stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SD 105/73 NOVEMBER

R. POTHARST

EEN FACTOR-ANALYSE IN VERBAND MET EEN ONDERZOEK
NAAR DE SAMENHANG TUSSEN HARTINFARCT EN WERKDRANG

2e boerhaavestraat 49 amsterdam

Een factor-analyse in verband met een onderzoek naar de samenhang tussen hartinfarct en werkdrang*)

door R. Potharst

Het waarnemingsmateriaal bestaat uit de scores van 1336 gezonde proefpersonen op 71 variabelen. Variabele 1 is de leeftijd van de desbetreffende proefpersoon in jaren, variabele 2 het opleidingsniveau (er zijn 3 opleidingsniveau's, genummerd van 1 t/m 3) en de overige 69 variabelen worden gevormd door de antwoorden op 69 aan de proefpersonen voorgelegde items uit een psychologische test, bekend onder de naam 'Beoordeling van Uitspraken Lijst' (BUL). Voor elk van deze items bestonden 6 geordende antwoordmogelijkheden, genummerd van 1 t/m 6.

Bij enkele (zeer weinig) proefpersonen ontbraken de scores op sommige van de 69 items; hier werd telkens de waarde $3\frac{1}{2}$ ingevuld. De gegevens over leeftijd en opleiding waren compleet.

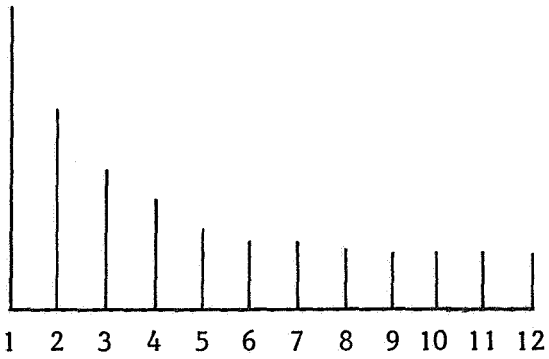
Allereerst zijn alle 2485 productmomentcorrelatie-coëfficienten tussen de 71 variabelen berekend. De in absolute waarde grootste van deze correlaties was 0,56. Verder waren er nog drie groter dan 0,50 en slechts acht tussen de 0,35 en 0,50. Verreweg het grootste deel der correlaties was in absolute waarde kleiner dan 0,30. Hoewel door het bijzonder grote aantal proefpersonen de nulhypothese van onafhankelijkheid der variabelen zeker verworpen zou worden, en er dus samenhang tussen de variabelen aantoonbaar is, is deze samenhang niet erg sterk te noemen. Dit wordt verder bevestigd door de multipele correlaties tussen elke variabele en alle andere, waarvan de kwadraten vrijwel allemaal beneden de 0,40 blijven: slechts acht van de 71 zijn groter dan 0,40, de grootste is 0,47. Zoals bekend, kan het kwadraat van een multipele correlatie-coëfficient tussen een variabele en een stel andere geïnterpreteerd worden als dat deel van de variantie van die ene variabele, dat verklaard wordt door de anderen.

Om de structuur van de samenhang tussen de 71 variabelen verder te onderzoeken, is vervolgens op bovengenoemde correlatie-matrix een factoranalyse volgens de methode van Jøreskog (1963) uitgevoerd.

---

*) Verschijnt als appendix van een artikel over dit onderwerp van drs. H. van Dijl.

Hierbij doet zich eerst het probleem voor van het bepalen van het aantal factoren. A priori was hierover niets bekend. Ook het verloop van de eigenwaarden van S^* (zie Jøreskog p. 38) geeft geen aanwijzingen omtrent het te kiezen aantal factoren: het onderstaande plaatje van de 12 grootste eigenwaarden vertoont het beeld van een vloeiende curve zonder duidelijke knik.



1	10,0	7	2,2
2	6,6	8	2,0
3	4,6	9	1,9
4	3,6	10	1,9
5	2,7	11	1,9
6	2,3	12	1,8

Tabel 1. De twaalf grootste eigenwaarden van S^*

Dat we uiteindelijk gestopt zijn bij 8 factoren kan worden beschouwd als een compromis tussen de volgende argumenten:

(i) De t-waarde, die ongeveer 1 dient te zijn (zie Jøreskog p. 38), is bij 8 factoren 1,024; dit pleit niet voor het opnemen van meer factoren; bij 5 factoren is t reeds 1,074, zodat we wat t betreft hadden kunnen volstaan met minder factoren.

(ii) Het percentage verklaarde variantie is bij 8 factoren 24,82, niet hoog dus en in overeenstemming met de verwachting wegens de over het algemeen lage correlaties tussen de variabelen. Bij 6 factoren is het percentage verklaarde variantie echter reeds 22,29 zodat het toevoegen van de twee laatste factoren nog slechts een verbetering van $2\frac{1}{2}$ % opleverde. Ook bij het toevoegen van nog meer factoren mogen we geen reële verbetering van het percentage verklaarde variantie verwachten.

(iii) Vóór het opnemen van meer factoren pleit echter het feit dat bij 8 factoren sommige restcorrelaties nog geenszins te verwaarlozen zijn. Er zijn er verscheidene in absolute waarde boven de 0,10, de hoogste is zelfs 0.15; het merendeel is echter niet onrustbarend.

(iv) De door Jøreskog (p. 109) voorgestelde procedure voor het toetsen van de hypothese dat k factoren (bij ons k=8) voldoende zijn, tegen het

alternatief dat er meer factoren opgenomen dienen te worden, leidt in ons geval tot verwerping van de nulhypothese dat 8 voldoende zou zijn: u_8=4670 met 2015 vrijheidsgraden.

(v) Een praktisch bezwaar van het toevoegen van nog meer factoren is dat de nieuwe factoren waarschijnlijk oninterpreteerbaar zouden zijn wegens te lage ladingen.

Samenvattende kunnen we zeggen dat, hoewel, met het oog op (iii) en (iv), de zaak waarschijnlijk gecompliceerder ligt dan bij een factormodel met 8 factoren, het niet onredelijk lijkt om zo'n model als eerste benadering aan te nemen.

De factor-ladingen boven de 0,35 in absolute waarde van de (met varimax) geroteerde 8 orthogonale factoren staan vermeld in onderstaande tabel.

Factor I		Factor II		Factor III		Factor IV		Factor V		Factor VI		Factor VII		Factor VIII	
9	38	8	-40	13	47	1	53	40	-36	6	58	44	-43	28	36
19	-40	57	42	27	53	15	35	52	-46	7	39	46	-39		
33	-47	60	55	34	53	25	49	62	-38	12	60				
45	-41	64	61	36	53	39	-36	71	-38						
49	36	65	49	43	35	54	48								
55	55	67	-35	44	-36										
58	55	68	59	46	-44										
59	67	69	55	50	-53										
61	62			65	-38										
63	-47														
70	-58														

Tabel 2. Factorladingen, in absolute waarde ≥ 0,35, van de acht-factor-oplossing, × 100

Bij elke factor staan telkens in de linkerkolom de nummers der variabelen met in de rechterkolom de bijbehorende factorlading, vermenigvuldigd met honderd.

Er zijn in totaal 31 variabelen, die op geen der factoren hoger dan 0,35 laden. Al deze variabelen hebben een multipele correlatie-coëfficient

kleiner dan 0,35 en slechts drie groter dan 0,25, zodat deze variabelen inderdaad weinig samenhangen met de rest.

Wat betreft de betrouwbaarheid van de factorladingen kan het volgende gezegd worden. Voor de ladingen van de ongeroteerde factoren zijn (volgens Jøreskog, p. 57) grove schattingen berekend van de standaardafwijkingen van deze ladingen. Behalve bij de factoren 6 en 7 zijn al deze schattingen van de grootte-orde 0,01 tot 0,10, zodat de ladingen waarschijnlijk redelijk nauwkeurig geschat zijn. Dat de standaardafwijkingen van de factorladingen van de factoren 6 en 7 hoger liggen (tussen zeg 0,10 en 0,25) is hoogstwaarschijnlijk te wijten aan het geringe verschil tussen de 6^{de} en de 7^{de} eigenwaarde van S^*. Dit hoeft ons niet te verontrusten, daar dit bij rotatie weer rechtgetrokken wordt (Jøreskog, p. 100).

De matrix van factorscore-coëfficienten is berekend volgens de methode van Anderson en Rubin (Jøreskog, p. 41). Voor een willekeurig proefpersoon, die de 71 vragen beantwoord heeft, kunnen we nu zijn (geschatte) factorscores op de 8 geëxtraheerde factoren berekenen door vermenigvuldiging van zijn ruwe scores met de factorscore-coëfficientenmatrix.

Theoretisch zouden de correlaties tussen de factorscores van elk paar factoren, berekend over de groep proefpersonen waarvoor de factoranalyse is uitgevoerd, exact 0 moeten zijn. Ter controle van de invloed van afrondingen hebben we de correlaties tussen de factorscores van de eerste zes factoren berekend. De uitslag is weergegeven in de volgende tabel:

2	0,003				
3	0,000	0,000			
4	-0,001	0,001	0,000		
5	-0,006	0,012	-0,001	-0,003	
6	0,005	-0,010	0,001	0,003	0,022
	1	2	3	4	5

Tabel 3. Correlaties tussen de factorscores van de eerste zes factoren

Deze afwijkingen van nul kunnen redelijkerwijs wel worden toegeschreven aan afrondingsfouten.

Bij het werk aan dit onderzoek zijn ook H. Elffers, G.J.F.P. Hanewald en J. Rijvordt betrokken geweest.

Literatuur

K.G. Jøreskog, "Statistical Estimation in Factor Analysis", Stockholm, 1963.