiënten, die de directe invloed van Z_j respectievelijk e_i op Z_i aangeven, waarbij alle indirecte effecten van Z_j via alle andere variabelen binnen de theorie worden uitgesloten. Wanneer er, zoals in het bovengegeven stelsel, sprake is van monocausale relaties, geldt: indien $p_{ij} \neq 0$ dan $p_{ji} = 0$.

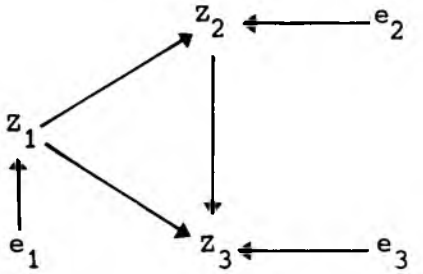
e_1 tot en met e_n: restvariabelen, die de invloed van niet in de theorie opgenomen variabelen op Z_1 tot en met Z_n representeren.

Het identificatieprobleem

De specificatie van een theorie via een vergelijkingenstelsel moet zodanig zijn, dat oplossing of identificering van de onbekende coefficiënten p_{ij} uit het stelsel structuurvergelijkingen mogelijk is; we hebben hier te maken met het zgn. identificatieprobleem. Gezien de definiëring van de coëfficiënten p_{ij} zou men kunnen veronderstellen, dat deze coëfficiënten identiek zijn aan de partiële gestandaardiseerde regressiecoëfficiënten $\beta_{ij.k..n}$. Deze β's

geven immers aan in welke mate de afhankelijke variabele Z_i verandert tengevolge van een verandering van één eenheid standaarddeviatie in de onafhankelijke variabele Z_j onder eliminering van de invloed van de overige variabelen. Om nu een antwoord te kunnen geven op de vraag of de p's identiek zijn aan de β's (en dus uit de empirische gegevens berekend zouden kunnen worden zonder ons verder om vergelijkingenstelsels te bekommeren) beschouwen we Diagram 3 en het daarbij behorende vergelijkingenstelsel:

Diagram 3



$Z_1 = p_{1e_1} e_1$

$Z_2 = p_{21} Z_1 + p_{2e_2} e_2$

$Z_3 = p_{31} Z_1 + p_{32} Z_2 + p_{3e_3} e_3$

Per definitie is nu $p_{23} = 0$, omdat $p_{32} \neq 0$. Hieruit volgt $p_{23} . p_{32} = 0$. Echter $\beta_{23.1} \cdot \beta_{32.1} = r^2_{23.1}$ (het kwadraat van de partiële correlatiecoëfficiënt). Daar de partiële correlatiecoëfficiënt in het algemeen ongelijk nul zal zijn, is het derhalve niet gerechtvaardigd het door ons gehanteerde model te identificeren met het algemene lineaire regressiemodel. De coëfficiënten p_{ij} zullen dus uit het vergelijkingenstelsel moeten worden bepaald. We lichten deze procedure toe aan de hand van het bij diagram 3 behorend vergelijkingenstelsel.

Vermenigvuldigen we de eerste vergelijking met e_1, sommeren we over alle waarnemingen en delen we door N dan krijgen we:

$\frac{\Sigma Z_1 e_1}{N} = p_{1e_1} \frac{\Sigma e_1 e_1}{N}$ waarin $\frac{\Sigma Z_1 e_1}{N} = r_{1e_1}$ (5) en

$\frac{\Sigma e_1 e_1}{N} = 1$ en p_{1e_1} de onbekende padcoëfficiënt voorstelt, die opgelost moet worden.

Hieruit volgt algemeen $r_{ie_i} = p_{ie_i}$. Echter $r_{1e_1} = 1$, omdat Z_1 op een constante vermenigvuldigingsfactor na (p_{1e_1}) gelijk is aan e_1. Algemeen geldt, indien Z_i een variabele is, die slechts door restvariabelen wordt beïnvloed, dat $p_{ie_i} = 1$. Dergelijke variabelen noemen we e x o g e n e variabelen. We schrijven veelal $Z_i = e_i$, vanwege het feit, dat de padcoëfficiënt van e_i gelijk 1 is. Variabelen, die niet alleen door restvariabelen worden beïnvloed noemen we e n d o g e n e variabelen. De onderscheiding exogeen/endogeen correspondeert ruwweg met wat in de wandeling onafhankelijk/afhankelijk wordt genoemd.

De tweede vergelijking behorend bij diagram 3 levert wat meer moeilijkheden. Indien we de vergelijking achtereenvolgens doorvermenigvuldigen met Z_1, Z_2 en e_2, sommeren over alle waarnemingen en delen door N, dan krijgen we:

$$\frac{\Sigma Z_1 Z_2}{N} = p_{21}\frac{\Sigma Z_1 Z_1}{N} + p_{2e_2}\frac{\Sigma Z_1 e_2}{N} \quad \text{ofwel } r_{12} = p_{21} + p_{2e_2} r_{1e_2} \quad (1)$$

$$\frac{\Sigma Z_2 Z_2}{N} = p_{21}\frac{\Sigma Z_1 Z_2}{N} + p_{2e_2}\frac{\Sigma Z_2 e_2}{N} \quad \text{ofwel } 1 = p_{21} r_{12} + p_{2e_2} r_{2e_2} \quad (2)$$

$$\frac{\Sigma Z_2 e_2}{N} = p_{21}\frac{\Sigma Z_1 e_2}{N} + p_{2e_2}\frac{\Sigma e_2 e_2}{N} \quad \text{ofwel } r_{2e_2} = p_{21} r_{1e_2} + p_{2e_2} \quad (3)$$

De vergelijkingen (1)-(3) en de hieronder volgende vergelijkingen (4)-(11) noemen we p a d c o ë f f i c i ë n t -vergelijkingen.

De vergelijkingen (1)-(3) vormen nu drie vergelijkingen in vier onbekenden t.w. p_{21}, p_{2e_2}, r_{1e_2} en r_{2e_2}. Er zijn derhalve geen oplossingen voor deze coëfficiënten mogelijk. We zeggen nu, dat de tweede vergelijking behorend bij diagram 3 ondergeïdentificeerd is. We zullen laten zien, hoe via het invoeren van de assumptie, dat Z_1 ongecorreleerd is met e_2, de vergelijking alsnog kan worden geïdentificeerd. Invulling van $r_{1e_2} = 0$ in (1)-(3) geeft:

$$r_{12} = p_{21} + 0 \quad (4)$$

$$1 = p_{21} r_{12} + p_{2e_2} r_{2e_2} \quad (5)$$

$$r_{2e_2} = 0 + p_{2e_2} \quad (6)$$

Uit de vergelijkingen (5) en (6) volgt:

$$1 = p_{21} r_{12} + p^2_{2e_2} \qquad (7)$$

Uit de vergelijkingen (4) en (7) kunnen we nu de twee onbekenden p_{21} en p_{2e_2} oplossen: $p_{21} = r_{12}$ en $p_{2e_2} = \sqrt{1 - r^2_{12}}$.

We zien hieruit, dat indien aan te nemen valt, dat Z_1, zijnde een oorzaak van Z_2, ongecorreleerd is met e_2, identificatie kan worden bewerkstelligd. We zullen nu nagaan of het bovenstaande ook voor de derde vergelijking behorend bij diagram 3 opgaat. Vermenigvuldigen we de vergelijking achtereenvolgens met Z_1, Z_2, Z_3 en e_3, dan geeft dit na sommering over alle waarnemingen en delen door N:

$$\frac{\Sigma Z_3 Z_1}{N} = p_{31} \frac{\Sigma Z_1 Z_1}{N} + p_{32} \frac{\Sigma Z_1 Z_2}{N} + p_{3e_3} \frac{\Sigma Z_1 e_3}{N} \quad \text{ofwel}$$

$$r_{13} = p_{31} + p_{32} r_{12} + p_{3e_3} r_{1e_3}$$

$$\frac{\Sigma Z_2 Z_3}{N} = p_{31} \frac{\Sigma Z_1 Z_2}{N} + p_{32} \frac{\Sigma Z_2 Z_2}{N} + p_{3e_3} \frac{\Sigma Z_2 e_3}{N} \quad \text{ofwel}$$

$$r_{23} = p_{31} r_{12} + p_{32} + p_{3e_3} r_{2e_3}$$

$$\frac{\Sigma Z_3 Z_3}{N} = p_{31} \frac{\Sigma Z_1 Z_3}{N} + p_{32} \frac{\Sigma Z_2 Z_3}{N} + p_{3e_3} \frac{\Sigma Z_3 e_3}{N} \quad \text{ofwel}$$

$$1 = p_{31} r_{13} + p_{32} r_{23} + p_{3e_3} r_{3e_3}$$

$$\frac{\Sigma Z_3 e_3}{N} = p_{31} \frac{\Sigma Z_1 e_3}{N} + p_{32} \frac{\Sigma Z_2 e_3}{N} + p_{3e_3} \frac{\Sigma e_3 e_3}{N} \quad \text{ofwel}$$

$$r_{3e_3} = p_{31} r_{1e_3} + p_{32} r_{2e_3} + p_{3e_3}.$$

We hebben hier vier vergelijkingen in zes onbekenden t.w. p_{31}, p_{32}, p_{3e_3}, r_{1e_3}, r_{2e_3} en r_{3e_3}. Evenals bij de tweede vergelijking van het stelsel hebben we hier te maken met een geval van onderidentificatie. Nemen we nu aan, dat Z_1 ongecorreleerd is met e_3, dan krijgen we:

$$r_{13} = p_{31} + p_{32} r_{12} + 0$$

$$r_{23} = p_{31} r_{12} + p_{32} + p_{3e_3} r_{2e_3}$$

$$1 = p_{31} r_{13} + p_{32} r_{23} + p_{3e_3} r_{3e_3}$$

$$r_{3e_3} = 0 + p_{32} r_{2e_3} + p_{3e_3}$$

Het aantal onbekenden is door invoering van deze assumptie gereduceerd tot vijf t.w. p_{31}, p_{32}, r_{2e_3}, p_{3e_3} en r_{3e_3}. Ook indien we de veronderstelling zouden invoeren, dat in plaats van Z_1 de variabele Z_2 ongecorreleerd is met e_3, blijven we zitten met vier vergelijkingen in vijf onbekenden. Eerst indien we beide assumpties tegelijkertijd invoeren, kunnen we de onbekende coefficiënten bepalen:

$$r_{13} = p_{31} + p_{32} r_{12} + 0 \tag{8}$$

$$r_{23} = p_{31} r_{12} + p_{32} + 0 \tag{9}$$

$$1 = p_{31} r_{13} + p_{32} r_{23} + p_{3e_3} r_{3e_3} \tag{10}$$

$$r_{3e_3} = 0 + 0 + p_{3e_3} \tag{11}$$

Uit de vergelijkingen (8) en (9) kunnen we nu p_{31} en p_{32} bepalen:

$$p_{31} = \frac{r_{13} - r_{12} r_{23}}{1 - r_{12}^2} \qquad \text{en} \qquad p_{32} = \frac{r_{23} - r_{12} r_{13}}{1 - r_{12}^2}$$

Uit de vergelijking (11) volgt $p_{3e_3} = r_{3e_3}$. Substitueren we dit samen met de gevonden waarden voor p_{31} en p_{32} in de vergelijking (10), dan vinden we:

$$1 = \left(\frac{r_{13} - r_{12} r_{23}}{1 - r_{12}^2}\right) r_{13} + \left(\frac{r_{23} - r_{12} r_{13}}{1 - r_{12}^2}\right) r_{23} + p_{3e_3}^2 \quad \text{ofwel na uitwerking:}$$

$$p_{3e_3} = \sqrt{1 - \frac{r_{13}^2 + r_{23}^2 - 2 r_{12} r_{23} r_{13}}{1 - r_{12}^2}}$$

We kunnen nu een tweetal opmerkingen maken:
I. Gegeven de volgende algemene vergelijking:
$Z_i = p_{i1} Z_1 + p_{i2} Z_2 + \cdots p_{i(i-1)} Z_{i-1} + p_{ie_i} e_i$
kunnen we tot de volgende algemene stelling komen: indien $r_{1e_i} =$

$r_{2e_i} = \ldots = r_{(i-1)e_i} = 0$ dan is identificatie van de vergelijking van Z_i mogelijk. Dit valt als volgt in te zien: doorvermenigvuldiging van de vergelijking van Z_i met respectievelijk Z_1, Z_2, ... Z_{i-1} en sommeren over het aantal waarnemingen geeft een stelsel van (i - 1) vergelijkingen in (i - 1) onbekenden, t.w. p_{i1}, p_{i2}, ... $p_{i(i-1)}$. Na oplossing van dit stelsel kunnen we met behulp van de opgeloste coëfficiënten de coëfficiënt p_{ie_i} bepalen uit de twee vergelijkingen, die ontstaan uit doorvermenigvuldigen van de vergelijking van Z_i met Z_i zelf en e_i. De hierboven gehanteerde procedure van doorvermenigvuldiging van een vergelijking met variabelen die ongecorreleerd zijn met de restvariabele in die vergelijking staat in de econometrie bekend als de 'instrumentele variabelen'-procedure. Uit het bovenstaande volgt zonder verdere moeilijkheden, dat indien we aannemen, dat tussen bepaalde variabelen geen directe causale relaties bestaan, het aantal onbekende coëfficiënten kleiner is dan (i - 1), aangezien elke weggelaten pijl een onbekende padcoëfficiënt minder geeft. Het aantal vergelijkingen blijft echter wel (i - 1), als gevolg waarvan we voor sommige coëfficiënten p_{ij} meerdere oplossingen zullen vinden. Indien het model juist is gespecificeerd zullen deze verschillende oplossingen bij benadering aan elkaar gelijk moeten zijn; indien de oplossingen aanmerkelijk van elkaar verschillen dienen we het gespecificeerde model te verwerpen. Wanneer er meer vergelijkingen dan onbekenden zijn, spreekt men van overidentificatie.

II. Met betrekking tot de tweede vergelijking uit het door ons gehanteerde stelsel hebben we gevonden dat $p_{21} = r_{12}$. Aangezien $r_{12} = \beta_{21}$ geldt hier dus $p_{21} = \beta_{21}$. Voor de derde vergelijking uit het stelsel vinden we soortgelijke resultaten:

$$p_{31} = \frac{r_{13} - r_{23}\, r_{12}}{1 - r_{12}^2} = \beta_{31.2} \quad \text{en} \quad p_{32} = \frac{r_{23} - r_{13}\, r_{12}}{1 - r_{12}^2} = \beta_{32.1}$$

We zien hieruit, dat onder bepaalde condities (het ongecorreleerd zijn van Z_i met e_j waarbij Z_j geen oorzaak van Z_i mag zijn) de coëfficiënten p_{ij} gelijk zijn aan de partiële gestandaardiseerde regressiecoëfficiënten $\beta_{ij.k..n}$. In aanvulling op een aan het begin van deze paragraaf gemaakte opmerking zien we dus, dat het door ons gehanteerde model en het lineaire regressiemodel tot dezelfde resultaten leiden, althans onder de eerder genoemde con-

dities. We zullen dit resultaat algemeen afleiden. We gaan daarbij uit van:

$$Z_i = p_{i1}Z_1 + p_{i2}Z_2 + \ldots p_{i(i-1)}Z_{i-1} + p_{ie_i}e_i \qquad (12)$$

Beschouwen we allereerst de oplossing via het padmodel. Toepassing van de 'instrumentele variabelen'-procedure veronderstelt:

$$\frac{\Sigma(Z_j e_i)}{N} = 0 \text{ met } j = 1, 2, \ldots (i-1) \qquad (13)$$

Werken we vergelijking (12) om tot:

$$e_i = \frac{Z_i - p_{i1}Z_1 - p_{i2}Z_2 - \ldots p_{i(i-1)}Z_{i-1}}{P_{ie_i}} \text{ en substitueren}$$

we dit in (13):

$$\frac{\Sigma(Z_j)(Z_i - p_{i1}Z_1 - p_{i2}Z_2 - \ldots p_{i(i-1)}Z_{i-1})}{Np_{ie_i}} = 0 \quad \text{ofwel}$$

$$\Sigma(Z_j)(Z_i - p_{i1}Z_1 - p_{i2}Z_2 - \ldots p_{i(i-1)}Z_{i-1}) = 0 \qquad (14)$$

Beschouwen we vervolgens de oplossing via het algemene regressiemodel, dan dienen we vergelijking (12) als regressievergelijking te interpreteren en via 'least squares' de partiële gestandaardiseerde regressiecoëfficiënten te berekenen. Hiertoe dienen we de uitdrukking $\Sigma(p_{ie_i}e_i)^2$ te minimaliseren, ofwel het minimaliseren van:

$$\Sigma(Z_i - p_{i1}Z_1 - p_{i2}Z_2 - \ldots p_{i(i-1)}Z_{i-1})^2$$

Differentiëren naar Z_j, waarbij $j = 1, 2, \ldots (i-1)$ levert:

$$\Sigma(Z_j)(Z_i - p_{i1}Z_1 - p_{i2}Z_2 - \ldots p_{i(i-1)}Z_{i-1}) = 0 \qquad (15)$$

De vergelijkingenstelsels (14) en (15) leveren (i - 1) identieke vergelijkingen in (i - 1) onbekenden, welke noodzakelijkerwijs ook dezelfde oplossingen geven.

Aan het hierboven afgeleide resultaat willen we tot slot nog een waarschuwing verbinden. De uitkomst dat onder hantering van de 'instrumentele variabelen'-procedure de coëfficiënten p_{ij} gelijk zijn aan de coëfficiënten $\beta_{ij.k..n}$, mag niet geïnterpreteerd worden als $\beta_{ji.k..n} = p_{ji} = 0$. Hier zou men een gedachtenfout maken: indien $p_{ij} \neq 0$ dan heeft p_{ji} in het monocausale model geen enkele betekenis. Op grond hiervan kan echter niet worden geconcludeerd, dat $\beta_{ji.k..n}$ binnen het lineaire regressiemodel geen betekenis zou

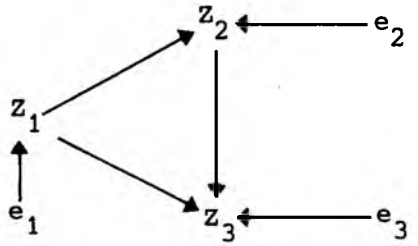
hebben. Dit zou betekenen, dat we twee qua structuur verschillende modellen met elkaar verwarren: beide β's zullen in het algemeen een waarde ongelijk nul hebben.

De Simon-Blalock-procedure versus padanalyse

Vóór de introductie van de padanalyse in de sociale wetenschappen waren vooral de elaboratie-procedure (via uitsplitsing van tabellen) en de partiële correlatie-analyse erg populair. De laatste, die bekend staat als de Simon-Blalock-procedure, steunt op het feit dat in het geval van overgeïdentificeerde recursieve stelsels voorspellingen kunnen worden afgeleid met betrekking tot het nul worden van bepaalde partiële correlaties. Als gevolg hiervan is de Simon-Blalock-procedure alleen toepasbaar op modellen, waarvan in het diagram minstens één pijl ontbreekt als gevolg van het afwezig zijn van een directe causale relatie.

Indien alle pijlen in een diagram zijn getrokken levert ook padanalyse geen toetsingsmogelijkheden meer. Wel kunnen we de numerieke waarden van de coëfficiënten p_{ij} bepalen. De Simon-Blalock-procedure kan derhalve gezien worden als een analyseprocedure voor modellen, die een deelverzameling vormen van de modellen, waarop de padanalyse-procedure van toepassing is. We merken nog op, dat de assumpties met betrekking tot modellen onderworpen aan de Simon-Blalock-procedure dezelfde zijn als die met betrekking tot modellen ontworpen aan de padanalyse. Het verschil tussen de Simon-Blalock-procedure en de padanalyse kan aan de hand van diagram 4 worden toegelicht.

Diagram 4



Het bij (pad)diagram 4 behorend padmodel is:

$Z_1 = p_{1e_1} e_1$

$Z_2 = p_{21} Z_1 + p_{2e_2} e_2$

$Z_3 = p_{31} Z_1 + p_{3e_3} e_3$

Op grond hiervan laten zich de volgende padcoëfficiënten-vergelijkingen opstellen:

$r_{12} = p_{21}$

$r_{13} = p_{31}$

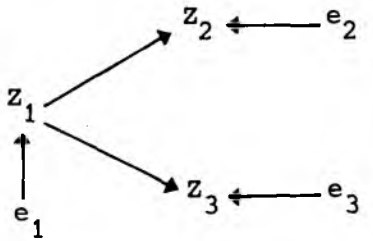
$r_{23} = p_{31} r_{12}$

We hebben hier een situatie van drie vergelijkingen in twee onbekenden. De laatste twee vergelijkingen leveren een toetsingsmogelijkheid. Beide vergelijkingen dienen immers bij benadering identieke schattingen voor p_{31} te leveren als het model juist is. Dit impliceert de voorspelling $r_{13} = r_{23}/r_{12}$ ofwel $r_{23} = r_{12} r_{13}$.

We kunnen direct opmerken, dat deze voorspelling identiek is aan de voorspelling, die we met behulp van de Simon-Blalock-procedure kunnen afleiden:

$$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}} = 0 \text{ ofwel } r_{23} - r_{12} r_{13} = 0$$

Diagram 5



Beschouwen we vervolgens (pad)diagram 5, dan is noch met behulp van de Simon-Blalock-procedure, noch met behulp van de padanalyse enige voorspelling af te leiden. Wel kunnen we met behulp van de padanalyse de padcoëfficiënten p_{21}, p_{31} en p_{32} uit het bij diagram 5 behorend padmodel berekenen.

Op grond van het voorafgaande kunnen we concluderen, dat wanneer we werken met overgeïdentificeerde modellen, d.w.z. modellen waarin minstens één pijl ontbreekt, er toetsingsmogelijkheden bestaan: uit bepaalde vergelijkingen kunnen we voorspellingen ten aanzien van de empirische resultaten opstellen. Dit soort voorspellingen is geheel en al vergelijkbaar met die welke af te leiden zijn met behulp van de Simon-Blalock-procedure. Werken

we daarentegen met precies geïdentificeerde modellen (alle $\binom{n}{2}$ pijlen getrokken), dan zijn er geen toetsingsmogelijkheden. Soms is het zinvol om aan de padanalyse een Simon-Blalock-procedure te laten voorafgaan. Namelijk in die situaties waar sprake is van simpele voorspellingen met betrekking tot het nul worden van partiële correlaties van de eerste of tweede orde. Worden dergelijke voorspellingen gefalsifieerd, dan heeft het geen zin padanalyse toe te passen op het ongemodificeerde model.

Interpretatie van de padcoëfficiënt

Ter ondersteuning van de verbale interpretatie, die aan padcoefficiënten wordt toegekend, geven we nu een interpretatie in termen van partiële varianties. We voeren hierbij wel de beperking in, dat we alleen padcoëfficiënten in precies geïdentificeerde vergelijkingen beschouwen, waarvoor geldt, zoals is aangetoond: $p_{ij} = \beta_{ij.k..n}$. Wellicht ten overvloede zij nog opgemerkt, dat de navolgende interpretatie direct afgeleid is van de interpretatie van de coëfficiënten in het lineaire regressiemodel.

We definiëren:

$$p^2_{ij} = \beta^2_{ij.1..ne_i} = \frac{S^2_{i.12..(j-1)(j+1)..ne_i}}{S^2_i} \cdot \frac{S^2_j}{S^2_{j.12..(i-1)(i+1)..ne_i}} \qquad (16)$$

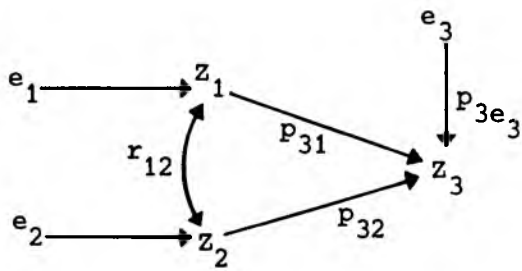
waarbij we de partiële variantie $S^2_{i.12..(j-1)(j+1)..ne_i}$ definieren als de variantie in Z_i onder constant houden van alle andere variabelen, inclusief e_i, doch uitgezonderd Z_j. Met andere woorden: de variantie in Z_i, waarvoor Z_j direct verantwoordelijk is. Het eerste gedeelte van het rechterlid van (16) kunnen we dan interpreteren als de proportie verklaarde variantie in Z_i, waarvoor Z_j direct verantwoordelijk is. Indien Z_j ongecorreleerd is met alle andere variabelen in het model, dan kan het kwadraat van de padcoëfficiënt opgevat worden als de proportie verklaarde variantie in de afhankelijke variabele Z_i. In het algemeen is Z_j echter gecorreleerd met andere variabelen in het model, waardoor de proportie verklaarde variantie, weergegeven door het eerste gedeelte in het rechter lid van (16) een onderschatting

geeft van de direct door Z_j verklaarde variantie. Daarom bepalen we de proportie autonome variantie in Z_j onder constant houden van alle variabelen uitgezonderd Z_i, en vermenigvuldigen we het eerste gedeelte van het rechter lid van (16) met de reciproke hiervan. Voor een uitvoeriger afleiding van de formule (16) verwijzen we naar Yule en Kendall (1964, p. 283 e.v.).

Een zeer eenvoudig model

Gegeven zij het volgende paddiagram:

Diagram 6



en het hierbij behorende padmodel:

$Z_1 = p_{1e_1} e_1$

$Z_2 = p_{2e_2} e_2$

$Z_3 = p_{31} Z_1 + p_{32} Z_2 + p_{3e_3} e_3$

Hierin zijn p_{1e_1} en p_{2e_2} beide gelijk aan 1, omdat de onafhankelijke variabelen Z_1 en Z_2 beide geacht worden volledig door variabelen buiten het model bepaald te zijn.

De padcoëfficiëntvergelijkingen kunnen uit deze structuurvergelijkingen als volgt bepaald worden: de vergelijking van Z_3 wordt vermenigvuldigd met Z_1; gesommeerd over alle waarnemingen en gedeeld door N, waarbij geassumeerd wordt dat e_3 ongecorreleerd is met Z_1:

$$\frac{\Sigma Z_1 Z_3}{N} = p_{31} \frac{\Sigma Z_1 Z_1}{N} + p_{32} \frac{\Sigma Z_1 Z_2}{N} + p_{3e_3} \frac{\Sigma Z_1 e_3}{N}$$

ofwel: $r_{13} = p_{31} + p_{32} r_{12} + 0$ (17)

Hieruit zien we, dat de correlatie r_{13} opgebouwd is uit een direct

effect van Z_1 op Z_3 (p_{31}) en een indirect effect ($p_{32}\,r_{12}$); hieruit volgt dat het indirecte effect van Z_1 op Z_3 gelijk is aan $r_{13} - p_{31}$.

Vergelijking (17) is een vergelijking met twee onbekende padcoefficiënten.

Vervolgens wordt de structuurvergelijking van Z_3 vermenigvuldigd met Z_2; gesommeerd over alle waarnemingen en gedeeld door N, waarbij geassumeerd wordt, dat e_3 ongecorreleerd is met Z_2:

$$\frac{\Sigma Z_2 Z_3}{N} = p_{31}\frac{\Sigma Z_1 Z_2}{N} + p_{32}\frac{\Sigma Z_2 Z_2}{N} + p_{3e_3}\frac{\Sigma Z_2 e_3}{N}$$

ofwel: $r_{23} = p_{31}\,r_{12} + p_{32} + 0$ (18)

Uit (17) en (18) kunnen, gegeven de correlatiecoëfficiënten, p_{31} en p_{32} worden opgelost.

In het algemeen geldt, indien Z_k oorzaak is van Z_j:

$$r_{kj} = \sum_{i=1}^{j-1} p_{ji}\,r_{ki}$$

Door het invoeren van verschillende waarden voor de correlaties in diagram 6 en het vervolgens oplossen van de padcoëfficiënten, verkrijgt men enig inzicht in de relaties tussen de verschillende grootheden en de toepassing van de padanalyse-techniek.

De onverklaarbare variantie

Hetzelfde eenvoudige model kan dienen om inzicht te krijgen in de aard van de restvariabele: vermenigvuldigen van de structuurvergelijking van Z_3 met Z_3, sommeren over alle waarnemingen en delen door N, geeft:

$$r_{33} = p_{31}\,r_{13} + p_{32}\,r_{23} + p^2_{3e_3} \qquad (19)$$

Substitutie van vergelijkingen (17) en (18) in (19) levert:

$$r_{33} = p_{31}(p_{31} + p_{32}r_{12}) + p_{32}(p_{31}r_{12} + p_{32}) + p^2_{3e_3}$$

$$= p^2_{31} + + p_{31}p_{32}r_{12} + p_{32}p_{31}\,r_{12} + p^2_{32} + p^2_{3e_3}$$

$$= p^2_{31} + p^2_{32} + 2p_{31}\,p_{32}\,r_{12} + p^2_{3e_3} = 1$$

$$p^2_{3e_3} = 1 - (p^2_{31} + p^2_{32} + 2p_{31}\,p_{32}\,r_{12})$$

Algemeen: $p^2_{ie_i} = 1 - \sum_{j=1}^{i-1} p^2_{ij} - 2 \sum_{j=1}^{i-1} \sum_{k=j+1}^{i-1} p_{ij} p_{ik} r_{jk}$

We zullen nu bewijzen, dat p_{3e_3} gelijk is aan de wortel uit de onverklaarde variantie van de afhankelijke variabele Z_3, ofwel:

$$p_{3e_3} = \sqrt{(1 - R^2_{3.12})}$$

In het geval van drie variabelen geldt:

$$R^2_{3.12} = r^2_{31} + r^2_{32.1}(1 - r^2_{31})$$

$$r_{32.1} = \frac{r_{32} - r_{31} r_{21}}{\sqrt{1 - r^2_{31}}\sqrt{1 - r^2_{21}}}$$

$$R^2_{3.12} = r^2_{31} + \left(\frac{r_{32} - r_{31} r_{21}}{\sqrt{1 - r^2_{31}}\sqrt{1 - r^2_{21}}}\right)^2 (1 - r^2_{31})$$

$$= r^2_{31} + \frac{r^2_{32} + r^2_{31} r^2_{21} - 2 r_{32} r_{31} r_{21}}{1 - r^2_{21}}$$

$$= \frac{r^2_{31} - r^2_{31} r^2_{21} + r^2_{32} + r^2_{31} r^2_{21} - 2 r_{32} r_{31} r_{12}}{1 - r^2_{21}}$$

$$= \frac{r^2_{31} + r^2_{32} - 2 r_{32} r_{31} r_{12}}{1 - r^2_{12}}$$

Substitutie van $r_{31} = p_{31} + p_{32} r_{12}$ en van $r_{32} = p_{32} + p_{31} r_{12}$ leidt tot:

$$R^2_{3.12} = \frac{p^2_{31} + p^2_{32} r^2_{12} + 2 p_{31} p_{32} r_{12} + p^2_{32} + p^2_{31} r^2_{12} + 2 p_{31} p_{32} r_{12}}{1 - r^2_{12}}$$

$$\frac{-2(p_{31} + p_{32} r_{12})(p_{32} + p_{31} r_{12}) r_{12}}{}$$

$$= \frac{p_{31}^2+p_{32}^2+4p_{31}p_{32}r_{12}+p_{32}^2r_{12}^2+p_{31}^2r_{12}^2-2p_{31}p_{32}r_{12}}{1-r_{12}^2}$$

$$\frac{-2p_{31}^2r_{12}^2-2p_{32}^2r_{12}^2-2p_{31}p_{32}r_{12}^2}{}$$

$$= \frac{p_{31}^2+p_{32}^2+2p_{31}p_{32}r_{12}-p_{32}^2r_{12}^2-p_{31}^2r_{12}^2-2p_{31}p_{32}r_{12}^3}{1-r_{12}^2}$$

$$= \frac{p_{31}^2(1-r_{12}^2)+p_{32}^2(1-r_{12}^2)+2p_{32}p_{31}r_{12}(1-r_{12}^2)}{1-r_{12}^2}$$

$= p_{31}^2 + p_{32}^2 + 2p_{32}p_{31}r_{12} = 1 - p_{3e_3}^2$ hetgeen te bewijzen was.

De algemene gedaante van de formule is:

$$R^2_{n.12\ldots(n-1)} = \sum_{i=1}^{n-1} p_{ni}^2 + \sum_{i=1}^{n-1}\sum_{\substack{j=1 \\ i\neq j}}^{n-1} p_{ni}p_{nj}r_{ij}$$

Tenslotte laten we nogmaals (zie ook het identificatieprobleem), nu aan de hand van het eenvoudige padmodel, zien, dat de padcoefficiënt p_{31} gelijk is aan de partiële gestandaardiseerde regressiecoëfficiënt $\beta_{31.2}$. Substitutie van (18) in (17) levert:

$p_{31} = r_{31} - (r_{32} - p_{31}r_{12})r_{12}$

$p_{31} = r_{31} - r_{32}r_{12} + p_{31}r_{12}r_{12}$

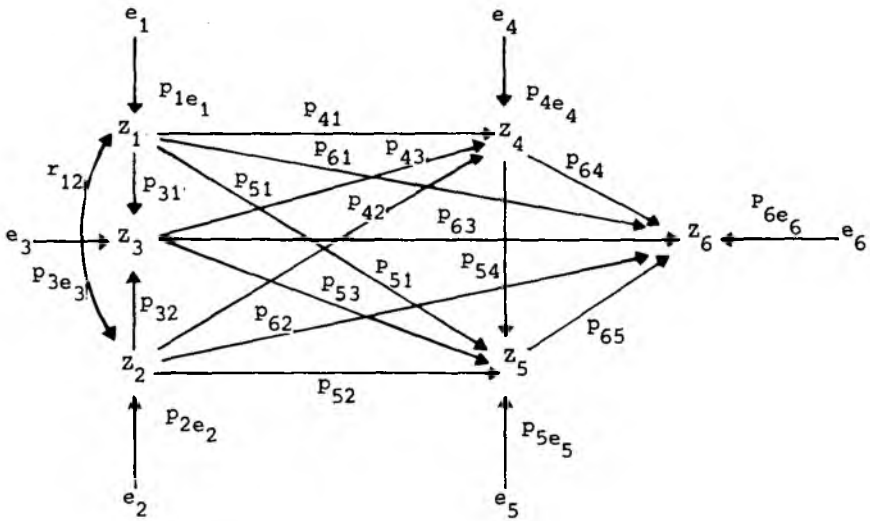
$p_{31} - p_{31}r_{12}r_{12} = r_{31} - r_{32}r_{12}$

$p_{31}(1 - r_{12}^2) = r_{31} - r_{32}r_{12}$

$$p_{31} = \frac{r_{31} - r_{32}r_{12}}{1 - r_{12}^2} = \beta_{31.2}$$

Hiertoe nemen we onderstaand paddiagram met de bijbehorende structuurvergelijkingen in beschouwing.

Diagram 7



$Z_1 = p_{1e_1} e_1$

$Z_2 = p_{2e_2} e_2$

$Z_3 = p_{31} Z_1 + p_{32} Z_2 + p_{3e_3} e_3$

$Z_4 = p_{41} Z_1 + p_{42} Z_2 + p_{43} Z_3 + p_{4e_4} e_4$

$Z_5 = p_{51} Z_1 + p_{52} Z_2 + p_{53} Z_3 + p_{54} Z_4 + p_{5e_5} e_5$

$Z_6 = p_{61} Z_1 + p_{62} Z_2 + p_{63} Z_3 + p_{64} Z_4 + p_{65} Z_5 + p_{6e_6} e_6$

Willen we b.v. de padcoëfficiënten in de structurele vergelijking van Z_4 bepalen, dan moeten we de vergelijking van Z_4 doorvermenigvuldigen met de variabelen die Z_4 veroorzaken, te weten Z_1, Z_2 en Z_3. Na doorvermenigvuldiging verkrijgen we:

$r_{14} = p_{41} r_{11} + p_{42} r_{12} + p_{43} r_{13}$

$r_{24} = p_{41} r_{21} + p_{42} r_{22} + p_{43} r_{23}$

$r_{34} = p_{41} r_{31} + p_{42} r_{32} + p_{43} r_{33}$

Aangezien $r_{11} = r_{22} = r_{33} = 1$, en de overige correlatiecoëfficienten bekend zijn, hebben we hier te maken met een stelsel vergelijkingen in drie onbekenden, waaruit p_{41}, p_{42} en p_{43} opgelost kunnen worden.

Algemeen gesteld komen er in de Z_k-structuurvergelijking k - 1 te bepalen padcoëfficiënten voor; er zijn m.a.w. k - 1 vergelijkingen nodig om de padcoëfficiënten te bepalen. Deze vergelijkingen verkrijgen we via doorvermenigvuldiging van Z_k met de variabelen Z_i (i = 1, 2 ... k-1) die Z_k veroorzaken, en door daarna dit stelsel vergelijkingen op te lossen naar de onbekende coëfficiënten. Voor k > 3 wordt dit laatste al spoedig een aangelegenheid die men beter aan de computer kan overdragen.

We geven vervolgens een indruk van de directe en indirecte effecten binnen het model door te laten zien hoe de correlatie r_{35} is opgebouwd. We stellen daartoe achtereenvolgens onderstaande padcoëfficiëntvergelijkingen op:

$$r_{35} = \frac{\Sigma Z_3 Z_5}{N} = \frac{\Sigma Z_3 (p_{51} Z_1 + p_{52} Z_2 + p_{53} Z_3 + p_{54} Z_4 + p_{5e_5} e_5)}{N}$$

$$= p_{51} \frac{\Sigma Z_3 Z_1}{N} + p_{52} \frac{\Sigma Z_3 Z_2}{N} + p_{53} \frac{\Sigma Z_3 Z_3}{N} + p_{54} \frac{\Sigma Z_4 Z_3}{N} +$$

$$p_{5e_5} \frac{\Sigma Z_3 e_5}{N} = p_{51} r_{31} + p_{52} r_{32} + p_{53} + p_{54} r_{34} \qquad (20)$$

$$r_{34} = \frac{\Sigma Z_3 Z_4}{N} = \frac{\Sigma Z_3 (p_{41} Z_1 + p_{42} Z_2 + p_{43} Z_3 + p_{4e_4} e_4)}{N}$$

$$= p_{41} \frac{\Sigma Z_3 Z_1}{N} + p_{42} \frac{\Sigma Z_3 Z_2}{N} + p_{43} \frac{\Sigma Z_3 Z_3}{N} + p_{4e_4} \frac{\Sigma Z_3 e_4}{N}$$

$$= p_{41} r_{31} + p_{42} r_{32} + p_{43} \qquad (21)$$

$$r_{32} = \frac{\Sigma Z_3 Z_2}{N} = \frac{\Sigma Z_2 (p_{31} Z_1 + p_{32} Z_2 + p_{3e_3} e_3)}{N}$$

$$= p_{31} \frac{\Sigma Z_2 Z_1}{N} + p_{32} \frac{\Sigma Z_2 Z_2}{N} + p_{3e_3} \frac{\Sigma Z_2 e_3}{N}$$

$$= p_{31} r_{12} + p_{32} \tag{22}$$

$$r_{31} = \frac{\Sigma Z_3 Z_1}{N} = \frac{\Sigma Z_1 (p_{31} Z_1 + p_{32} Z_2 + p_{3e_3} e_3)}{N}$$

$$= p_{31} \frac{\Sigma Z_1 Z_1}{N} + p_{32} \frac{\Sigma Z_1 Z_2}{N} + p_{3e_3} \frac{\Sigma Z_1 e_3}{N}$$

$$= p_{31} + p_{32} r_{12} \tag{23}$$

Uit (20) tot en met (23) volgt:

$$r_{35} = p_{51}(p_{31} + p_{32} r_{12}) + p_{52}(p_{31} r_{12} + p_{32}) + p_{53} + p_{54}(p_{41} (p_{31} + p_{32} r_{12}) + p_{42} (p_{31} r_{12} + p_{32}) + p_{43})$$

$$= p_{51} p_{31} + p_{51} p_{32} r_{12} + p_{52} p_{31} r_{12} + p_{52} p_{32} + p_{53} + p_{54} p_{41} p_{31} + p_{54} p_{41} p_{32} r_{12} + p_{54} p_{42} p_{31} r_{12} + p_{54} p_{42} p_{32} + p_{54} p_{43}$$

$$= p_{53} + p_{54} p_{43} + p_{54} p_{42} p_{32} + p_{54} p_{42} p_{31} r_{12} + p_{54} p_{41} p_{32} r_{12} + p_{54} p_{41} p_{31} + p_{52} p_{32} + p_{52} p_{31} r_{12} + p_{51} p_{32} r_{12} + p_{51} p_{31}.$$

Bovenstaande schrijfwijze geeft ons informatie over de directe invloed van Z_3 op Z_5 (uitgedrukt in p_{53}) en over de indirecte invloed (6), langs alle causale paden van Z_3 naar Z_5. De som van de indirecte invloeden is gelijk aan $r_{35} - p_{53}$. Op overeenkomstige wijze kunnen we de correlatie van elk paar variabelen binnen het model uitdrukken in een directe en een indirecte invloed van de ene op de andere via een formule waarin alleen padcoëfficiënten en de correlatie(s) tussen de exogene variabelen zijn opgenomen (7).

Tenslotte merken we op, dat een correlatie tussen exogene variabelen een correlatie tussen de restvariabelen van deze variabelen impliceert. Deze kan b.v. veroorzaakt zijn door een gemeenschappelijke oorzaakvariabele buiten het model. Wellicht kent men een dergelijke 'impliciete' variabele zelfs, maar wanneer men hier niet in geïnteresseerd is, kan men deze uiteraard zonder meer buiten het model houden. Uit de afleidingen van de

voorspellingen zal duidelijk geworden zijn, dat wij geen assumptie terzake van het ongecorreleerd zijn van de restvariabelen van de exogene variabelen onderling behoeven, dit i.t.t. de relaties tussen de restvariabelen van de endogene variabelen onderling en de restvariabelen van exogene en endogene variabelen t.o.v. elkaar.

Voorts wijzen we er op, dat in het onderhavige model alle pijlen zijn getrokken, zodat er geen toetsingsmogelijkheden zijn; elke verzameling van empirisch gevonden correlaties 'past' in het model.

Toepassing

Voor een toepassing van de padanalyse ontlenen we de gegevens aan het vrijetijdsbestedingsonderzoek van Wippler (1968). De door ons geselecteerde variabelen zijn de volgende:

Z_1 : leeftijd (oud ⟶ jong)

Z_2 : opleiding (laag ⟶ hoog)

Z_3 : burg. staat (gehuwd ⟶ ongehuwd)

Z_4 : perceptie promotiemogelijkheden (geen ⟶ wel)

Z_5 : vitaal-expansief vrijetijdsgedrag (geen ⟶ frequent)

De correlatiematrix is als volgt:

	Z_2	Z_3	Z_4	Z_5
Z_1	.27	.61	.35	.54
Z_2	-	.24	.43	.43
Z_3		-	.35	.49
Z_4			-	.45
Z_5				-

Het is interessant om te laten zien, hoe het model stapsgewijs opgebouwd kan worden; beginnend met één, daarna twee, vervolgens drie en vier onafhankelijke variabelen. In feite doet Wippler dit ook bij zijn stapsgewijs opgebouwde multipele correlaties, maar zijn conclusie is misleidend ten aanzien van de relatieve bijdragen van de verschillende onafhankelijke variabelen, omdat een en ander een artefact is van de techniek: of een toegevoegde

variabele veel of weinig verklaart hangt bij deze techniek af van de variabelen die er al zijn ingestopt. Bij de padanalyse is hetzelfde het geval, maar wanneer het model compleet is geven de padcoëfficiënten een objectief beeld van de relatieve bijdrage van elk der onafhankelijke variabelen. Dit laatste kan men met de multipele correlatierekening niet bereiken.

We kunnen een en ander als volgt toelichten. Gewoonlijk wordt de analyse gestart met die onafhankelijke variabele, welke de hoogste correlatie met de afhankelijke vertoont. In ons geval betreft dit de variabele leeftijd (Z_1). Het paddiagram en het bijbehorende padmodel ziet er als volgt uit:

Diagram 8

$$Z_5 = p_{51} Z_1 + p_{5e_5} e_5$$

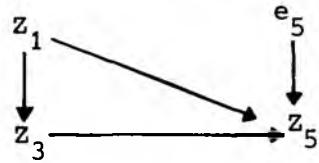
De padcoëfficiëntvergelijking luidt: $r_{15} = p_{51} = .54$.

De verklaarde variantie is 29%; $p_{5e_5} = \sqrt{1 - .29} = .84$.

Was de analyse gestart met de variabele burgerlijke staat (Z_3), dan hadden we verkregen $p_{53} = r_{35} = .49$; een verklaarde variantie van 24% en een $p_{5e_5} = .87$.

We construeren vervolgens het paddiagram met de bijbehorende vergelijkingen voor de variabelen Z_1, Z_3 en Z_5.

Diagram 9



$$Z_1 = p_{1e_1} e_1$$

$$Z_3 = p_{31} Z_1 + p_{3e_3} e_3$$

$$Z_5 = p_{51} Z_1 + p_{53} Z_3 + p_{5e_5} e_5$$

De voorspellingen vanuit de padcoëfficiëntvergelijkingen

$r_{13} = p_{31}$

$r_{15} = p_{51} + p_{53} r_{13}$

$r_{35} = p_{51} r_{13} + p_{53}$

leveren op: $p_{31} = .61$, $p_{51} = .38$ en $p_{53} = .26$.

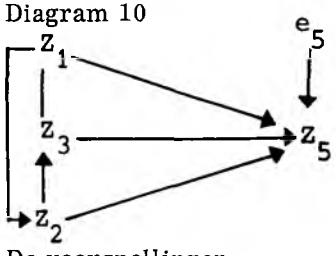
Verder is $R^2_{5.13} = p^2_{51} + p^2_{53} + 2p_{51}p_{53} r_{13} = .33$; de verklaarde variantie is dus 33%, en $p_{5e_5} = .82$.

We zien, dat p_{51} daalt ten opzichte van het eerste model. De verklaring hiervoor wordt bij vergelijking van de vergelijkingen zonder meer duidelijk; terwijl eerst gold: $r_{15} = p_{51}$, geldt nu: $r_{15} = p_{51}$ + 'iets', m.a.w. p_{51} wordt nu kleiner (uitgaande van positieve waarden). De correlatie tussen Z_1 en Z_5 wordt in het tweede geval niet alleen door het directe causale pad $Z_1 \longrightarrow Z_5$ veroorzaakt, maar bovendien door een tweede pad via Z_3. Naarmate r_{13} hoger is, zou de causale invloed in twee meer gelijke delen gesplitst worden.

Een tweede interessant punt is, dat het van geen belang is of de causale invloed van Z_1 naar Z_3 loopt of van Z_3 naar Z_1; de voorspellingen blijven gelijk.

We voegen vervolgens een nieuwe variabele toe:

Diagram 10



$$Z_1 = p_{1e_1} e_1$$

$$Z_3 = p_{31} Z_1 + p_{32} Z_2 + p_{3e_3} e_3$$

$$Z_2 = p_{21} Z_1 + p_{2e_2} e_2$$

$$Z_5 = p_{51} Z_1 + p_{53} Z_3 + p_{52} Z_2 + p_{5e_5} e_5$$

De voorspellingen

$$r_{12} = p_{21}$$

$$r_{13} = p_{31} + p_{32} r_{12}$$

$$r_{23} = p_{31} r_{12} + p_{32}$$

$$r_{15} = p_{51} + p_{53} r_{13} + p_{52} r_{12}$$

$$r_{25} = p_{51} r_{12} + p_{53} r_{23} + p_{52}$$

$$r_{35} = p_{51} r_{13} + p_{53} + p_{52} r_{23}$$

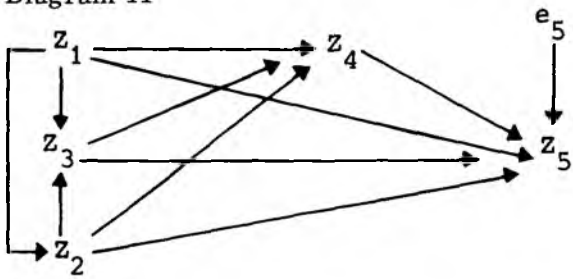
leveren: $p_{21} = .27$; $p_{31} = .59$; $p_{32} = .07$; $p_{51} = .32$; $p_{53} = .23$; $p_{52} = .29$.

Weer is nu p_{51} in waarde gedaald; bovendien is ook p_{53} kleiner (de correlatie r_{35} is niet alleen spurious via Z_1 maar ook via Z_2, zoals uit dit kleiner worden blijkt).

$R^2_{5.123} = .41$; de verklaarde variantie op basis van deze drie onafhankelijke variabelen is dus 41%; $p_{5e_5} = .77$.

Wederom breiden we het model uit, nu met de onafhankelijke variabele 'perceptie van promotiemogelijkheden', Z_4.

Diagram 11



$Z_1 = p_{1e_1} e_1$

$Z_2 = p_{21} Z_1 + p_{2e_2} e_2$

$Z_3 = p_{31} Z_1 + p_{32} Z_2 + p_{3e_3} e_3$

$Z_4 = p_{41} Z_1 + p_{42} Z_2 + p_{43} Z_3 + p_{4e_4} e_4$

$Z_5 = p_{51} Z_1 + p_{52} Z_2 + p_{53} Z_3 + p_{54} Z_4 + p_{5e_5} e_5$

De voorspellingen:

$r_{12} = p_{21}$

$r_{13} = p_{31} + p_{32} r_{12}$

$r_{23} = p_{31} r_{12} + p_{32}$

$r_{14} = p_{41} + p_{42} r_{12} + p_{43} r_{13}$

$r_{34} = p_{41} r_{13} + p_{42} r_{23} + p_{43}$

$r_{24} = p_{41} r_{12} + p_{42} + p_{43} r_{23}$

$r_{15} = p_{51} + p_{52} r_{12} + p_{53} r_{13} + p_{54} r_{14}$

$r_{35} = p_{51} r_{13} + p_{52} r_{23} + p_{53} + p_{54} r_{34}$

$r_{25} = p_{51} r_{12} + p_{52} + p_{53} r_{23} + p_{54} r_{24}$

$r_{45} = p_{51} r_{14} + p_{52} r_{24} + p_{53} r_{34} + p_{54}$

leveren: $p_{21} = .27$; $p_{31} = .59$; $p_{32} = .07$; $p_{41} = .15$; $p_{42} = .35$; $p_{43} = .18$; $p_{51} = .30$; $p_{52} = .23$; $p_{53} = .19$; $p_{54} = .18$.

Zowel p_{51} als p_{53} als p_{52} zijn in waarde gedaald. De correlatie

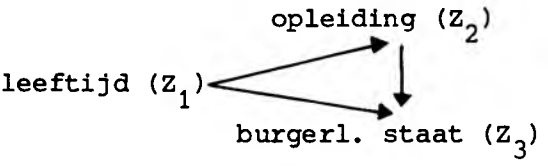
r_{35} is ten dele spurious (via Z_1, Z_2 en Z_4).

$R^2_{5.1234} = .44$; de verklaarde variantie op basis van de vier onafhankelijke variabelen is 44%; $p_{5e_5} = .75$. Toevoeging van Z_4 heeft weinig bijgedragen in het verklaren van de variantie in de afhankelijke variabele. Vergelijk $R^2_{5.123} = .41$.

In het door ons weergegeven model wordt Z_5 als direct afhankelijk van alle andere variabelen voorgesteld. In feite doet Wippler dit ook bij zijn berekening van de multipele correlatie; het ligt dan ook voor de hand dat de resultaten m.b.t. de totale verklaarde variantie daarom niet uiteenlopen. Anders zou het zijn, wanneer wij pijlen hadden geëlimineerd. De Simon-Blalock-procedure als explorerende voorfase van de padanalyse gaf hier echter geen aanleiding toe. Een voordeel van de toepassing van de padanalyse is echter wèl, zoals gesteld, de mogelijkheid om de sterkte van de diverse invloeden te schatten op een wijze die onafhankelijk is van de volgorde van introductie van variabelen in de analyse.

Verder zijn wij via de padanalyse gekomen tot een schatting van de indirecte effecten, doordat wij vooraf een theorie terzake van a l l e relaties tussen de betrokken variabelen hebben gespecificeerd.

We merken daarbij wel op, dat b.v. het door ons aldus gespecificeerde theoriegedeelte



geen toetsing van de causale richtingen toelaat; het betreffende stelsel van vergelijkingen is precies geïdentificeerd. We zijn hier dus - voor wat de berekening van de coëfficiënten betreft - volledig afhankelijk van onze keuze vooràf - op grond van onderzoek en/of voorwetenschappelijk inzicht - van het model. Duidelijk zal overigens zijn, dat met de Simon-Blalock-procedure in een dergelijk geval in het geheel niets bereikt kan worden.

Noten

Wij danken Drs. W. E. Saris en de leden van de vakgroep Theorie en Methodologie i.o. voor hun opmerkingen naar aanleiding van een eerdere versie van dit artikel.

1. Uitgesloten worden hier elementaire uitspraken van het type 'b% van groep A heeft kenmerk X'.
2. Zie paragraaf over identifikatie.
3. Dit betreft geen principieel punt, maar vereenvoudigt de berekeningen.
4. We gaan er van uit, dat p_{31} en p_{32} voor elke respondent dezelfde waarde hebben.
5. $\Sigma Z_i Z_j / N = r_{ij}$ omdat Z_i en Z_j in standaardskores zijn uitgedrukt.
6. Voor een nadere specifikatie van het begrip 'effekt' pleiten auteurs als Finney (1972) en Saris (1974). Zo wordt de term $p_{54}p_{43}$ beschouwd als een 'echt' indirekt effekt; de term $p_{52}p_{32}$ als een 'spurious' effekt, teweeggebracht door de gemeenschappelijke oorzaak Z_2.
7. Vgl. ook Heise (1969), die, uitgaande van de struktuurvergelijkingen, een matrixvermenigvuldigingsmethode heeft ontworpen om hetzelfde doel te bereiken.

Literatuur

Blalock, H. M., 'Causal inferences, closed populations and measures of association', in: *American Political Science Review,* 61, 1967, pp. 130-136.

Boudon, R., 'A method of linear causal analysis: dependence analysis, in: *American Sociological Review,* 30, 1965, pp. 365-374.

Duncan, O. D., 'Path analysis: sociological examples', in: *American Journal of Sociology,* 72, 1966, pp. 1-16.

Finney, J. M., 'Indirect effects in path analysis", in: *Sociological Methods & Research,* 1, 1972, pp. 175-187.

Heise, D. R., 'Problems in path analysis and causal inferences', in: E. F. Borgatta (Ed.), *Sociological Methodology 1969,* 1969, pp. 38-74.

Land, K. C., 'Principles of path analysis', in: E. F. Borgatta (Ed.), *Sociological Methodology 1969,* 1969, pp. 3-38.

Li, C. C., 'The concept of pathcoefficients and its impact on population genetics', in: *Biometrics,* 12, 1956, pp. 190-210.

Saris, W. E., 'An approach to the problem of systematic measurement error in survey research', in: *Mens en Maatschappij,* 49, 1974, pp. 29-50.

Wippler, R., *Sociale determinanten van het vrijetijdsgedrag,* Assen, 1968.

Wright, S., 'Correlation and causation', in: *Journal of Agricultural Research,* 20, 1921, pp. 557-585.

Wright, S., 'The method of path coefficients', in: *Annals of Mathematical Statistics,* 5, 1934, pp. 161-215.

Wright, S., 'Path coefficients and path regressions: alternative or complementary concepts?', in: *Biometrics,* 16, 1960, pp. 189-202.

Yule, G. U. and M. G. Kendall, *An introduction to the theory of statistics,* Londen, 1964.