KM 2(1981)2
pag 156-160

MULTIVARIATE ANALYSEMETHODEN VOOR DISCRETE VARIABELEN, EEN VERVOLG: KIJKEN NAAR RESIDUEN

Albert Verbeek ')

Samenvatting

Deze notitie zet de informele analyse voort uit de vergelijking van analysemethoden, gepresenteerd door Israëls e.a. We zoeken voor de tabel Gemiddeld Inkomen x (Scholing x Leeftijdsklasse x Bedrijfsklasse) een simpel model en een goede presentatie, waarbij het bekijken van de residuen een belangrijk hulpmiddel is.

1. Inleiding

In /2/ presenteren Israëls e.a. een boeiende vergelijking van een aantal van de belangrijkste analysemethoden voor multivariate discrete data. Als doel stellen zij zich (zie sectie 2.2) "om verschillende methoden te kunnen tonen en met elkaar te vergelijken aan de hand van een willekeurig bestand ... De analyse had niet tot doel om een relevante bijdrage te leveren tot het onderzoek naar de loonstructuur." Bovendien was bij het leeuwedeel der gepresenteerde analyses niet bekend of de te verklaren variabele I ordinaal dan wel nominaal was. Hier gebruiken we dat I ordinaal is, en we kiezen een iets ander doel: namelijk om te laten zien hoe een eenvoudige analyse van residuen het beeld kan aanvullen en verfijnen dat geschetst wordt in de hoofdstukken "Kijken naar meerdimensionale tabellen" en "Regressieanalyse". In hoeverre deze informatie ook te krijgen is uit andere hoofdstukken wordt aan de lezer ter beoordeling overgelaten.

We vatten de draad op bij tabel 3.5. (zie pag. 101). Deze beschrijft de invloed van SxLxB op het "gemiddelde" van I. Dat wil zeggen de ordinale variabele I kan de waarden 10, 20, 30, 40 en 50 aannemen, we vatten deze vervolgens op als "echte getallen", en tabel 3.5. geeft voor elk van de 3x4x6 cellen van SxLxB 1000xhet gemiddelde van I in die cel. *(Als we voor de presentatie gebruik maken van de kennis omtrent de interpretatie van I,S,L en B zullen we dit cursief doen. Het juist genoemde gemiddelde is een ruwe benadering voor het gemiddelde inkomen in duizenden guldens.)* We maken hier dus zeer nadrukkelijk gebruik van onze kennis dat I ordinaal is. Een gegeven, dat voor de Ree in hoofdstuk 3 eerst zeer kundig gereconstrueerd was, en dat bij de presentatie van tabel 3.5 - in de terminologie van /2/ - "ook al onthuld was".

2. Gevolgde methode

Onze analyse verloopt nu als volgt:

a. Pas Ehrenbergs richtlijnen voor het presenteren van tabellen toe. Zie /1/, Chapters 1 & 2, en vooral 1.9.
b. Negeer de aantallen waarnemingen per cel, en vat de tabel op als de data van een drieweg variantie-analyse met 1 waarneming/cel. Bereken de hoofdeffecten en de residuen t.o.v. deze hoofdeffecten.
c. Zie of de hoofdeffecten interessant zijn (d.w.z. "groot genoeg") en of de residuen nog interessante interactiepatronen vertonen.

Ad a. Ehrenberg

Hoewel de richtlijnen van Ehrenberg door velen voor triviaal gehouden worden, kunnen zeer veel gepubliceerde tabellen m.b.v. zijn richtlijnen veel informatiever gemaakt worden. De volgende punten kunnen we met succes

') Sociologisch Inst. & Inst. v. Math. Stat. Rijksuniversiteit Utrecht

op tabel 3.5 loslaten: (de nummering volgt /1/)
(i) See if the number of digits shown can be reduced. (Mental arithmetic is difficult with more than two significant digits, i.e. ones which vary).
(iv) Use averages (hier de genoemde hoofdeffecten) to help focus the eye when examining any array of numbers.
(vii) See if the columns or rows can be ordered ... to reflect their average size (Hiermee ordenen we de categorieën van S, L en B. We zullen "later" - na onthulling - wel zien of we deze ordening kunnen interpreteren).
(x) Summarise irregular aspects of the data statistically, e.g. by a measure of the average size of the residuals. (D.w.z. probeer een simpel model te vinden, bekijk de residuen t.o.v. dit model, en als daar nog "duidelijke" patronen in zitten, verfijn dan het model. Zijn de residuen "volledig random", dan hoeven ze niet meer apart vermeld te worden, maar is alleen het feit dat ze er willekeurig uitzien + hun gemiddelde absolute grootte o.i.d. interessant).

Ad b. Drieweg variantie-analyse met 1 waarneming/cel.
Doorslaggevend voor het negeren van het aantal waarnemingen per cel is de eenvoud van een (meervoudige) variantie-analyse met 1 waarneming/cel waarbij de nadruk ligt op schatten, niet op toetsen. Goede andere redenen zijn:
- Gezien de grote aantallen zijn we niet erg geïnteresseerd in toetsen, en zelfs nauwelijks in betrouwbaarheidsintervallen. Berekenen van enkele standaard fouten $s_{\bar{I}}$ m.b.v. de bijlage uit /2/ leert dat deze tussen ongeveer 0.05 en 1.62 liggen. Slechts zes cellen hebben een standaardfout groter dan 0.5, namelijk de SxLxB combinaties 312, 322, 324, 332, 342 en 344. Deze standaardfouten zijn zo klein dat voor een eenvoudig model de systematische afwijkingen waarschijnlijk groter zijn dan deze stochastische afwijkingen.
- Variantie-analyse op de oorspronkelijke waarnemingen levert schattingen die optimaal zijn (o.a. in de zin van een minimaal risico bij voorspellen onder een kwadratische verliesfunctie) t.o.v. de oorspronkelijke populatie. Daarbij tellen fouten in cellen met veel waarnemingen zwaarder, dan fouten in cellen met weinig waarnemingen. Bij een beschrijving van $\bar{I}$ als functie van S, L en B zou het kunnen zijn dat men alle cellen van SxLxB gelijk wil behandelen. Dat is wat hier gebeurt.

Een van de spelregels uit "Kijken naar meerdimensionale kruistabellen" was dat bij de analyse slechts gebruik gemaakt mocht worden van een zakrekenmachientje. Deze regel is ook hier gehanteerd.

3. Analyse

Tabel 1 is een beter leesbaar gemaakte versie van de te analyseren tabel. Hiertoe zijn o.a. de getallen afgerond op twee "significante" (d.w.z. variërende) cijfers. Voor verdere berekeningen maken we natuurlijk gebruik van minder afgeronde getallen (drie cijfers is voldoende). In de marges zijn het gemiddelde en de hoofdeffecten aangegeven, zodat gemakkelijk afgelezen kan worden of een cel naar boven of naar beneden afwijkt t.o.v. het model met alleen hoofdeffecten (in gangbare notatie:)

$$\mu_{slb} = \mu + \alpha^{S}_{s} + \alpha^{L}_{l} + \alpha^{B}_{b} \qquad (s=1,2,3;\ l=1,\ldots 4;\ b=1,\ldots 6)$$

De afwijkingen $\bar{I}_{slb} - \hat{\mu}_{slb}$ sataan in tabel 2. Terwijl de $\bar{I}_{slb}$ in tabel 1 een spreiding hebben van 7.1, hebben de residuen in tabel 2 nog slechts een spreiding van 1.8. In termen van verklaarde variatie: deze bedraagt $1-(18/71)^2 =$ $= 0.94$. In /2/ kan men een soortgelijk resultaat terugvinden in tabel 9.1 (p 138):

$$R^2_{S+L+B} \;/\; R^2_{SxLxB} = 0.344/0.355 = 0.97 \ .$$

Het verschil met de hier gevolgde procedure is, dat in /2/ gewerkt wordt met verschillende aantallen waarnemingen/cel, en met gewone regressieanalyse en canonische regressieanalyse in plaats van, zoals hier, variantieanalyse. Nu is gewone regressieanalyse natuurlijk wat restrictiever dan het variantieanalysemodel (1 parameter per verklarende variabele tegen 1 parameter per

categorie van een verklarende variabele), terwijl canonische regressieanalyse minder restrictief is, namelijk met 1 parameter per categorie zowel van de verklarende variabelen als van de te verklaren variabele.

Tabel 1

		B 2 *kleding leer*	4 *basis metaal*	5 *metaal, transport*	6 *voeding*	3 *overig*	1 *textiel*	gemiddelde + hoofd-effecten van S en L
S	L							
1	1	16	15	19	18	18	19	15.4 = 28.0 - 5.0 - 7.6
	2	23	21	24	24	26	25	23.6 = " " + 0.6
	3	23	23	25	26	27	28	26.4 = " " + 3.4
	4	23	25	26	26	27	28	26.7 = " " + 3.7
2	1	17	18	20	20	20	21	18.8 = 28.0 - 1.6 - 7.6
	2	26	25	25	27	28	28	27.0 = " " + 0.6
	3	30	28	30	30	31	31	29.8 = " " + 3.4
	4	29	27	30	29	31	33	30.1 = " " + 3.7
3	1	23	26	25	24	25	24	27.1 = 28.0 + 6.7 - 7.6
	2	37	37	34	35	36	35	35.3 = " " + 0.6
	3	33	39	39	40	41	41	38.1 = " " + 3.4
	4	33	37	40	41	42	44	38.4 = " " + 3.7
hoofdeffecten B		-1.8	-1.3	0.1	0.4	1.2	1.7	

Net als tabel 3.5 uit /2/ (deze KM pag 101) bevat bovenstaande tabel voor elke combinatie van S, L en B *(Scholing, Leeftijdsklasse en Bedrijfsklasse)* $\bar{I}$, het gemiddelde van I *(Inkomen, getransformeerd naar een vijfpuntsschaal)*, in die cel. De getallen zijn afgerond op twee cijfers. De categorieen var. S, L en B (waarvan niet bekend is of ze ordinaal zijn!) zijn gerangschikt naar opklimmende gemiddelde waarde van I.

De ontbinding van KS_{tot} in $KS_S + KS_L + KS_B + KS_{res}$ geeft hier:

$$3582 = 1733 + 1505 + 119 + 225.$$

De standaarddeviatie van de 72 $\bar{I}$'en in de tabel is dus $\sqrt{(3582/72)} = 7.1$. Ter vergelijking hebben we uit de bijlage van /2/ afgeleid, dat de 72 standaard deviaties van de verdelingen van I binnen elk der cellen liggen tussen 1.4 en 3.6.

Tabel 2.

		B 2 *kleding leer*	4 *basis metaal*	5 *metaal, transport*	6 *voeding*	3 *overig*	1 *textiel*	rijgemiddelde
S	L							
1	1	3	1	4	2	1	2	2.2
	2	1	-2	0	0	1	-1	0.0
	3	-2	-2	-2	-1	-1	-1	-1.2
	4	-2	-1	-1	-1	-1	-1	-1.1
2	1	1	0	1	1	0	1	0.5
	2	1	-1	-2	-1	0	-1	-0.6
	3	2	-1	0	1	0	0	0.2
	4	1	-2	0	-1	0	1	-0.2
3	1	-2	0	-2	-4	-3	-5	-2.7
	2	4	3	-2	-1	-1	-1	0.5
	3	-3	3	1	2	2	1	1.0
	4	-3	0	2	2	2	4	1.2

Bovenstaande tabel bevat de residuen uit tabel 1 ten opzichte van het variantieanalysemodel zonder interacties. De standaardafwijking van de 72 gktallen in tabel 2 is $\sqrt{(225/72)} = 1.8$.

De residuen in tabel 2 vertonen weinig opzienbarends. Er zijn geen uitschieters bij. Er zijn wat kleine regelmatigheden zoals een eerste rij van louter +'en, een tekenvastheid die we ook vinden bij de vierde, vijfde en negende rij, en in de groepen B4, S3 en B2, S2. Dat wil zeggen dat de inkomens in dergelijke groepen systematisch iets (hooguit 2 à 3) afwijken van wat het hoofdeffectenmodel aangeeft. Deze en andere interacties zijn duidelijker te zien als we de tweedimensionale marginalen van tabel 2 berekenen. Voor SxL is dit reeds in de marge van tabel 2 gedaan. Voor SxB en LxB doen we dit in tabel 3 hieronder, behalve dan dat we voor de vergelijkbaarheid niet de rijsommen etc. berekenen, maar de rijgemiddelden. Het enige andere punt dat mij nog opviel in tabel 2, is dat binnen elke SxB groep de residuen een tamelijk monotoon verloop tonen, behalve bij S3, B2. Dit monotone verloop correspondeert met de gesignaleerde tekenvastheid van verschillende rijen, en is ook in de rijgemiddelden goed terug te vinden. Bij S1 is de stijging in $\bar{I}$ langs L1-L4 duidelijk minder dan bij S3. De oorzaak van de afwijking bij S3,B2 kunnen we gemakkelijk herkennen in tabel 1: $\bar{I}$=37 bij S3, L2, B2 is verrassend hoog t.o.v. de getallen eronder en erboven. Inderdaad is het de enige plaats in de tabel waar L2 een gemiddeld hogere I heeft dan L3. Inspectie van de bijlage uit /2/ leert dat L2 en L3+L4 (bij S3,B2) elk op ongeveer 100 waarnemingen berusten, terwijl de standaardfout van deze $\bar{I}$'en ongeveer 1 is. Als men bedenkt dat die honderd waarnemingen dan (mogelijk) een opgehoogd aantal betreffen, dan zou het kunnen dat deze eigenaardigheid een steekproefeffect is. (We kunnen dit effect ook herkennen in fig. 4.1 van /2/).

Tabel 3.

De volgende twee tabellen zijn ontstaan door de (niet afgeronde) residuen te middelen over L, respectievelijk over S.

tabel 3.a — B

	2 *kleding leer*	4 *basis metaal*	5 *metaal, transport*	6 *voeding*	3 *overig*	1 *textiel*
S 1	0.1	-0.9	0.3	0.3	0.2	0.1
2	1.0	-0.8	-0.1	-0.1	-0.1	0.1
3	-1.1	1.6	-0.2	-0.1	-0.1	-0.3

tabel 3.b — B

	2 *kleding leer*	4 *basis metaal*	5 *metaal, transport*	6 *voeding*	3 *overig*	1 *textiel*
L 1	0.4	0.4	0.8	-0.2	-0.7	-0.7
2	2.0	0.2	-1.0	-0.3	0.0	-1.1
3	-0.6	0.0	-0.2	0.6	0.4	0.1
4	-0.8	-0.7	0.3	0.0	0.2	1.5

In beide tabellen 3 valt allereerst op dat deze gemiddelden van de residuen aanzienlijk kleiner zijn dan bij SxL (vermeld in de marge van tabel 2). Wellicht zou het nog de moeite waard zijn om na te gaan of er iets bijzonders aan de hand is bij B4, S3 (als we bij een echt onderzoek over die gegevens beschikten; codeer- of rekenfout, toevallige afwijkingen in de steekproef, of "echte" bijzondere omstandigheden?)

In tabel 3.b zien we hetzelfde effect als bij SxL, alleen zeker 2x zo zwak: in de B-categorieën met een klein hoofdeffect is de stijging van $\bar{I}$ langs L1-L4 kleiner dan bij categorieën met een groter hoofdeffect. Dit soort interactie is vaak goed te modelleren met één vrijheidsgraag. Als je de (mogelijk nominale, mogelijk ordinale) variabelen S, L, en B opvat als ordinaal, met de categorieën gerangschikt naar grote van de hoofdeffecten, en vervolgens vat je de rangnummers van de categorieën op als gewone scores, dan kun je deze interactie op twee manieren beschrijven. Je kunt zeggen dat de hellingen van de regressielijnen van $\bar{I}$ op L binnen de groepen S1, S2 en S3 onderling verschillen; ze zijn des te groter naarmate S groter is. De tweede manier is om het product SL (of (S-2)(L-2½) voor meer symmetrie) als verklarende variabele in het model op te nemen. Vergelijk Tukeys "One degree of freedom for non-additivity" /3/, en /4/:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \lambda\alpha_i\beta_j + \underline{\text{residu}}_{ij}.$$

Hoewel dit model niet lineair is, geeft Tukey een handzame toets voor de nulhypothese van geen interactie: $\lambda = 0$.

4. Conclusies van de analyse

De te verklaren variabele I is gemiddeld 28.0 met een spreiding $s_I = 7\frac{1}{2}$. De gemiddelden $\bar{I}$ per combinatie van S, L en B hebben een spreiding $s_{\bar{I}} = 7.1$ en variëren van 15 (B4, S1, L1) tot 44 (B1, S3, L4). S en L hebben ongeveer even grote invloed; het gemiddelde verschil in I tussen eenheden uit de uiterste categorieën van S bedraagt:

$$\text{effect } S_3 - \text{effect } S_1 = 12,$$

en voor L is dit

$$\text{effect } L_4 - \text{effect } L_1 = 11.$$

Beide variabelen samen verklaren ongeveer 90% van de variatie in $\bar{I}$. Daarna zijn de belangrijkste effecten: de verschillen tussen de B-categorieën en de SxL-interactie, die elk ongeveer 3% variatie verklaren. (Vergelijk tabel 9.1 uit /2/).

Meer formeel en (dus) meer gedetailleerd zouden we de conclusies aldus kunnen formuleren.

We kunnen voor tabel 1 de volgende, steeds ingewikkelder en beter passende modellen opstellen.

$I_{SLB} + \bar{I}_{SLB} + res^{(0)}_{SLB}$, waarbij de spreiding $s_{res^{(0)}_{SLB}}$ van I binnen cel SLB ligt tussen 1.4 eb 3.6, en $S_{\bar{I}_{SLB}} = 7.1$

$$I_{SLB} = \begin{cases} 23.0 \text{ als } S = 1 \\ 26.4 \text{ als } S = 2 \\ 34.7 \text{ als } S = 3 \end{cases} + \begin{cases} -7.6 \text{ als } L = 1 \\ +0.6 \text{ als } L = 2 \\ -3.4 \text{ als } L = 3 \\ +3.7 \text{ als } L = 4 \end{cases} + res^{(1)}_{SLB}, \text{ met}$$

$$s_{res^{(1)}_{SLB}} = 2.1$$

Dit model kunnen we verder verfijnen:

$$res^{(1)}_{SLB} = (S\text{-}2). \begin{cases} -2.4 \text{ als } L = 1 \\ 0 \text{ als } L = 2 \\ 1.1 \text{ als } L = 3 \text{ of } 4 \end{cases} + \begin{cases} -1.8 \text{ als } B = 2 \\ -1.3 \text{ als } B = 4 \\ +0.1 \text{ als } B = 5 \\ +0.4 \text{ als } B = 6 \\ +1.2 \text{ als } B = 3 \\ +1.7 \text{ als } B = 7 \end{cases} + res^{(2)}_{SLB}$$

waarbij $s_{res^{(2)}_{SLB}} \approx 1.4$

Opmerkelijk zijn dan tenslotte nog de hoge waarden van $\bar{I}$ binnen B4, S3, en het verschil tussen L2 en de overige L-categorieën binnen B2, S3.

Literatuur

/1/ A.S.C. Ehrenberg - Data Reduction, John Wiley (1975)

/2/ A.Z. Israëls, J.G. Bethlehem, J. van Driel, M.E. Jansen, J. Pannekoek, S.J.M. de Ree & D. Sikkel - Multivariate Analysemethoden voor Discrete Variabelen, KM 2 (1981) 2, pag

/3/ P.R. Krishnaiah & M.G. Yochmowitz - Inference on the Structure of Interaction in Two-way Classification Model. Hfdst. 25, pag. 973-994 van P.R. Krishnaiah (ed.) - Handbook of Statistics, Vol. 1, North-Holland (1980)

/4/ J.W. Tukey - One degree of freedom for non-additivity. Biometrics, 5 (1949), 232.