Over multilevel analyse

Jan de Leeuw
Vakgroep Datatheorie FSW/RUL
Ita Kreft
Pedagogisch Didaktisch Instituut UvA

Kommentaar op P. van den Eeden & W. E. Saris, Empirisch Onderzoek naar Multilevel uitspraken, Mens en Maatschappij, 1984, 59, 165-178.

INLEIDING

In een <u>multilevel uitspraak</u> worden relaties aangegeven tussen eenheden van een verschillend (aggregatie)niveau. Een databestand bevat multilevel gegevens wanneer het informatie bevat over eenheden op verschillend niveau. Een data analyse is een multilevel analyse wanneer in de analyse variabelen die eenheden op verschillende aggregatieniveau's beschrijven betrokken worden. Een statistisch model is een multilevel model wanneer het een relatie tussen variabelen, die eenheden op verschillende niveau's beschrijven, formaliseert. Dus: een multilevel statistisch model is een bepaald soort multilevel uitspraak. Aanpassen van een multilevel model doet men met een multilevel analyse, waarvoor men multilevel gegevens nodig heeft.
Binnen het kader van deze definities is nog steeds een aanzienlijke verwarring mogelijk, omdat bijvoorbeeld de term 'variable' niet in eenduidige zin gebruikt wordt. Onder een variabele verstaan we een afbeelding of functie. Ieder variabele heeft een domein (de verzameling van onderzoekseenheden waarop de variabele gedefinieerd is) en een bereik (de verzameling mogelijke waarden die de variabele aanneemt). Bij multilevel gegevens is er sprake van variabelen met een verschillend domein. Over het algemeen zijn deze verschillende domeinen genest. Dat wil zeggen dat de domeinen gerangschikt kunnen worden op zo'n manier dat de elementen van ieder domein deelverzamelingen zijn van de elementen van het voorafgaande domein. Provincies bestaan uit gemeenten, gemeenten uit wijken, wijken uit individuen. Schooltypen bestaan uit scholen, scholen uit klassen, klassen uit leerlingen, enzovoorts. We kunnen ook zeggen dat de elementen van het domein van een variabele van niveau k+1 aggregaties zijn van de elementen van niveau k, en omgekeerd zijn de elementen van niveau k disaggregaties van de elementen van niveau k+1.
Een steeds terugkerende vorm van verwarring in multilevel analyse komt voort uit het feit dat variabelen onder één domein gebracht worden. Wanneer we aan iedere leerling het type toekennen van de school waarop hij zit, dan is daarmee schooltype een individuele variabele geworden. Het is duidelijk dat op deze manier alle variabelen uit een bestand met multilevel gegevens op individueel niveau gebracht kunnen worden. Wanneer dat gebeurt is, dan bevat volgens onze definities dit aangepaste bestand geen multilevel gegevens meer. Alle domeinen zijn gelijk

gemaakt, multilevel model en multilevel analyse zijn niet meer mogelijk. De analyses van bijvoorbeeld Dronkers en Schijf (1984) zijn in deze zin geen multilevel analyses. Het lijkt ons strikt genomen niet mogelijk om op basis van dergelijke analyses multilevel uitspraken te doen. Het onderscheid tussen bijvoorbeeld schooltype-op-niveau-schooltype, schooltype-op-niveau-school, schooltype-op-niveau-individu moet duidelijk gehandhaafd blijven. Het gaat hier om verschillende variabelen. Dat ze natuurlijk wel degelijk iets met elkaar te maken hebben, via de eenvoudige relatie van aggregatie of disaggregatie, wordt nu juist geformaliseerd in het multilevel model.

Multilevel uitspraken zijn van groot belang in de sociologie en in de onderwijsresearch. We wijzen wat betreft de sociologie naar het klassieke artikel van Lazarsfeld en Menzel (1961), en wat betreft de onderwijsresearch naar de overzichtsartikelen van Burstein (1980) of Langbein (1977). In de Nederlandse onderwijsresearch heeft het schoolloopbanen onderzoek (overzicht in Tesser, 1981, Peschar, 1983) zich voornamelijk beziggehouden met analyses op individueel niveau. De laatste tijd komt daarin enige verandering. Dronkers en Schijf (1984) nemen bijvoorbeeld gedisaggregeerde wijk- en schoolvariabelen in hun individuele schoolloopbaanmodel op, met het expliciete doel om daardoor multilevel uitspraken te kunnen doen. We hebben al kort aangegeven dat tegen een dergelijke procedure methodologische bezwaren ingebracht kunnen worden, dit nog afgezien van het feit dat gebruik van gedisaggregeerde variabelen op individueel niveau op een aantal meer technische problemen stuit (die samenhangen met de 'ecologische fout' en het 'Robinson effect', zie bijvoorbeeld Hannan (1970) voor een overzicht).

In een recent artikel voeren Van den Eeden en Saris (1984) een multilevel analyse uit op de GALO-gegevens. Het artikel heeft vooral een technische strekking, de analyse wordt gebruikt om een bepaalde tweestaps procedure te illustreren. In het GALO materiaal zitten onder meer de schoolnummers van de ongeveer dertig scholen in het onderzoek. De eerste stap van de procedure van Van den Eeden en Saris bestaat uit het aanpassen van een zelfde regressiemodel binnen alle scholen afzonderlijk. Dit zijn dus dertig analyses op individueel niveau. Vervolgens worden de schattingen van de regressiegewichten op homogeniteit onderzocht. Gebrek aan homogeniteit impliceert dat er in ieder geval schooleffecten zijn. In de tweede stap worden de dertig schattingen

van de regressieparameters opgevat als een nieuwe data matrix, dat wil zeggen als een aantal variabelen op schoolniveau die door regressie analyse, een modelmatige vorm van aggregatie, uit individuele variabelen zijn verkregen. Deze schoolvariabelen worden vervolgens, ook weer door regressie, gerelateerd aan andere schoolvariabelen. De 'onafhankelijke' schoolvariabelen in de tweede stap zijn in het voorbeeld van Van den Eeden en Saris allemaal direkt geaggregeerde individuele variabelen. De enige 'echte' schoolniveau-variabele in het GALO materiaal die gebruikt wordt is dus het schoolnummer.
Het werk van Van den Eeden en Saris is de directe aanleiding tot dit artikel. Uit onze analyse van het artikel blijkt dat de door hen aanbevolen procedure in algemene zin zeker nuttig is. Een zorgvuldig opgesteld twee-staps model, met een aan de model-assumpties aangepaste statistische analysetechniek, geeft een solide basis voor het doen van multilevel uitspraken. We moeten echter constateren dat er op de uitwerking van dit algemene idee door Van den Eeden en Saris veel aangemerkt kan worden. Onze technische opmerkingen hebben onder andere tot gevolg dat een deel van hun conclusies niet langer volgehouden kan worden.

HET INTERVALNIVEAU

Ons eerste punt van kritiek is verre van origineel, maar het kan niet vaak genoeg herhaald worden. Van den Eeden en Saris passen op hun 31 scholen een enkelvoudig regressiemodel aan, met IQ als predictor van onderwijzersadvies. De variabele advies werd 'in navolging van Blok en Saris (1980) en De Leeuw en Stoop (1979) behandeld als een intervalniveau variabele' (lc, pag. 168). Bij dit citaat zijn een aantal kanttekeningen nodig.
In De Leeuw en Stoop worden de 'Van Jaar tot Jaar' gegevens van het ITS geheranalyseerd met behulp van optimale schalingstechnieken. In dat artikel wordt advies nergens als interval-variabele behandeld. De variabele heeft in het ITS onderzoek overigens slechts vier mogelijke waarden in zijn bereik, te weten vglo, lbo, ulo, mms/vhmo. De optimale schalingen laten zien, dat de schaalwaarden van vglo en lbo zo weinig verschillen dat ze gevoeglijk samen genomen kunnen worden. In het GALO-materiaal is de regressie tussen advies en IQ met vergelijkbare

technieken bestudeerd door Meester en De Leeuw (1983, pag. 53-60). Ook de daar gevonden schalingen wijzen er op dat een niet-lineaire transformatie van advies de regressies beter lineariseert.
Een tweede kanttekening is wat fundamenteler van aard. Waarom willen onderzoekers toch zo graag duidelijk nominale variabelen, zoals advies, behandelen als interval variabelen? Dat komt, omdat ze geleerd hebben dat lineaire regressie, en a forteriori LISREL, alleen maar gebruikt mogen worden bij interval variabelen. Maar een dergelijke gebodsbepaling blijkt bij nader inzien vaag of zelfs betekenisloos te zijn. Als lineaire regressie alleen op interval variabelen 'mag', dan 'mag' discriminant analyse in ieder geval niet. Daar immers is de afhankelijke variabele binair! Regressie-analyse (en LISREL) geven zuivere schatters van de parameters van het model wanneer de regressie van de afhankelijke op de onafhankelijke variabele(n) lineair is. Wanneer de regressie homoscedastisch is, dan zijn de schatters optimaal in een beperkte klasse van schatters, en wanneer de storingen bovendien normaal verdeeld zijn, dan zijn de gebruikelijke schatters de beste in een brede klasse van schatters. In het laatste geval kunnen we ook de gebruikelijke modeltoetsen toepassen, in de andere gevallen moeten de modeltoetsen wat aangepast worden (vergelijk bijvoorbeeld Van Praag, 1981). Nergens in deze opmerkingen is zelfs maar sprake van het begrip 'intervalvariabele'.
De vraag is niet welk meetniveau men veronderstelt, maar de vraag is of men aan de waarden in het bereik van de variabele getallen toekent. Als men dit doet, dan zal men zijn keuze hetzij op inhoudelijke, hetzij op statistische gronden, moeten verdedigen. Toekennen van de getallen 1 t/m 7 aan de categorieën van onderwijzersadvies is op inhoudelijke gronden niet te verdedigen, en blijkt op statistische gronden flink voor verbetering vatbaar. Maar toekennen van scores is in de analyse van Van den Eeden en Saris nergens voor nodig, en kan volgens ons daarom beter achterwege blijven.

LINEAIRE REGRESSIE

Wanneer we aan de waarden van advies geen scores toekennen, dan kunnen we ook geen lineaire regressie technieken gebruiken in de eerste stap. Voordat we bespreken wat dan wel mogelijk is, maken we een paar opmerkingen bij de door Van den Eeden en Saris uitgevoerde analyses. De drie modellen op pag. 169 hebben in feite drie typen parameters: de intercepten, de hellingen, en de variantie van de storingsterm. Over deze laatsten wordt niets meegedeeld, maar ook over deze varianties kan aangenomen worden dat ze òf gelijk zijn òf variëren in contexten. Dat men LISREL gebruikt om regressieanalyse te doen heeft een aantal nadelen. De belangrijkste is, dat de exacte toetsing door middel van F-ratios vervangen wordt door benaderende chi-kwadraat toetsing, zonder dat dit nodig is. Als men normaliteit aanneemt, dan kan men het beste natuurlijk alle consequenties ervan gebruiken. Bij gebruik van chi-kwadraat verliest men veel onderscheidend vermogen (power).

Bestudering van tabel 2 laat nog iets interessants zien. De hellingkolom en de intercept-kolom zijn bijna perfect gecorreleerd. Dit betekent dat de regressielijnen allemaal bij benadering door het punt gaan dat overeenkomt met het totaal-gemiddelde over alle scholen. Het is echter niet uit te maken of we hier een interessant empirisch gegeven ontdekt hebben (gemiddelde intelligentie krijgt op alle scholen een gemiddeld advies), of dat dit feit een gevolg is van de poging om de regressiecurve over het gehele IQ-continuum met een rechte lijn te benaderen.

Het door ons voorgestelde alternatief is om de regressieanalyses in de eerste stap van de multilevel procedure over te doen met een techniek die geen a priori scores voor de waarden van advies gebruikt. Men kan hierbij denken aan gegroepeerd continue methoden (McCullagh, 1980, De Leeuw, 1984), of aan logistische methoden (Meester en De Leeuw, 1983, appendix C, Anderson, 1983). En eventueel zelfs aan vormen van correspondentieanalyse (Meester en De Leeuw, 1983, appendix A). Bij gegroepeerd continue en logistische methoden zijn dezelfde modeltoetsen mogelijk als bij LISREL, zonder dat we lineaire regressie, normaliteit, of zelfs a priori scoring nodig hebben. Bij correspondentieanalyse kan lineariteit van de regressie onderzocht en getoetst worden, en is homogeniteit van de regressie tussen scholen ook eenvoudig vast te stellen.

Samenvattend kunnen we hier zeggen, dat zelfs wanneer de LISREL-modeltoets toepasbaar zou zijn de gebruikelijke OLS-toetsing via F-toetsen betere informatie zou geven. Er is echter geen sprake van lineaire regressies.

Van den Eeden en Saris springen bovendien nogal slordig met de storingstermen om. Die worden als het ware toegevoegd omdat dat nu eenmaal hoort, maar wat er precies over gespecificeerd wordt in het model blijft geheel in het vage.

DE TWEEDE STAP

De regressiecoëfficienten uit de eerste stap worden nu gerelateerd aan schoolvariabelen, in deze toepassing overigens allemaal geaggregeerde individuele variabelen. Van den Eeden en Saris proberen de hellingen te voorspellen, maar vanwege de grote correlatie tussen hellingen en intercepten zou voorspellen van de intercepten niet erg verschillende resultaten opleveren. Het is niet verrassend, dat Van den Eeden en Saris in de tweede stap de minder gelukkige keuzen uit de eerste fase herhalen. Zo worden beroep, opleiding, schoolwens allemaal van integerscores voorzien voor ze geaggregeerd worden. Dat de resulterende variabelen geen duidelijke samenhang vertonen met de regressie op schoolniveau kan weer een empirisch interessant gegeven zijn, maar het kan ook liggen aan de eigenaardige constructie van de predictoren. In tabel 3 van Van den Eeden en Saris zien we, tot onze verbazing, dat ook het gemiddelde advies gebruikt is om de regressiecoëfficiënt te voorspellen. Bovendien blijkt uit de discussie dat uit de tweede stap het gemiddeld IQ als beste voorspeller van de regressiecoëfficiënt naar voren komt. Het lijkt ons dat de gevolgde procedure de gemiddelde binnen-scholen regressie en de tussen-scholen regressie van advies op IQ op een uiterst ingewikkelde manier door de war husselt. En dat terwijl de twee soorten regressies toch op een uiterst eenvoudige manier apart te berekenen zijn (zie het werk van Cronbach, bijvoorbeeld).
Wat er precies gebeurt is moeilijk in statistische termen te beschrijven, omdat met name de eigenschappen van de storingstermen niet vastgelegd worden. Wel is het duidelijk zo, dat de hellingen in de eerste stap parameters zijn, dat wil zeggen vaste constanten die geschat moeten worden. In de modellen van de tweede stap zijn de hellingen stochasten die verklaard moeten worden. Wanneer we veronderstellen dat de hellingen in de tweede stap de schattingen van de parameters uit de eerste stap zijn, dan zijn deze specificaties consistent. Maar dan is het natuurlijk wel zo, dat de helling voor een bepaalde school een variantie heeft die afhangt van het aantal leerlingen op de school, van de IQ-variantie op de school, en van de variantie van de sto-

ringsterm op die school in de eerste fase. In ieder geval is de variantie niet constant over scholen, zodat de gebruikelijke statistische informatie uit OLS-regressie-programma's onbruikbaar is.
Nadere inspectie van het GALO-materiaal laat zien dat, als we gebruik maken van de integer-scoring voor advies, er een hoge correlatie is tussen gemiddeld advies per school en variantie van advies per school. Hoe hoger de school, hoe gedifferentieerder het advies. Zonder gebruik te maken van scoring is het direkt uit het GALO-materiaal te destilleren dat 'lage' adviezen samengaan met een relatief hoge IQ-variantie. De variantie van het advies per school is een bijna even goede voorspeller van de helling van de regressiecoëfficiënt als het gemiddeld IQ per school. Al deze eenvoudige effecten, die direkt uit descriptieve statistieken te voorschijn komen, hebben natuurlijk invloed op de regressieanalyses, en maken ze verwarrend en nauwelijks bruikbaar.

INTEGREREN DER STAPPEN

Van den Eeden en Saris proberen op pagina 171-176 van hun artikel de uitkomsten van de twee stappen te combineren. Dit gaat voor een deel theoretisch, voor een deel via schattingen. De afleiding, die uit vergelijking (8) via (9) naar (10) leidt, is curieus. De gemiddelde storingsterm, een stochast dus, wordt verondersteld gelijk aan nul te zijn. Zowel de betekenis als de functie van deze aanname zijn ons niet duidelijk. Vergelijking (11) wordt plotsklaps van een nieuwe storingsterm voorzien, blijkbaar anders dan die in (8), een storingsterm die 'alle weggelaten variabelen representeert' (lc, pag. 175). Ook van deze ingreep is de betekenis onduidelijk, te meer omdat na een nieuwe substitutie in (12) weer de oude storingsterm te voorschijn komt. Zoals we al eerder aangaven, wordt er in het artikel slordig omgesprongen met de storingstermen. Dit is natuurlijk riskant, omdat in de gebruikelijke behandelingen van regressie de statistische eigenschappen van tests en schatters geheel door de aannamen omtrent storingstermen bepaald worden. Dit zijn immers de enige stochastische componenten in het systeem.
Wanneer we eerste en tweede stap in één vergelijking (model) combineren, dan moeten we niet alleen rekening houden met algebraïsche combinatie van predictoren, maar ook met de invloed van de combinatie op de eigenschappen van de storingsterm. Als we in simultane vergelijkings-

modellen overgaan op de gereduceerde vorm, dan wordt ook de storingsterm meegetransformeerd. Een zelfde resultaat geldt hier natuurlijk ook. Als we rekening houden met de aard van de storingsterm, dan kunnen we de gebruikelijke OLS-schatters corrigeren voor de door Van den Eeden en Saris genoemde onzuiverheid. Wanneer het model bekend is, dan kan men de juiste schattingsprocedure erbij uitzoeken. Het is hierbij van geen belang of het model nu in twee afzonderlijke stappen, of (door combinatie van de twee) direkt in één stap gespecificeerd wordt.

CONCLUSIES

Gebruik van een twee-staps procedure voor multilevel onderzoek is nuttig, en theoretisch valt er weinig op aan te merken. Niettemin is het noodzakelijk, dat de eigenschappen van het model op ieder van de niveau's zo nauwkeurig mogelijk gespecificeerd worden, en dat men op hogere aggregatieniveau's (bij latere stappen, dus) rekening houdt met de aannamen op lagere niveau's (bij eerdere stappen). Het is bovendien zaak om direkt bij de eerste stap al een geschikt model, plus een geschikte techniek, te kiezen, omdat keuzes in de eerste stap vanzelfsprekend in de verdere stappen doorwerken.
Ieder van de stappen op zich opereert op slechts één niveau, en is daarom geen multilevel techniek. Combinatie van de stappen achteraf kan tot multilevel uitspraken leiden. Het is evenwel niet zo, dat de twee-traps schatters over het algemeen beter zijn dan de één-traps-schatters in een goed gespecificeerd model. Evenmin als schatten voor iedere vergelijking apart in simultane modellen noodzakelijk optimaal is. Uiteindelijk kan de maximale hoeveelheid informatie gebruikt worden, wanneer alle submodellen op de juiste manier gecombineerd worden.
Van den Eeden en Saris bevelen aan het multilevel model in twee stappen af te leiden. Men kan deze aanbeveling interpreteren op zuiver theoretisch niveau: leidt deelmodellen af voor beide niveau's, en combineer ze vervolgens tot een compleet multilevel model. Men kan de aanbeveling ook lezen als een aanmoediging tot 'data snooping' of 'data massage'. Bij deze laatste lezing bevat de aanbeveling duidelijke gevaren, omdat de risico's van kanskapitalisatie bij a posteriori selectie van predictoren zeer groot zijn.
Onze belangrijkste conclusie is echter, dat modellen voor kwantitatieve variabelen in het schoolloopbaanonderzoek een beperkte toepasbaar-

heid hebben. Voordat men overgaat tot het fitten van regressie, factor, of LISREL modellen is het over het algemeen aan te bevelen goed de eenvoudige descriptieve statistieken te bestuderen, en naar de vorm van de regressiefunctie te kijken. Wanneer men een variabele numeriek maakt, door één of andere vorm van a priori scoring te gebruiken, dient men zich goed te realiseren waarom men dit doet, en waarom men juist deze scoring kiest. Alle verdere modelspecificaties zullen immers van de gekozen scoring afhankelijk zijn. Tenslotte willen wij er nogmaals, wellicht ten overvloede, op wijzen dat de storingsterm een integraal deel van het regressiemodel is, en dat men bij combinatie of transformatie van modellen de storingsterm natuurlijk mee moet transformeren of combineren.
In het artikel van Van den Eeden en Saris wordt volgens ons slordig omgesprongen met het meetniveau van de variabelen en met de storingstermen van de regressievergelijkingen. Bovendien wordt tussen de twee stappen een (in de contextuele analyse min of meer gebruikelijke) geheimzinnige transmutatie uitgevoerd, die parameters uit de eerste stap veranderen in variabelen van de tweede stap. Bij combinatie tot een samengesteld multilevel model levert dit een regressiemodel op met stochastische regressiegewichten, of een niet-lineair regressiemodel. Op grond van onze discussie kunnen wij het onmogelijk eens zijn met het vermoeden dat twee-staps schatters beter zullen zijn dan één-staps schatters. De inhoudelijke conclusie dat de 'wijze waarop het schoolhoofd het IQ van de individuele leerling hanteert bij het vormen van zijn oordeel over studie-advies in belangrijke mate afhankelijk is van het IQ-gemiddelde en in mindere mate van het advies-gemiddelde van zijn school' vinden wij door de analyse van Van den Eeden en Saris zelfs niet aannemelijker geworden.

REFERENTIES

Anderson, J.A. Regression and ordered categorical variables. Journal of the Statistical Society, B, 1984, 46, 1-30.

Blok, H. & Saris, W.E. Relevante variabelen bij het doorverwijzen na de lagere school; een structureel model. Tijdschrift voor Onderwijsresearch, 1980, 5, 63-79.

Burstein, L. The analysis of multilevel data in educational research and evaluation. Review of Research in Education, 1980, 8, 158-233.

De Leeuw, J. Discrete normal linear regression models. In Dijkstra, T. (Ed.), Misspecification Analysis, New York, Springer Verlag, in press.

De Leeuw, J. & Stoop, I. Secondaire analyse 'Van Jaar tot Jaar' met behulp van niet-lineaire multivariate technieken. In Peschar, J.L. (ed.), Van Achteren naar Voren, Den Haag, Staatsuitgeverij, 1979.

Dronkers, J. & Schijf, H. Buurten, scholen en individuele onderwijsloopbanen; een beter analyse-model? Paper voor de Sociologendagen, 1984.

Hannan, M.T. Aggregation and Disaggregation in Sociology. Lexington, Mass, Heath, 1971.

Langbein, L.I. Schools or students: aggregation problems in the study of student achievement. Evaluation studies review annual, 2, 1977, 270-298.

Lazarsfeld, P.F. & Menzel, H. On the relation between individual and collective properties. In Etzioni, A. (Ed.), Complex Organizations a Sociological Reader. New York, Holt, Rinehart, & Winston, 1961.

McCullagh, P. Regression models for ordinal data. Journal of the Royal Statistical Society, B, 1980, 42, 109-142.

Meester, A.C. & De Leeuw, J. Intelligentie, Sociaal Milieu, en de Schoolloopbaan. Leiden, Vakgroep Datatheorie FSW/RUL, 1983.

Peschar, J.L. Schoolloopbanen in het voortgezet onderwijs. Den Haag, Staatsuitgeverij, 1983.

Tesser, P. Schoolloopbaanonderzoek in Nederland. Nijmegen, ITS, 1981.

Van den Eeden, P. & Saris, W.E. Empirisch onderzoek naar multilevel uitspraken. Mens en Maatschappij, 1984, 59, 165-178.

Van Praag, B.M.S. Model-free regression. Economics Letters, 1981, 7, 139-144.