KM 2(1981)2
pag 87-149

centraal bureau voor de statistiek

Hoofdafdeling Statistische Methoden
Postbus 959, 2270 AZ VOORBURG

MULTIVARIATE ANALYSEMETHODEN VOOR DISCRETE VARIABELEN

door

A.Z. Israëls (eindredaktie)
J.G. Bethlehem
J. van Driel
M.E. Jansen
J. Pannekoek
S.J.M. de Ree
D. Sikkel

Dit rapport werd gepresenteerd op de Economisch Statistische Dag 1980.
De in dit rapport weergegeven opvattingen zijn die van de auteurs en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek of met de opvattingen van andere stafmedewerkers van het C.B.S. De schrijvers danken J.A.A. de Beer, A. ten Cate, W.J. Keller en T.J. Wansbeek voor hun stimulerende opmerkingen.

februari 1981

MULTIVARIATE ANALYSEMETHODEN VOOR DISCRETE VARIABELEN

Samenvatting

Het C.B.S. houdt zich naast het leveren van getallen in toenemende mate bezig met het analyseren van grote data-bestanden. Voor het analyseren van dergelijke data-bestanden bestaan talloze technieken. De hoofdafdeling Statistische Methoden heeft een experiment uitgevoerd door een data-bestand te analyseren zonder kennis van de betekenis der variabelen. Het betrof hier een 3 x 4 x 6 x 5 frequentietabel van 4 discrete variabelen. Aan de onderzoekers was slechts bekend dat de variabele met 5 categorieën als de afhankelijke variabele kon worden beschouwd. Diverse technieken zijn op dit 'geheime bestand' losgelaten: 'kijken naar meerdimensionale tabellen', ridit-analyse, stapsgewijze selectiemethoden m.b.v. het G^2-criterium, log-lineaire analyse, clusteranalyse, correspondentieanalyse en regressie-analyse.

INHOUD

1. INLEIDING

Veel informatie over economische en sociologische verschijnselen komt beschikbaar in de vorm van kruistabellen. Vooral het C.B.S. produceert jaarlijks grote hoeveelheden cijfers in tabelvorm, waarbij dikwijls sprake is van grote en op het oog moeilijk analyseerbare meer-dimensionale tabellen. In dit rapport wordt aan de hand van een min of meer willekeurige, vier-dimensionale, tabel getoond welke methoden zoal beschikbaar zijn om in een dergelijke situatie verbanden op te sporen en te typeren.

Juist tegenover een gehoor van economen en sociologen lijkt het geboden de zinvolheid van het gebruik van dit soort analysemethoden te rechtvaardigen. In de sociale wetenschappen en zeker in de economie bestaat nl. soms een zeker wantrouwen tegen exploratieve data-analyse. Immers, zo gaat de redenering, het onderzoeken van het verband tussen maatschappelijke verschijnselen dient te geschieden op basis van een model waarin, uitgaande van veronderstellingen omtrent het menselijk gedrag, afhankelijke en onafhankelijke variabelen worden aangegeven, hun functionele relatie wordt afgeleid, en een onderliggend kansmodel gespecificeerd. De parameters van een dergelijk model worden vervolgens geschat, waarbij zich dikwijls technische problemen zoals dynamiek en simultaniteit voordoen. Deze problemen worden nog vergroot wanneer de voor het schatten benodigde gegevens slechts in de vorm van een kruistabel, dus in de vorm van gegroepeerde data beschikbaar zijn. Zijn deze technische problemen opgelost, dan verkrijgt men in principe asymptotisch rake schatters van de modelparameters en kan men de juistheid van het model toetsen en voorspellingen doen over de effecten van een verandering in de onafhankelijke op de afhankelijke variabelen.

In de praktijk is de situatie gewoonlijk een stuk gecompliceerder. Los van de al genoemde technische problemen is de aanpak een stuk minder rechttoe-rechtaan dan beschreven; een belangrijk knelpunt is het gewoonlijk ontbreken van een gedragstheorie die voldoende rijk en specifiek is om ook maar in enig detail de formulering van het te schatten model te kunnen dicteren. Voordat men aan een dergelijke aanpak toekomt kan het gewenst zijn eerst enig zicht te krijgen op de patronen die in de waarnemingen aanwezig zijn, zodat men bij het specificeren van een model enig houvast heeft.

Voordat men, anders gezegd, aan de confirmatieve fase toekomt, dient dikwijls eerst een exploratieve fase plaats te vinden. Wanneer men over

gegevens in tabelvorm beschikt, houdt deze exploratieve fase in dat gezocht moet worden naar de meest saillante relaties tussen de variabelen. Wanneer het gaat om een eenvoudige twee-dimensionale kruistabel hoeft dat nog geen grote problemen op te leveren; als het echter gaat om grote, meer-dimensionale tabellen is dat meestal een gecompliceerde bezigheid. Het doel van dit rapport is om te laten zien welke methoden zoal kunnen worden toegepast bij het analyseren van meer-dimensionale tabellen teneinde de waargenomen werkelijkheid te beschrijven met een klein aantal relaties.

De nadruk ligt daarbij op het 'typeren', dus op data-reductie. 'Toetsen' op significantie van de opgespoorde effecten speelt daarbij een ondergeschikte rol; werkt men, zoals bij het C.B.S. veelal gebruikelijk, met grote databestanden, dan zal een bepaalde a priori restrictie vrijwel zeker worden verworpen.

Dit probleem is o.a. door Berkson (1938) behandeld en door Leamer (1978, p.89) geparafraseerd: "Since a large sample is presumably more informative than a small one, and since it is apparently the case that we will reject the null hypothesis in a sufficiently large sample, we might as well begin by rejecting the hypothesis and not sample at all."

2. HET BESTAND

2.1. Het geheime bestand

Een groot probleem dat zich bij het toepassen van exploratieve technieken voordoet is de omstandigheid dat het voor de onderzoeker moeilijk is objectief te werk te gaan. De vraag is in hoeverre hij zich bij de interpretatie van de resultaten laat leiden door zaken die hijzelf interessant vindt, die hij er uit had willen krijgen of die hem toevalligerwijs opvallen. Om een idee te krijgen in hoeverre exploratieve technieken vrij zijn van dergelijke subjectieve elementen hebben wij een 3 x 4 x 6 x 5 frequentietabel (zie de bijlage) geanalyseerd, in eerste instantie zonder te weten uit welk onderzoek de data afkomstig waren en welke variabelen het betrof. Er was slechts bekend dat de variabele met 5 klassen als afhankelijke variabele kon worden beschouwd, en dat 3 van de 4 discrete variabelen een ordinaal karakter hadden. Omdat niet bekend was welke 3 variabelen dit betrof (de klassen van deze variabelen waren vooraf gepermuteerd), werden aanvankelijk slechts technieken gebruikt die van toepassing zijn op nominale variabelen. Nadat de analyses op dit 'geheime

bestand' waren voltooid, werd de betekenis der variabelen onthuld, waarna nog enkele aanvullende analyses werden verricht.
De toegepaste technieken waren

- Kijken naar meerdimensionale tabellen (hoofdstuk 3)
- Ridit-analyse (" 4)
- Stapsgewijze selectiemethoden m.b.v. G^2 (" 5)
- Log-lineaire analyse (" 6)
- Clusteranalyse (" 7)
- Correspondentieanalyse (" 8)
- Regressie-analyse (" 9)

De technieken beschreven in hoofdstuk 4 (ridit analyse) en het eerste deel van hoofdstuk 9 (gewone regressieanalyse) werden pas toegepast nadat zekerheid was verkregen over de ordinaliteit van de afhankelijke variabele en de bijbehorende rangorde. Ook het laatste gedeelte van hoofdstuk 3 vond plaats na de 'onthulling' van het geheime bestand. De overige technieken werden alle toegepast op het geheime bestand. De uitkomsten hiervan behoefden geen revisie na de onthulling van de betekenis van de variabelen; de conclusies bleven gehandhaafd. Het bleek bij dit bestand van 4 variabelen dus mogelijk om de belangrijkste conclusies te trekken zonder kennis omtrent de variabelen en ordeningen van de klassen. We moeten hierbij opmerken dat de analyse slechts een voorbeeld betrof (met slechts 4 variabelen, ieder met een beperkt aantal klassen), zodat niet in zijn algemeenheid kan worden gesteld dat de gebruikte technieken tot juiste interpretaties leiden.

In dit rapport gaan we uit van de terminologie die bij het 'onthulde bestand' behoort, omdat aanduidingen als 'klasse 2 van variabele 3' de lezer weinig zullen aanspreken. We zullen daarom eerst de oorsprong van het bestand bekendmaken.

2.2. Het onthulde bestand

Bij de analyse van het geheime bestand werd aangenomen dat de gegevens waren op te vatten als een enkelvoudige aselecte steekproef uit een oneindige populatie. Na onthulling van het geheime bestand bleek dat de cijfers afkomstig waren uit het 'Loonstructuuronderzoek 1972', dat op een met ongelijke kansen getrokken steekproef was gebaseerd, omvattende 48 000 van de in totaal 231 833 werknemers in de industrie. De resultaten werden ons ter beschikking gesteld in de vorm van opgehoogde aantallen omdat, terecht, werd verondersteld dat de exploratieve technieken beter op de opgehoogde aantallen konden worden toegepast aangezien deze aantallen de populatie beter representeren. De lezer dient in gedachte te houden dat, waar wordt gesproken over 'significantie', wordt uitgegaan van een enkelvoudig aselecte steekproef van 231 833 werknemers uit een oneindige populatie.

De afhankelijke variabele was Inkomen (bruto-jaarloon), ingedeeld in 5 klassen. De klasse-indeling, met bedragen in duizenden guldens, en de bijbehorende frequenties staan in tabel 2.1. (De volledige frequentie-verdeling is te vinden in de bijlage).

Tabel 2.1. Frequentie-verdeling van de afhankelijke variabele

Variabele	Klasse-indeling	Aantal mensen
Inkomen (I) (in duizenden guldens)	1. < 12	9 219
	2. 12 - < 22	101 646
	3. 22 - < 32	80 832
	4. 32 - < 42	24 966
	5. ≥ 42	15 170
		231 833

De 3 onafhankelijke variabelen zijn beschreven in tabel 2.2.

Tabel 2.2. Frequentie-verdelingen van de onafhankelijke variabelen

Variabele	Klasse-indeling	Aantal mensen
Scholing (S)	1. laag	42 171
	2. middelbaar	156 283
	3. hoog	33 379
		231 833
Leeftijd (L)	1. 21 - 29 jaar	56 710
	2. 30 - 39 jaar	63 021
	3. 40 - 49 jaar	62 790
	4. 50 - 64 jaar	49 312
		231 833
Bedrijfstak (B)	1. textiel	10 993
	2. kleding, leer	7 563
	3. overig	75 195
	4. basis metaal	9 459
	5. metaal, transportmiddelen, elektrotechnisch	100 216
	6. voedings- en genotmiddelen	28 407
		231 833

In het verdere verloop van het rapport geven we de variabelen inkomen, scholing, leeftijd en bedrijfstak weer met de symbolen I, S, L en B. Klassen van variabelen worden veelal voorgesteld door een letter-cijfer combinatie, bijv. B6 bij de bedrijfstak voedings- en genotmiddelen. Deze klassen zijn op te vatten als dummy-variabelen.

Bij de meeste analyses beschouwt men de 3 x 4 x 6 x 5-frequentietabel als een (3 x 4 x 6) x 5-tabel, ofwel een 72 x 5-tabel, omdat er sprake is van 2 sets variabelen, t.w. de combinaties van de onafhankelijke variabelen, S x L x B, en de afhankelijke, I. De tabellen uit de bijlage kunnen worden opgevat als 3 x 4 x 6 x 5-tabellen, en tevens als 72 x 5-tabellen. De 72 x 5-frequentietabel wordt kortweg weergegeven door (SxLxB)xI. Door in deze tabel te sommeren over de 6 bedrijfstakken ontstaat een 12 x 5-frequentietabel, die wordt geschreven als (SxL)xI. Evenzo is bijv. S x I de 3 x 5-frequentietabel scholing x inkomen. Naast deze tabellen, waarin alle onderlinge verbanden zijn meegenomen tussen de beschouwde onafhankelijke variabelen, worden ook tabellen behandeld zonder interacties tussen de onafhankelijke variabelen. Wanneer we bijv. de drie variabelen S, L en B allen als onafhankelijke variabelen meenemen, leidt dit tot drie frequentietabellen van resp. 3 x 5, 4 x 5 en 6 x 5 die kunnen worden samengevoegd tot een (3 + 4 + 6) x 5-tabel, ofwel een 13 x 5-tabel. We geven deze tabel weer als (S + L + B) x I.

Voor de goede orde zij opgemerkt dat het doel van de analyse is om verschillende methoden te kunnen tonen en met elkaar te vergelijken aan de hand van een willekeurig bestand. De vergelijking vindt voornamelijk plaats in hoofdstuk 10. De analyse had niet tot doel om een relevante bijdrage te leveren tot het onderzoek naar de loonstructuur. Was dat het geval dan zouden meer onafhankelijke variabelen in het onderzoek zijn meegenomen, en zou bedrijfstak niet als zuiver onafhankelijke variabele zijn beschouwd. Het beperkt houden van het aantal variabelen vergemakkelijkt echter onze hoofddoelstelling, nl. het vergelijken van de resultaten van de verschillende technieken. Tevens is vanwege de opzet van het onderzoek nauwelijks gebruik gemaakt van het feit dat de categorieën van de variabelen I, S en L a priori een numerieke betekenis hebben. We merken in dit verband op dat inkomen , scholing en leeftijd 'harde' variabelen zijn. De in dit rapport gepresenteerde methoden zijn ook toepasbaar op 'zachte' variabelen als tevredenheid, gemeten via categorieën die lopen van 'zeer ontevreden' tot 'zeer tevreden'.

2.3. Enkele algemene opmerkingen

In dit rapport stelt n_{ij} de frequentie voor in cel (i,j) van een tweedimensionale tabel, p_{ij} de kans dat iemand in cel (i,j) zit in het achterliggende kansmodel, en $\hat{p}_{ij}$ de fractie in cel (i,j), d.w.z. $\hat{p}_{ij} = \frac{n_{ij}}{n}$ met n de totale frequentie. Randtotalen worden weergegeven met n_{i+}, n_{+j}, p_{i+} etc., waarbij '+' op de plaats van een index aangeeft dat over die index wordt gesommeerd. In de bijlage wordt de 72 x 5-datamatrix (SxLxB)xI tevens gegeven in de vorm van relatieve rijfrequenties n_{ij}/n_{i+} (in procenten uitgedrukt), aangevuld met rijtotalen n_{i+}.

Stochastische grootheden worden onderstreept, terwijl schatters en schattingen van parameters van een dakje (bijv. $\hat{p}_{ij}$) worden voorzien.

De beschrijvingen van de 7 technieken zijn door verschillende auteurs gemaakt. Er is meestal een zeer globale beschrijving van de techniek gegeven, waarna de techniek is toegepast. Elk hoofdstuk is afzonderlijk leesbaar; in enkele gevallen evenwel konden verwijzingen naar andere hoofdstukken niet uitblijven, vanwege het nauwe verband tussen de technieken.

Het bleek niet mogelijk om één bepaald boek over discrete multivariate analyse te vermelden waarin alle gebruikte technieken zijn te vinden. Enkele boeken over discrete multivariate analyse, die echter ieder bepaalde facetten ervan behandelen, zijn: Andersen (1980), Benzecri e.a. (1976), Bishop e.a. (1975), Fienberg (1977), Gifi (1980), Goldstein en Dillon (1978), Goodman (1978) en Lebart e.a. (1977).

3. KIJKEN NAAR MEERDIMENSIONALE TABELLEN:

Analyse van percentages, gemiddelden en gestandaardiseerde gemiddelden.

3.1. Algemeen

Bij het analyseren van een meerdimensionale kruistabel is het goed om eerst eens rustig te kijken hoe de tabel er uit ziet, alvorens allerlei ingewikkelde analyse-technieken toe te passen. In deze eerste fase van de analyse is het gewenst de tabel zo weer te geven dat de eigen aardigheden die de tabel bevat als het ware zo in het oog springen. Een gebruikelijke wijze van presenteren is de celfrequenties uit te drukken in percentages. Bij een meerdimensionale tabel kan dat op vele manieren gebeuren, namelijk op evenveel manieren als er (sub)totalen kunnen worden gevormd. Niet elke mogelijkheid is echter even interessant. Zie bijvoorbeeld Van den Ende (1971). Wanneer naast percentages ook de frequenties van de (sub)totalen waarop gepercenteerd is worden weergegeven, blijft alle informatie uit de oorspronkelijke tabel behouden. Een tweede transformatie die zonder verlies aan informatie kan geschieden is het veranderen van de volgordes waarin de klassen van de variabelen zijn weergegeven. Beide methoden, percenteren en permuteren, zijn toegepast om meer inzicht te krijgen in het geheime bestand. Nadat de betekenis van de variabelen onthuld was werd een nieuw analyseplan uitgevoerd, waarbij behalve percentages ook rekenkundige gemiddelden en direct en indirect gestandaardiseerde gemiddelden van het inkomen zijn berekend. De standaardisaties zijn uitgevoerd om gemiddelden van verschillende subpopulaties met verschillende verdelingen over achtergrondkenmerken vergelijkbaar te maken. Zie bijvoorbeeld ook Fleiss (1973).

3.2. Analyse van het geheime bestand

Van het geheime bestand was gegeven dat I de te verklaren variabele was. Dit gegeven bepaalde de beslissing om de vierdimensionale tabel weer te geven als een tweedimensionale tabel (SxLxB)xI, waarbij de 5 kolommen betrekking hebben op de klassen van I en de 3x4x6 rijen betrekking hebben op de klassen van de verklarende variabelen S,L en B. Er is vervolgens horizontaal gepercenteerd waardoor de verdelingen over de klassen van I voor de verschillende combinaties van S,L en B eenvoudig met elkaar kunnen worden vergeleken. Het resultaat is afgedrukt als tabel 3.2.
Regelmatige patronen blijken nog niet zichtbaar. Een reden kan zijn dat door vooraf uitgevoerde permutaties van de natuurlijke volgordes der klassen eventueel bestaande patronen versluierd zijn. Daarom is bekeken of er wel patronen zichtbaar worden na geschikte permutaties van de klassen van de

4 variabelen. Om wat meer orde aan te brengen zijn de marginale frequentietabellen SxI, LxI en BxI geconstrueerd. Het resultaat is vermeld in tabel 3.1. De volgordes van de klassen komen overeen met de bij het geheime bestand gegeven ordening. De klasseaanduiding komt overeen met het onthulde bestand. (Tussen haakjes het rangnummer in de kolom (per variabele)).

Tabel 3.1. De verdelingen van S,L en B naar I

	I2	I4	I1	I3	I5	Totaal = 100%
			% (rangnummers)			
S3	16 (1)	25 (3)	0 (1)	37 (3)	22 (3)	33 379
S1	59 (3)	4 (1)	6 (3)	28 (1)	3 (1)	42 171
S2	46 (2)	9 (2)	4 (2)	36 (2)	4 (2)	156 283
L2	41 (3)	11 (2)	1 (2)	43 (3)	4 (2)	63 021
L4	35 (2)	14 (3)	1 (2)	39 (2)	11 (3)	49 312
L1	72 (4)	1 (1)	15 (4)	12 (1)	0 (1)	56 710
L3	28 (1)	16 (4)	1 (2)	44 (4)	11 (4)	62 790
B1	34 (1)	15 (6)	1 (1)	42 (6)	8 (5)	10 993
B2	53 (6)	6 (1)	12 (6)	25 (1)	4 (1)	7 563
B3	40 (2)	13 (5)	4 (3)	35 (4)	8 (6)	75 195
B4	48 (5)	7 (2)	10 (5)	30 (2)	5 (2)	9 459
B5	46 (3)	10 (3)	2 (2)	36 (5)	6 (3)	100 215
B6	47 (4)	10 (4)	5 (4)	32 (3)	6 (4)	28 407
Totaal	101 646	24 966	9 219	80 832	15 170	231 833

Bij het bekijken van de rangnummers valt op dat de eerste kolom praktisch hetzelfde verloop heeft als de derde kolom, terwijl de overige kolommen grotendeels het tegengestelde verloop te zien geven. Dit betekent dat de categorieën van variabele I gesplitst kunnen worden in twee groepen nl. 1 en 3 en in 2, 4 en 5. Verder geldt dat de fluctuaties in de percentages in kolom 3 sterker zijn dan in kolom 1 en die van kolom 4 zwakker zijn dan die van kolom 2 en die weer zwakker dan die van kolom 5. Categorie 3 (onthuld I1) en categorie 5 (onthuld I5) lijken dus randcategorieën te zijn. Dit suggereert een ordening van de kolommen van I als volgt: 3, 1, 4, 2, 5. Achteraf bleken deze respectievelijk overeen te komen met de klassen I1, I2, I3, I4 en I5. Geef aan de klasse I1 en I2 de kwalificatie 'laag' dan zijn op grond hiervan ook de klassen van S,L en B te ordenen. De volgorde van S komt overeen met de rijen 2,3 en 1, welke klassen achteraf overeen bleken te komen met S1, S2 en S3; de gekozen volgorde van L werd

Tabel 3.2. Rijpercentages en -totalen van de (3x4x6x5)-tabel volgens de gegeven ordening

		%			totaal
6	37	0	46	12	752
7	25	0	39	28	127
9	32	0	42	17	4195
8	33	0	36	23	277
8	24	0	57	11	6004
16	34	0	34	15	888
1	32	0	12	55	318
34	32	0	17	17	47
5	28	0	21	47	1642
14	19	0	33	34	187
4	26	0	29	40	2201
6	19	0	28	46	354
60	3	0	37	0	424
77	1	0	21	1	81
54	10	3	33	0	1964
38	0	0	62	0	105
57	5	1	37	0	3224
63	6	4	27	1	599
1	35	0	25	38	612
23	17	0	40	19	77
3	33	0	24	40	3323
2	41	0	29	28	279
3	30	0	34	33	4908
4	31	0	26	38	791
52	4	2	41	0	326
73	6	4	17	0	494
57	4	2	30	7	2132
79	0	8	13	0	428
68	3	0	28	1	3374
69	5	0	23	2	1387
40	10	0	45	4	891
70	2	8	14	5	497
48	9	1	37	5	3760
54	4	3	36	2	903
57	7	0	32	4	5849
53	5	1	35	5	2196

(vervolg)		%			totaal
84	1	12	3	0	281
62	0	37	1	0	475
68	1	27	4	0	2813
49	0	51	0	0	433
74	0	19	7	0	2823
64	1	28	7	0	1641
37	6	0	54	3	649
71	3	5	18	3	578
43	9	0	45	3	3004
67	2	4	26	1	605
59	3	0	35	2	4770
50	6	2	40	3	1862
33	7	0	58	1	1728
51	6	3	36	4	1608
36	11	0	49	3	13981
57	6	2	34	2	1452
53	4	0	41	1	18766
43	9	2	44	3	5102
19	26	0	44	12	1400
43	12	2	33	11	870
28	19	1	42	11	10426
47	8	2	36	7	1326
29	14	0	48	8	12476
37	15	1	39	8	3969
81	0	5	14	0	1521
58	0	35	7	0	1752
74	1	15	10	0	14469
54	0	35	11	0	1673
80	1	10	9	0	17241
74	1	15	10	0	5191
17	20	0	59	4	2091
26	11	1	54	8	957
26	18	0	47	10	13486
40	8	0	45	6	1791
29	15	0	51	5	18580
29	17	1	45	9	4427

Tabel 3.3. Rij-percentages en -totalen van de (3x4x6x5)-tabel volgens de gevonden ordening (zie ook de tabel in de bijlage, die op grond van het onthulde bestand is gecreëerd).

		%			totaal
37	62	1	0	0	475
51	49	0	0	0	433
28	64	7	1	0	1641
19	74	7	0	0	2823
27	68	4	1	0	2813
12	84	3	1	0	281
4	73	17	6	0	494
8	79	13	0	0	428
0	69	23	5	2	1387
0	68	28	3	1	3374
2	57	30	4	7	2132
2	52	41	4	0	326
8	70	14	2	5	497
3	54	36	4	2	903
1	53	35	5	5	2196
0	57	32	7	4	5849
1	48	37	9	5	3760
0	40	45	10	4	891
5	71	18	3	3	578
4	67	26	2	1	605
2	50	40	6	3	1862
0	59	35	3	2	4770
0	43	45	9	3	3004
0	37	54	6	3	649
35	58	7	0	0	1752
35	54	11	0	0	1673
15	74	10	1	0	5191
10	80	9	1	0	17241
15	74	10	1	0	14469
5	81	14	0	0	1521
3	51	36	6	4	1608
2	57	34	6	2	1452
2	43	44	9	3	5102
0	53	41	4	1	18766
0	36	49	11	3	13981
0	33	58	7	1	1728

(vervolg)		%			totaal
2	43	33	12	11	870
2	47	36	8	7	1326
1	37	39	15	8	3969
0	29	48	14	8	12476
1	28	42	19	11	10426
0	19	44	26	12	1400
1	26	54	11	8	957
0	40	45	8	6	1791
1	29	45	17	9	4427
0	29	51	15	5	18580
0	26	47	18	10	13486
0	17	59	20	4	2091
0	77	21	1	1	81
0	38	62	0	0	105
4	63	27	6	1	599
1	57	37	5	0	3224
3	54	33	10	0	1964
0	60	37	3	0	424
0	7	39	25	28	127
0	8	36	33	23	277
0	16	34	34	15	888
0	8	57	24	11	6004
0	9	42	32	17	4195
0	6	46	37	12	752
0	34	17	32	17	47
0	14	33	19	34	187
0	6	28	19	46	354
0	4	29	26	40	2201
0	5	21	28	47	1642
0	1	12	32	55	316
0	23	40	17	19	77
0	2	29	41	28	279
0	4	26	31	38	791
0	3	34	30	33	4908
0	3	24	33	40	3323
0	1	25	35	38	612

L1, L2, L4 en L3 en de volgorde van B werd B2, B4, B6, B5, B3 en B1. Achteraf bleek de volgorde van de inkomensklassen juist te zijn. Ook de volgorde van de scholingsklassen kwam ongeschonden tevoorschijn. Bij leeftijd bleken de klassen 3 en 4 te zijn verwisseld, hetgeen aangeeft dat de inkomenspositie van de oudste groep weer iets verslechtert. De variabele bedrijfstak bleek achteraf de nominale variabele te zijn. De gevonden volgorde heeft natuurlijk toch wel betekenis.
Op grond van de gevonden volgordes kan tabel 3.2 worden herschreven. Het resultaat is gegeven als tabel 3.3. Deze tabel is veel overzichtelijker dan de eerste. Er blijkt een duidelijk patroon en afwijkingen hiervan kunnen eenvoudig worden opgespoord. Vooral de variabelen S en L blijken een groot deel van de verschillen in de inkomensverdelingen te verklaren. Toch mag de invloed van de bedrijfstak niet worden genegeerd. Ook treden er duidelijke interacties op. Tabel 3.3 komt praktisch overeen met de bijlage die op basis van het onthulde bestand is gecreëerd.

3.3. Analyse van het onthulde bestand

Zoals in 3.1 is gesteld, zijn niet alle mogelijkheden om te percenteren even interessant. Bij een geheim bestand is het moeilijk een interessante keuze te maken. Immers, het resultaat kan niet in de werkelijkheid vertaald worden. Alleen met het gegeven dat I de te verklaren variabele was kon iets gedaan worden.

Nadat de betekenis van de variabelen was onthuld werden de resultaten van de analyse, die in paragraaf 3.2 is beschreven, duidelijk. Naast datgene wat iedereen wel weet, namelijk dat het inkomen met leeftijd en scholing samenhangt, bleek het feit dat het inkomen bij eenzelfde klasse van scholing x leeftijd nog sterk varieert over de verschillende bedrijfstakken.
Alvorens een poging te doen om de verschillen in inkomensverdelingen van de verschillende bedrijfstakken weer te geven, lijkt het zinvol om de structuur qua leeftijd en scholing te beschrijven. Daarvoor beschouwen we de (3x4)x6 tabel (SxL)xB, welke verticaal is gepercenteerd (tabel 3.4). Hieraan is nog een zevende kolom toegevoegd die de situatie over alle bedrijfstakken gesommeerd weergeeft.

Vervolgens is voor iedere cel uit tabel 3.4 het gemiddelde inkomen bepaald. (Voor het gemak op basis van de aanname dat het gemiddelde inkomen in de k^{de} klasse van I gelijk is aan k x f 10 000.) Het resultaat van deze berekeningen is te lezen in tabel 3.5. Helaas bevatten de tabellen 3.4 en 3.5 minder informatie dan de oorspronkelijk gegeven tabel: de inkomensverdelingen per cel zijn vervangen door gemiddelden. Beide tabellen vullen elkaar aan.

Tabel 3.4 Procentuele verdelingen naar scholing x leeftijd (SxL) per bedrijfstak en voor de gehele populatie

S	L	Bedrijfstak (B) 1	2	3	4	5	6	Gehele populatie
					%			
1	1	3	6	4	5	3	6	4
	2	3	7	3	5	3	5	4
	3	6	8	4	6	5	7	5
	4	8	7	5	10	6	8	6
2	1	14	23	19	18	17	18	18
	2	16	21	19	15	19	18	18
	3	19	13	18	19	19	16	18
	4	13	12	14	14	12	14	13
3	1	4	1	3	1	3	2	3
	2	7	2	6	3	6	3	5
	3	6	1	4	3	5	3	4
	4	3	1	2	2	2	1	2
Totaal		100	100	100	100	100	100	100

Tabel 3.5 Gemiddelde inkomen per scholing x leeftijdsklasse (SxL) per bedrijfstak en voor de gehele populatie

S	L	Bedrijfstak (B) 1	2	3	4	5	6	Gehele populatie
1	1	19217	16463	17952	14896	18767	17995	18034
	2	24724	22591	25755	20607	23672	24009	24090
	3	27504	22855	27127	23074	24688	25859	25499
	4	27744	22515	26715	24762	25743	25956	25985
2	1	21019	17277	19802	17645	19984	19699	19716
	2	27737	25678	28065	24807	25333	26735	26489
	3	31133	29875	31030	27929	29539	30452	30142
	4	33079	28805	31086	27074	30198	29191	30327
3	1	24198	22716	25046	26190	24705	23489	24662
	2	35386	37480	35578	37256	33721	34842	34660
	3	40997	33247	40870	39391	39336	40329	39981
	4	44057	33191	41681	37273	40236	40593	40832

De op grond van de laatste kolom van tabel 3.5 vast te stellen ordening van de leeftijdsklassen stemt, in tegenstelling tot de in 3.2 gevonden ordening, overeen met de natuurlijke ordening. Dit geldt min of meer ook voor de afzonderlijke bedrijfsklassen. Slechts in zeven van de achttien gevallen wint leeftijdsklasse 3 het van 4.

Op basis van deze tabellen zijn tenslotte nog drie soorten gemiddelden per bedrijfstak berekend, namelijk het rekenkundige, het direct gestandaardiseerde en het indirect gestandaardiseerde gemiddelde. De gestandaardiseerde gemiddelden zijn berekend om de directe invloed van de specifieke structuur per bedrijfstak van scholing en leeftijd op het gemiddelde uit te schakelen. Als standaardpopulatie is gekozen de totale populatie.

Deze gemiddelden worden als volgt berekend:
Zij q_{ij} het percentage van bedrijfstak j, dat betrekking heeft op de i^{de} combinatie van scholing maal leeftijd en zij q_{i+} het overeenkomstige percentage van de gehele populatie.
Zij x_{ij} en x_{i+} de bijbehorende gemiddelde inkomens.
q_{ij}, q_{i+}, x_{ij} en x_{i+} zijn af te lezen uit de tabellen 3.4 en 3.5.
Het rekenkundige gemiddelde van bedrijfstak j is eenvoudig

$$\bar{x}_j = \Sigma q_{ij} x_{ij}/100 \qquad (3.1)$$

Het rekenkundige gemiddelde voor de gehele populatie is

$$\bar{x}_+ = \Sigma q_{i+} x_{i+}/100 \qquad (3.2)$$

Het direct gestandaardiseerde gemiddelde voor bedrijfstak j is

$$\bar{x}_j^{d.s} = \Sigma q_{i+} x_{ij}/100 \qquad (3.3)$$

Het indirect gestandaardiseerde gemiddelde voor bedrijfstak j is

$$\bar{x}_j^{i.s} = ((\Sigma q_{ij} x_{ij})/(\Sigma q_{ij} x_{i+}))\bar{x}_+ \qquad (3.4)$$

De formules spreken voor zich. De twee methodes van standaardiseren leveren doorgaans praktisch gelijke resultaten. Zie ook Fleiss (1973).
De uitkomsten van de berekeningen staan in tabel 3.6.

Tabel 3.6 Gemiddelde inkomens per bedrijfstak

Soort gemiddelde	Bedrijfstak (B) 1	2	3	4	5	6	Totale populatie
rekenkundig	29 411	23 709	28 080	24 747	27 046	26 352	27 206
direct gestandaardiseerd	28 616	25 565	28 064	25 350	26 786	27 153	27 206
indirect gestandaardiseerd	28 574	25 423	28 043	25 109	26 766	27 137	27 206

Na standaardisatie blijken de grote verschillen belangrijk kleiner te worden. Verder treedt er een tweetal wijzigingen op in de volgorde der bedrijfstakken, als deze worden geordend volgens de gemiddelde inkomens.

4. RIDIT-ANALYSE (relative to an identified distribution)

4.1. Techniek

Ridit-analyse is een techniek bedoeld voor ordinale variabelen. Deze techniek biedt de mogelijkheid om na te gaan in hoeverre frequentieverdelingen van een ordinale variabele bij verschillende populaties ten opzichte van elkaar verschoven liggen. We kunnen de techniek dan ook beschouwen als het ordinale analogon voor het vergelijken van gemiddelden bij variabelen op nogere meetniveaus. De centrale grootheid bij ridit-analyse is de zgn. 'gemiddelde ridit', die twee frequentieverdelingen met elkaar vergelijkt. De ene frequentieverdeling hoort bij de zgn. referentiegroep, de andere bij de onderzoekgroep. Gemakshalve nemen we aan dat de ordinale variabele, in het vervolg ridit-variabele genoemd, conceptueel een continue variabele is, die bijvoorbeeld bij gebrek aan geschiktere meetapparatuur slechts als een discrete variabele op ordinaal niveau is waargenomen. Literatuur over ridit-analyse vindt men in Bross (1958), Jansen (1980a) en Jansen (1980b).

De gemiddelde ridit is - zoals de naam al zegt - een gemiddelde van ridits. Voor iedere klasse van de ridit-variabele wordt met behulp van de referentiegroep een ridit berekend die als schaalwaarde te beschouwen is voor de ridit-variabele. In het onderstaande zal worden aangetoond dat de gemiddelde ridit kan worden geïnterpreteerd als de kans dat een aselekt gekozen element uit de referentiegroep lager op de ridit-variabele scoort dan een aselekt gekozen element uit de onderzoekgroep. Hiertoe voeren we enige notaties in:

X	referentiegroep
$N_i(X)$	aantal elementen uit groep X behorend tot klasse i van de ridit-variabele
$N(X)$	totaal aantal elementen in groep X
Y	onderzoekgroep
$N_i(Y)$	aantal elementen uit groep Y behorend tot klasse i van de ridit-variabele
$N(Y)$	totaal aantal elementen van groep Y
$R_i(X)$	ridit van klasse i van de ridit-variabele, berekend m.b.v. de frequentieverdeling van groep X.

$R_i(X)$ is gedefinieerd als de kans dat een aselekt gekozen element uit de referentiegroep X lager scoort op de ridit-variabele dan een aselect gekozen element uit groep X behorend tot klasse i van de ridit-variabele. In formule:

$$R_i(X) = P\{\underline{x}_1 \leq \underline{x}_2 | \underline{x}_2 \text{ in klasse } i\}. \qquad (4.1)$$

Rekening houdend met het continue karakter van de variabele kunnen we $R_i(X)$ als volgt berekenen

$$R_i(X) = [0,5\ N_i(X) + \sum_{j=1}^{i-1} N_j(X)]/N(X) \quad . \tag{4.2}$$

Aan iedere klasse van de riditvariabele is aldus een schaalwaarde $R_i(X)$ toegekend, die bij benadering overeenkomt met de cumulatieve verdelingsfunctie. De gemiddelde ridit van de onderzoekgroep Y, gegeven de frequentieverdeling van de referentiegroep X is gedefinieerd als:

$$R(Y|X) = \sum_i R_i(X) \frac{N_i(Y)}{N(Y)} \tag{4.3}$$

Veronderstellen we dat

$$P\ \{\underline{x} \leq \underline{y} | \underline{x} \text{ en } \underline{y} \text{ in klasse i}\} = 0,5 \tag{4.4}$$

dan geldt:

$$R(Y|X) = P\ \{\underline{x} \leq \underline{y}\} \tag{4.5}$$

Indien veronderstelling (4.4) niet juist is en geen rekening wordt gehouden met het continue karakter van de variabele, dan geldt:

$$R(Y|X) = P(\underline{x} < \underline{y}) + 0,5\ P(\underline{x} = \underline{y}) \tag{4.6}$$

$R(Y|X)$ is een gemiddelde (over de frequentieverdeling van $\underline{y}$) van ridits. Wanneer we op intervalniveau gemiddelden van twee groepen vergelijken is het neutraal element (geen verschil tussen beide groepen) gelijk aan nul. In onze situatie volgt uit (4.5) dat het neutraal element $R(X|X)$ gelijk is aan 0,5.

Met een ridit-analyse kan men de samenhang tussen de ridit-variabele en andere, onafhankelijke, variabelen onderzoeken. Stel dat aan de elementen van de onderzoekgroep behalve de ridit-variabele ook één of meer onafhankelijke variabelen zijn gemeten. Men kan de onderzoekgroep dan splitsen in subgroepen die corresponderen met de klassen van de onafhankelijke variabelen. De afwijking van de gemiddelde ridits van de subgroepen t.o.v. de gemiddelde ridit van de totale onderzoekgroep geeft informatie over de

invloed van de onafhankelijke variabelen op de ridit-variabele. In Jansen (1980a) vindt men hoe betrouwbaarheidsintervallen voor gemiddelde ridits kunnen worden bepaald. Deze problemen blijven hier verder buiten beschouwing.

4.2. Analyse

Een ridit-analyse zoals hier toegepast bestaat uit het vergelijken van gemiddelde ridits van subgroepen van de onderzoekgroep van 231 833 individuen. Deze subgroepen zijn bepaald door de variabelen Scholing (S), Leeftijd (L) en Bedrijfstak (B). De ridit-variabele is hier het Inkomen (I). Als referentiegroep is de totale onderzoekgroep gebruikt; de gemiddelde ridit voor deze groep is dus 0,5. De ridits $R_i(X)$ voor de 5 inkomensklassen zijn: $R_1(X) = 0{,}0199$, $R_2(X) = 0{,}2590$, $R_3(X) = 0{,}6525$, $R_4(X) = 0{,}8807$ en $R_5(X) = 0{,}967$. We kunnen nu m.b.v. (4.3) voor iedere subgroep de gemiddelde ridit bepalen. De resultaten hiervan vindt men in fig. 4.1. Hierin is eerst de totale groep uitgesplitst naar bedrijfstak. De resulterende 6 groepen zijn uitgesplitst naar leeftijd. De resulterende 24 groepen zijn vervolgens weer uitgesplitst naar scholing. Figuur 4.1 is dus toegesneden op de volgende vraagstelling: hoe varieert het inkomen per bedrijfstak? Hoe varieert binnen de bedrijfstakken het inkomen per leeftijdsgroep? Hoe varieert binnen de leeftijdsgroepen per bedrijfstak het inkomen met de scholing? Afgezien van deze vragen kan men aan fig. 4.1 direct zien dat het inkomen gemiddeld hoger is in de groepen met hogere scholing en hogere leeftijd, met uitzondering van de hoogste leeftijdsgroep, waar weer een lichte daling optreedt. (Met gemiddeld inkomen wordt telkens de gemiddelde ridit van het inkomen bedoeld.) Wanneer we de gemiddelde ridits per bedrijfstak (stippellijnen) vergelijken blijkt dat in de bedrijfstak textiel (B1) het gemiddeld inkomen aanzienlijk hoger is dan in de totale groep (0,57 vs. 0,50). De bedrijfstak kleding/leer (B2) vertoont een duidelijk effect in tegenovergestelde richting.

Vervolgens kan worden gekeken naar de invloed van leeftijd op het inkomen binnen de verschillende bedrijfstakken (de gestreepte lijnen van fig. 4.1). Uniform geldt dat de groep 40-49 jr. het hoogste inkomen heeft, gevolgd door 50-64 jr. Weer wat lager staat de groep 30-39 jr., terwijl de groep 20-29 jr. relatief weinig verdient.

Het gecombineerde effect van S, L en B wordt weergegeven door de trapfuncties in fig. 4.1. Dit zijn de gemiddelde inkomen-ridits van de scholingsgroepen binnen een gegeven klasse van L x B. Binnen deze klassen stijgt het inkomen met de scholing. De mate van stijging is echter niet binnen alle klassen

even groot. Zo is het inkomen van de groep S3 binnen leeftijdsklasse 30-39 jr. en bedrijfstak kleding/leer zodanig groot dat hiermee de volgorde van de leeftijdsklassen m.b.t. de inkomens verstoord wordt. Een indruk van het effect van opleiding kan men krijgen door naar tabel 4.1 te kijken, waarin de verschillen in gemiddelde ridit voor de bedrijfstakken en leeftijdsklassen worden gegeven.

Tabel 4.1. Verschil van de gemiddelde ridits van S3 en S1 binnen een gegeven leeftijdsklasse en gegeven bedrijfstak

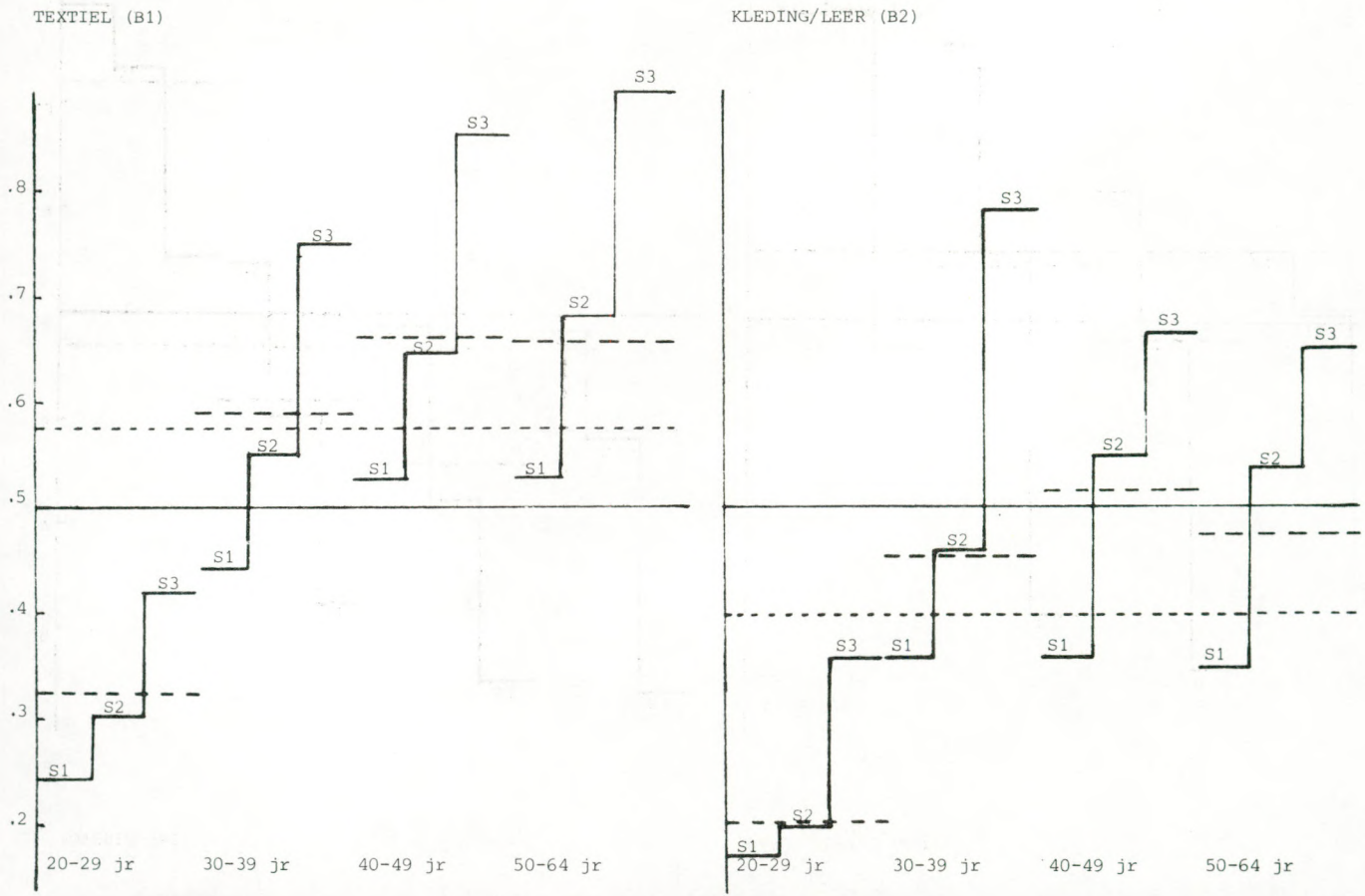
leeftijd (L) / bedr.tak (B)	20-29 jr.	30-39 jr.	40-49 jr.	50-64 jr.
1. textiel	0,17	0,30	0,32	0,37
2. kleding/leer	0,18	0,42	0,30	0,30
3. overig	0,23	0,32	0,33	0,33
4. basismetaal	0,36	0,53	0,44	0,32
5. metaal e.a.	0,20	0,31	0,39	0,37
6. voeding/genot	0,17	0,32	0,35	0,36

We zien dat een verschil in scholing vooral bij de ouderen wordt weerspiegeld in het inkomen. Voor de groep 30-39 jr. is scholing vooral belangrijk in de bedrijfstakken kleding/leer en basismetaal.

Tenslotte bekijken we nog een schijnbare tegenstelling in figuur 4.1. Het blijkt dat in de bedrijfstakken textiel en metaal e.a. de groep 50-64 jr. voor elk scholingsniveau meer verdient dan de groep 40-49 jr. Toch is het zo, dat in beide gevallen de groep 50-64 jr. in het totaal minder verdient dan de groep 40-49 jr. Deze situatie ontstaat omdat binnen de oudste leeftijdsgroep zich relatief meer laaggeschoolden bevinden. Deze informatie is in fig. 4.1. nog impliciet aanwezig omdat, wanneer we de groepen wegen met hun absolute aantallen, het zwaartepunt van S1, S2 en S3 juist op de corresponderende gestreepte lijn valt.

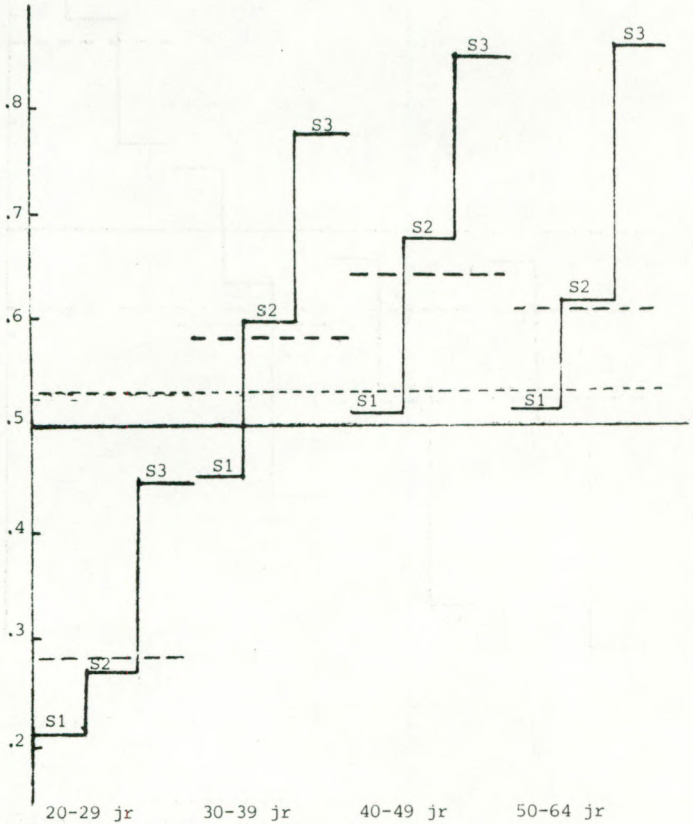
Uit de analyses blijkt dat het gezamenlijke effect van de variabelen S, L en B op de variabele I niet zonder meer gelijk is aan het totaal van de effecten van de variabelen afzonderlijk. Bij de verschillende bedrijfstakken wordt het inkomen door de interacties van de variabelen opleiding en leeftijd in verschillende mate beïnvloed. Door het berekenen van gemiddelde ridits werd het duidelijk dat er interacties zijn tussen de variabelen S, L en B die het inkomen beïnvloeden.

Figuur 4.1. Gemiddelde ridits met betrekking tot inkomen

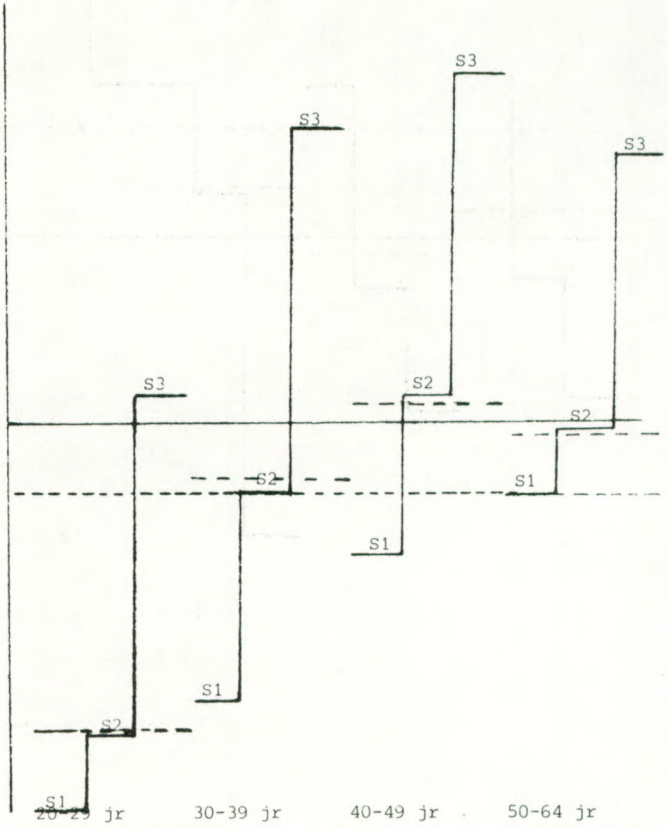


Figuur 4.1. vervolg

OVERIG (B3)

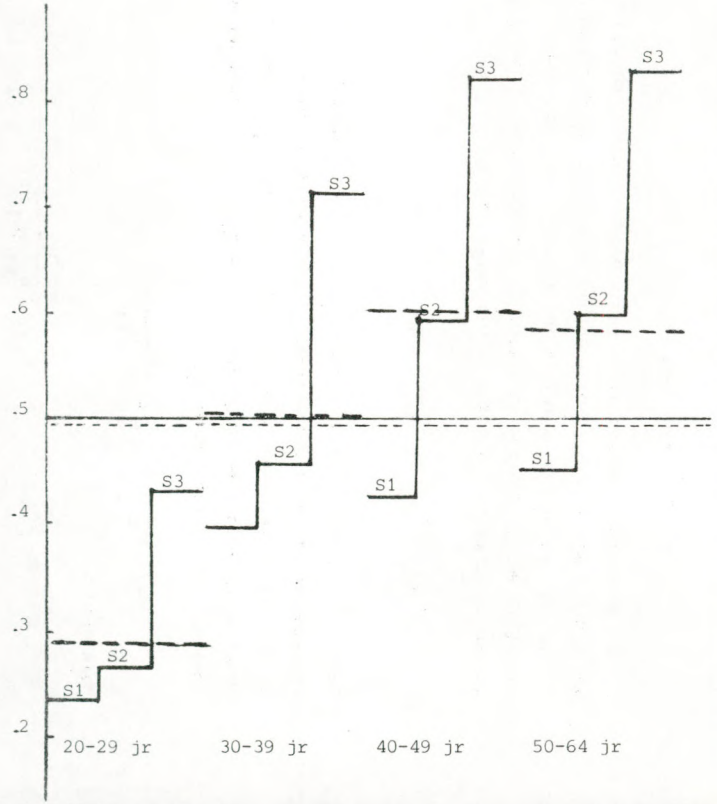


BASIS METAAL (B4)

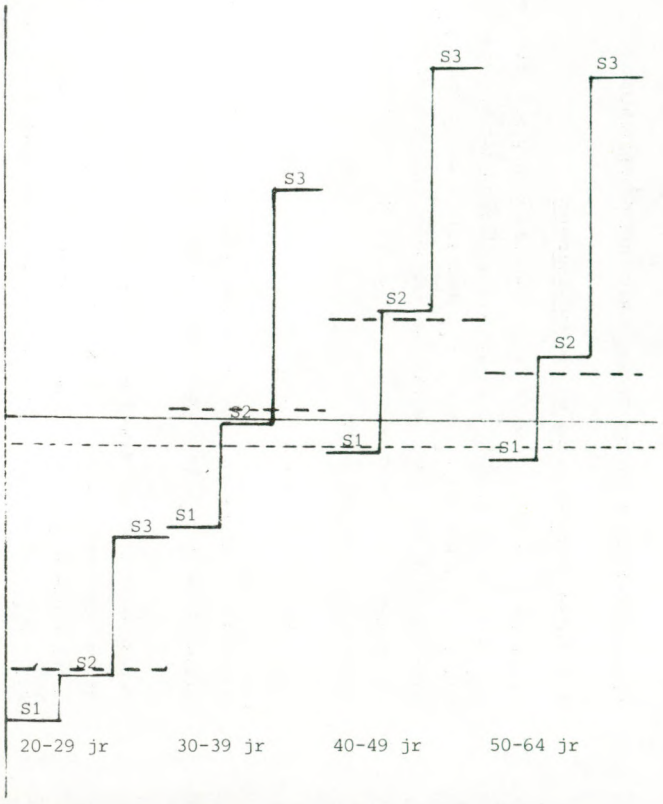


Figuur 4.1. vervolg

METAAL E.A. (B5)



VOEDING/GENOT (B6)



5. STAPSGEWIJZE SELECTIEMETHODEN M.B.V. HET G^2-CRITERIUM

5.1. Maten voor samenhang en voorspelbaarheid

Uitgangspunt zijn 2 discrete kansvariabelen $\underline{x}$ en $\underline{y}$ met simultane kansverdeling p_{ij} (i=1, ..., M en j=1, ..., N). Het is bekend dat bij onafhankelijkheid van $\underline{x}$ en $\underline{y}$ geldt dat $p_{ij} = p_{i+}\, p_{+j}$ voor alle i en j. Zodra $p_{ij} \neq p_{i+}\, p_{+j}$ voor zekere i en j, dan bestaat er samenhang tussen $\underline{x}$ en $\underline{y}$. Deze samenhang kan worden gemeten met de uit de informatietheorie afkomstige maatstaf (zie bijv. Kullback (1959))

$$J(\underline{x},\underline{y}) = \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{i+}p_{+j}}, \tag{5.1}$$

waarbij 'ln' de natuurlijke logaritme voorstelt. Merk op dat bij onafhankelijkheid van $\underline{x}$ en $\underline{y}$ geldt dat $J(\underline{x},\underline{y}) = 0$.

Uit de maatstaf voor samenhang ,J, kan ook een maat worden afgeleid voor de voorspelbaarheid van $\underline{y}$ uit $\underline{x}$; we zeggen dat $\underline{y}$ volledig voorspelbaar is uit $\underline{x}$ wanneer $\underline{y}$ bij elke gegeven x slechts één waarde kan aannemen, m.a.w. $p_{ij} = p_{i+}$ voor één j, en elders 0. In dit geval gaat (5.1) over in

$$H(\underline{y}) = \sum_{j=1}^{N} p_{+j} \ln \frac{1}{p_{+j}} \tag{5.2}$$

Dit is de maximale waarde die J kan aannemen. Totale onvoorspelbaarheid van $\underline{y}$ uit $\underline{x}$ wordt bereikt bij onafhankelijkheid, waarbij geldt dat $J(\underline{x},\underline{y}) = 0$. Kennis van $\underline{x}$ levert in dit geval geen informatie over de verdeling van $\underline{y}$. Als maatstaf voor de voorspelbaarheid van $\underline{y}$ uit $\underline{x}$ kan worden gebruikt (zie Israëls en Van Driel (1980))

$$\psi^2 = \frac{J(\underline{x},\underline{y})}{H(\underline{y})} \qquad (0 \leq \psi^2 \leq 1). \tag{5.3}$$

Over het algemeen zijn de kansen p_{ij} niet bekend maar beschikken we over een steekproef van n waarnemingen waarvan we de resultaten kunnen weergeven in een M x N frequentietabel met celfrequenties n_{ij}. Als schatters voor p_{ij} nemen we de relatieve frequenties $\hat{p}_{ij} = n_{ij}/n$. De (meest aannemelijke) schatter voor $J(\underline{x},\underline{y})$ wordt nu

$$\hat{J}(\underline{x},\underline{y}) = \frac{1}{n} \sum_i \sum_j n_{ij} \ln \frac{n_{ij}}{n_{i+}n_{+j}/n}. \tag{5.4}$$

Merk op dat

$$G^2 = 2n\, \hat{J}(\underline{x},\underline{y}) = 2 \sum_i \sum_j n_{ij} \ln \frac{n_{ij}}{n_{i+}\, n_{+j}/n}, \tag{5.5}$$

waarbij $\underline{G}^2$ de bekende likelihood ratio toetsingsgrootheid is bij het toetsen op onafhankelijkheid tussen $\underline{x}$ en $\underline{y}$, die bij onafhankelijkheid asymptotisch χ^2 is verdeeld met (M-1)(N-1) vrijheidsgraden. Evenzo wordt $H(\underline{y})$ geschat door

$$\hat{\underline{H}}(\underline{y}) = \frac{1}{n} \sum_j \underline{n}_{+j} \ln \frac{n}{\underline{n}_{+j}} \quad . \tag{5.6}$$

ψ^2 kan nu worden geschat door

$$\hat{\underline{\psi}}^2 = \frac{\underline{G}^2}{2n\, \hat{\underline{H}}(\underline{y})} \quad . \tag{5.7}$$

5.2. Decompositie van G^2

Een M x N frequentietabel kan worden ontbonden in een aantal M_k x N deeltabellen met $\sum_{k=1}^{K} M_k = M$. De M rijen worden aldus in K groepen ingedeeld.

We kunnen nu G^2 berekenen:

1. In de M x N tabel: Dit geeft $\underline{G}^2_{\text{Totaal}}$.
2. Tussen de K groepen, ofwel in de K x N tabel waarin de k^e rij bestaat uit de kolomtotalen van de k^e deeltabel: Dit geeft $\underline{G}^2_{\text{Tussen}}$.
3. Binnen groep k, d.w.z. in de k^e deeltabel: Dit geeft $\underline{G}^2_{\text{Binnen k}}$. Wanneer we dit binnen iedere groep k toepassen vinden we: $\underline{G}^2_{\text{Binnen}} = \sum_k \underline{G}^2_{\text{Binnen k}}$.

Tussen deze 3 grootheden bestaat de volgende optelregel:

$$\underline{G}^2_{\text{Totaal}} = \underline{G}^2_{\text{Tussen}} + \underline{G}^2_{\text{Binnen}} = \underline{G}^2_{\text{Tussen}} + \sum_k \underline{G}^2_{\text{Binnen k}} . \tag{5.8}$$

Onder de nulhypothese van onafhankelijkheid tussen rijen en kolommen zijn de 3 grootheden asymtotisch χ^2 verdeeld met resp. (M-1)(N-1), (K-1)(N-1) en $\sum_k (M_k-1)(N-1) = (M-K)(N-1)$ vrijheidsgraden (degrees of freedom, in de tabellen aangegeven met: df).

Analoog aan de splitsing in (5.8) verkrijgen we, na deling van de termen door $2n\, \hat{\underline{H}}(\underline{y})$:

$$\hat{\underline{\psi}}^2_{\text{Totaal}} = \hat{\underline{\psi}}^2_{\text{Tussen}} + \hat{\underline{\psi}}^2_{\text{Binnen}} . \tag{5.9}$$

We concluderen hieruit dat ook de verbetering qua voorspelbaarheid van de kolomvariabele ten gevolge van kennis van de rij op eenzelfde manier kan worden gesplitst.

5.3. Stapsgewijze methoden voor het analyseren van afhankelijkheid in tabellen

We gaan nu verder uit van de (3 x 4 x 6) x 5-tabel met verklarende variabelen S, L en B en afhankelijke variabele I. We willen de 72 rijen van deze 72 x 5-tabel zodanig tot K groepen samenvoegen dat de voorspelbaarheid van I gegeven de K groepen maximaal is, bij vaste K. We zoeken dus naar een K x 5-tabel met maximale $\hat{\psi}^2$. Dit komt neer op maximalisatie van $\hat{\psi}^2_{Tussen}$ uit (5.9). Omdat bij iedere te vormen K x 5-tabel zowel n als $\hat{H}(\underline{y})$ dezelfde zijn is dit identiek aan maximalisatie van G^2_{Tussen} uit (5.8), ofwel minimalisatie van G^2_{Binnen}. We zouden nu G^2 kunnen berekenen voor alle K x 5-tabellen (K vast) om de maximale G^2_{Tussen} of $\hat{\psi}^2_{Tussen}$ te bepalen. Wij gaan hier echter uit van voorwaartse stapsgewijze methoden, waarbij per stap een maximalisatie van G^2_{Tussen} plaatsvindt. Dit leidt tot eventueel suboptimale oplossingen voor de K x 5-tabel ($K \geq 3$). We behandelen hier 2 methoden:

1. selectie van gehele variabelen (S, L of B)
2. selectie van binaire splitsingen van variabelen.

Met behulp van deze methoden kunnen we de G^2_{Totaal} van de 72 x 5-tabel ontbinden in termen G^2 die op kleinere tabellen betrekking hebben. Voor de 72 x 5-tabel is G^2 gelijk aan 114 068 bij 71 x 4 = 284 vrijheidsgraden.

5.4. Selectie van gehele variabelen (methode 1)

Stap 1. We bekijken welke variabele het inkomen zo goed mogelijk voorspelt. Hiertoe bepalen we G^2 voor de tabellen S x I, L x I en B x I.

Tabel 5.1. Selectie van 1 verklarende variabele

Variabele	Afmeting v.d.tabel	G^2	$\hat{\psi}^2$	df
S	3 x 5	29 729	.050	8
L	4 x 5	65 498	.111	12
B	6 x 5	5 595	.009	20

Leeftijd blijkt de beste voorspeller en wordt in de 1e stap geselecteerd. Van de totale G^2 van 114 068 'verklaart' leeftijd 65 498.

Stap 2. We bekijken welke variabele het meest bijdraagt aan de waarde van G^2, gegeven dat L reeds is geselecteerd. Per leeftijdsklasse bekijken we hiervoor een 3 x 5 en een 6 x 5-subtabel voor resp. S en B.

Tabel 5.2. Selectie van een tweede verklarende variabele (gegeven L)

Toegevoegde variabele	Aantal tabellen	Afmeting per tabel	G^2	$\hat{\psi}^2$	df
S	4	3 x 5	40 401	.068	32
B	4	6 x 5	7 056	.012	80

Het blijkt dat scholing de variabele is die binnen de leeftijdsklassen het meeste invloed heeft op het inkomen. De 12 x 5-tabel (L x S) x I heeft een G^2 van 65 498 + 40 401 = 105 899 bij 12 + 32 = 44 vrijheidsgraden. Het weglaten van B uit de tabel betekent een verlies aan G^2 van 114 068 - 105 899 = 8 169 bij 284 - 44 = 240 vrijheidsgraden. Bedrijfstak blijkt dus een relatief kleine, maar toch nog zeer significante invloed te hebben.

In hoofdstuk 6 wordt een alternatief gegeven voor de in deze paragraaf geschetste methode, nl. logit-analyse. Daarbij is niet het doel de tabel tot een kleinere tabel te comprimeren, maar om effecten op te sporen die het inkomen beïnvloeden; waar hier telkens gehele variabelen worden toegevoegd, daar wordt in hoofdstuk 6 de mogelijkheid gegeven slechts marginale effecten van een variabele of bepaalde interacties toe te voegen.

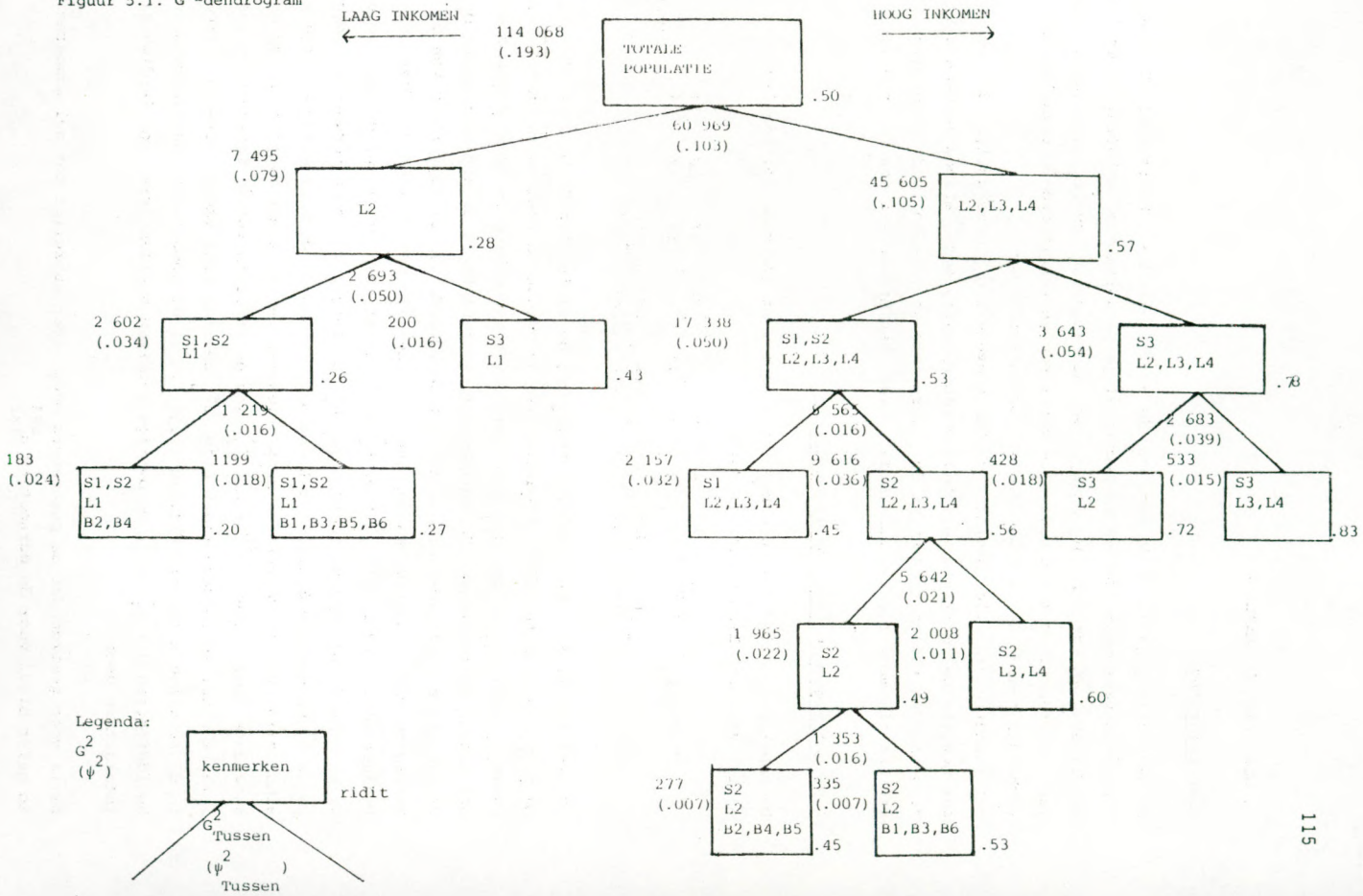
5.5. Selectie van binaire splitsingen van variabelen (methode 2)

De optimale binaire splitsing op grond van één van de verklarende variabelen is de splitsing in de groepen L1 en (L2, L3, L4). Deze splitsing leidt tot een 2 x 5-tabel met een G^2 van 60 969 hetgeen zeer hoog is vergeleken met de waarde van 65 498 voor G^2 bij de 4 x 5-tabel L x I en t.o.v. de G^2 van 114 068 van de oorspronkelijke 72 x 5-tabel. Beide groepen kunnen verder worden gesplitst. De grootste G^2 vinden we bij het splitsen van (L2, L3, L4) in (S1, S2; L2, L3, L4) en (S3; L2, L3, L4). Deze splitsing levert een G^2 op van 24 624. Beide splitsingen tezamen resulteren in een 3 x 5-tabel met een G^2 van 60 969 + 24 624 = 85 593. Aldus vervolgend kan men een dendrogram tekenen, zie fig. 5.1. In deze figuur staan binnen iedere rechthoek de kenmerken van de groep, links erboven de G^2-waarde in deze groep met tussen haakjes de bijbehorende ψ^2. Onder de rechthoeken staan de waarden van G^2_{Tussen} die bij elke splitsing worden verklaard (ψ^2_{Tussen} tussen haakjes). De figuur is afgekapt waar een verdere splitsing een bijdrage aan G^2 zou leveren van minder dan 700. De 8 getoonde splitsingen leveren tezamen een 9 x 5-tabel op met een G^2 van 106 748. De variabele bedrijfstak is slechts verantwoordelijk voor de laatste 2 van deze 8 splitsingen, terwijl deze variabele maar liefst 6 klassen heeft. Deze relatief geringe invloed namen we ook bij methode 1 waar.

Bij strikt nominale afhankelijke variabelen met meer dan 2 klassen is het dendrogram niet zonder meer eenduidig bepaald omdat er geen criterium is om te kiezen of een tak bij een splitsing links of rechts moet worden gezet. In ons geval, waar we de ordinale variabele inkomen als afhankelijke variabele hebben, kan dit wel door de afhankelijke variabele numeriek te maken via het toekennen van schaalwaarden aan iedere inkomensklasse, en daarmee aan de individuen uit die klassen. Wij hebben als schaalwaarden voor de klassen I1 t/m I5 resp. 0,0199, 0,2590, 0,6525, 0,8807 en 0,9673 genomen. Deze schaalwaarden zijn de in hoofdstuk 4 gedefinieerde ridits, dat zijn getallen die aangeven ten opzichte van hoeveel mensen men een hoog inkomen bezit (in fracties uitgedrukt). Bij iedere splitsing in fig. 5.1 hebben we voor beide ontstane groepen een gemiddelde inkomensscore (gemiddelde ridit) berekend en hebben telkens de groep met de hoogste score rechts geplaatst. Deze scores zijn rechts onder de rechthoeken geplaatst. Het blijkt dat bij iedere splitsing 2 groepen ontstaan die sterk in inkomen verschillen, ondanks het feit dat bij het splitsen geen rekening is gehouden met de ordening der klassen. De reden voor de keuze van ridits als schaalwaarden, in plaats van bijv. klassemiddens was de onbekendheid met de betekenis van variabele I bij het analyseren van het geheime bestand.

Figuur 5.1 bevestigt het vermoeden dat de intelligente lezers reeds hadden, nl. dat er een positieve samenhang bestaat tussen leeftijd en inkomen (waarbij L3 en L4 overigens nergens gesplitst worden) en tussen scholing en inkomen. Bedrijfstak speelt een relatief kleine rol.

Figuur 5.1. G^2-dendrogram



6. LOG-LINEAIRE ANALYSE

6.1. Inleiding

De data zijn gegeven in de vorm van de vierdimensionale kruistabel. De vier dimensies vertegenwoordigen respectievelijk de variabelen scholing, leeftijd, bedrijfstak en inkomen. Bij analyse met het algemene log-lineaire model wordt geen onderscheid gemaakt tussen te verklaren en verklarende variabelen. De verbanden tussen de vier variabelen worden beschouwd
met inbegrip van alle interacties. Omdat evenwel bij onze analyse de variabele inkomen als te verklaren variabele wordt opgevat, wordt alleen gekeken naar de verbanden tussen inkomen en de andere drie variabelen. Door deze inperking kan het algemene log-lineaire model worden geschreven als een logit-model.

6.2. Het algemene log-lineaire model

De tabel kan volledig worden beschreven door het (verzadigde) log-lineaire model (zie o.a. Bishop e.a. (1975)):

$$\begin{aligned} \text{Ln } m_{ijk\ell} = u + u_i^S + u_j^L + u_k^B + u_\ell^I + u_{ij}^{SL} + u_{ik}^{SB} + u_{i\ell}^{SI} + u_{jk}^{LB} + u_{j\ell}^{LI} \\ + u_{k\ell}^{BI} + u_{ijk}^{SLB} + u_{ij\ell}^{SLI} + u_{ik\ell}^{SBI} + u_{jk\ell}^{LBI} + u_{ijk\ell}^{SLBI} \end{aligned} \qquad (6.1)$$

(De indices i, j, k en ℓ behoren resp. bij de variabelen S, L, B en I).
Ln $m_{ijk\ell}$ is de natuurlijke logaritme van de verwachte frequentie in cel (i,j,k,ℓ), dus $m_{ijk\ell} = n\, p_{ijk\ell}$ met $p_{ijk\ell}$ de kans dat een trekking in deze cel komt, en n het aantal waarnemingen. De u-parameters representeren de effecten op ln $m_{ijk\ell}$. De parameter u is gedefinieerd als het gemiddelde van de logaritmen van de verwachte celfrequenties en representeert het algemene niveau van deze celfrequenties. Verder zijn u_i^S t.m. u_ℓ^I de hoofdeffecten van resp. de variabelen S, L, B en I. Zo representeert bijv. u_i^S de afwijking van de verwachte frequentie in de i^e klasse van S t.o.v. het algemene niveau u. De hoofdeffecten zijn slechts afhankelijk van de univariate marginalen van de tabel met logaritmen van de verwachte frequenties. u_{ij}^{SL} t.m. $u_{k\ell}^{BI}$ zijn de effecten van de interacties tussen 2 variabelen. Zo representeert bijv. u_{ij}^{SL} de afwijking van de verwachte frequentie in de scholing x leeftijdsklasse (i,j) t.o.v. de combinatie van algemene en hoofdeffecten $u + u_i^S + u_j^L$. De overige u-termen stellen drie- en vierdimensionale interacties voor.

Er is voor gezorgd dat de parameters over ieder subscript tot nul sommeren. Zo geldt bijv. voor de parameters $u_{ik\ell}^{SBI}$

$$\sum_i u^{SBI}_{ik\ell} = \sum_k u^{SBI}_{ik\ell} = \sum_\ell u^{SBI}_{ik\ell} = 0$$

Alle parameters kunnen worden geschat door maximum-likelihood-schatters. Substitueren we de parameters van (6.1) door hun schatters dan verkrijgen we

$$\text{Ln } \underset{\sim}{n}_{ijk\ell} = \hat{\underset{\sim}{u}} + \hat{\underset{\sim}{u}}^S_i + \ldots + \hat{\underset{\sim}{u}}^I_\ell + \hat{\underset{\sim}{u}}^{SL}_{ij} + \ldots + \hat{\underset{\sim}{u}}^{BI}_{k\ell} + \hat{\underset{\sim}{u}}^{SLB}_{ijk} + \ldots + \hat{\underset{\sim}{u}}^{LBI}_{jk\ell} + \hat{\underset{\sim}{u}}^{SLBI}_{ijk\ell} \qquad (6.2)$$

Hypothesen over het al dan niet nul zijn van effecten kunnen worden gesteld in de vorm van modellen waarin een aantal van de u-parameters gelijk aan nul worden gesteld. Deze modellen kunnen worden beschreven in termen van marginale frequenties. Als voorbeeld nemen we het model

$$\ln m_{ijk\ell} = u + u^S_i + u^L_j + u^B_k + u^I_\ell + u^{SL}_{ij} . \qquad (6.3)$$

(Analoog aan (6.1) kunnen we de verwachte frequenties en de parameters in dit model schatten.)

In dit model zijn behalve de interactie tussen variabelen S en L alle interacties gelijk aan nul gesteld. Anders gezegd: Volgens model (6.3) laat de vier-dimensionale tabel met verwachte frequenties zich beschrijven door een tabel, gevormd door de variabelen S en L, die de bivariate marginale verwachte frequenties bevat, en de vier univariate marginalen van variabelen S, L, B en I. De tabel met bivariate marginalen van variabelen S en L bevat ook de informatie over de univariate marginalen van variabelen S en L. De minimaal voldoende statistische grootheden voor het schatten van de celfrequenties onder model (6.3) zijn nu de gerealiseerde univariate marginalen van variabele B en van variabele I en de gerealiseerde bivariate marginalen van variabelen S en L. Het model (6.3) zullen we daarom in het vervolg noteren als SL B I.

Stel we beschikken over een model H^*. Zij model H° verkregen uit H^* door weglating van een set parameters. We kunnen nu toetsen of de verwachte frequenties onder beide modellen van elkaar verschillen (ofwel de weggelaten parameters nul zijn) m.b.v. de toetsingsgrootheid

$$\underset{\sim}{G}^2(H^\circ;H^*) = 2 \sum\sum\sum\sum \underset{\sim}{n}_{ijk\ell} (\ln \hat{\underset{\sim}{m}}^*_{ijk\ell} - \ln \hat{\underset{\sim}{m}}^\circ_{ijk\ell}), \qquad (6.4)$$

waarin $\hat{\underset{\sim}{m}}^*_{ijk\ell}$ de geschatte frequenties zijn onder model H^* en $\hat{\underset{\sim}{m}}^\circ_{ijk\ell}$ de geschatte frequentie onder model H°. Als H^* het verzadigde model (6.1) is, gaat formule (6.4) over in

$$\underset{\sim}{G}^2(H^\circ) = 2 \sum\sum\sum\sum \underset{\sim}{n}_{ijk\ell} (\ln \underset{\sim}{n}_{ijk\ell} - \ln \hat{\underset{\sim}{m}}^\circ_{ijk\ell}) . \qquad (6.5)$$

Als de onder H^o verwachte frequenties gelijk zijn aan de onder het verzadigde model verwachte frequenties, is $G^2(H^o)$ asymptotisch χ^2 verdeeld met het aantal vrijheidsgraden gelijk aan het aantal cellen van de tabel verminderd met het aantal geschatte parameters onder model H^o. $G^2(H^o;H^*)$ is ook asymptotisch χ^2 verdeeld met het aantal vrijheidsgraden gelijk aan het aantal parameters onder H^* verminderd met het aantal parameters onder H^o.

6.3. Het logitmodel

Als één van de variabelen als afhankelijke variabele wordt opgevat, zoals bij ons met inkomen het geval is, en we zijn alleen geïnteresseerd in de effecten die de andere variabelen op deze afhankelijke variabele hebben, kunnen we het algemene log-lineaire model schrijven als een logitmodel. Meestal wordt het logitmodel alleen beschreven voor één afhankelijke variabele met twee categorieën. Goodman (1971) laat echter zien dat dit model is te generaliseren naar meer categorieën voor de afhankelijke variabele. Voor het geheim bestand beschouwen de logits

$$\Phi_{ijk\ell} = (\ln m_{ijk\ell} - \ln m_{ijk1})/2 \quad \text{voor } \ell = 2,3,4,5 . \tag{6.6}$$

Dus voor ieder niveau (i, j, k) van de variabelen scholing, leeftijd en bedrijfstak zijn er vier logits ($\frac{1}{2} \ln (m_{ijk\ell}/m_{ijk1})$ voor $\ell = 2,\ldots,5$) voor de variabele inkomen.

Uit (6.1) en (6.6) volgt

$$\Phi_{ijk\ell} = v^I_\ell + v^{SI}_{i\ell} + v^{LI}_{j\ell} + v^{BI}_{k\ell} + v^{SLI}_{ij\ell} + v^{SBI}_{ik\ell} + v^{LBI}_{jk\ell} + v^{SLBI}_{ijk\ell} \quad \text{voor } \ell=2,..,5, \tag{6.7}$$

waarin de v-parameters corresponderen met de verschillen van overeenkomstige u-parameters, bv.:

$$v^{LI}_{j\ell} = (u^{LI}_{j\ell} - u^{LI}_{j1})/2 \text{ , voor } \ell=2,..,5.$$

Dus de v-parameters zijn functies van de u-parameters. Voor de v-parameters geldt ook dat zij sommeren tot nul over ieder subscript i, j en k.

Model (6.7) is het verzadigde logit-model. Als één of meer van de v-parameters gelijk aan nul worden verondersteld, ontstaat een onverzadigd logitmodel. De v-parameters van een onverzadigd logitmodel kunnen worden berekend uit de u-parameters van het corresponderende onverzadigde log-lineaire model. De G^2-waarde voor de aanpassing van het onverzadigde logitmodel aan de data is gelijk aan de G^2-waarde voor de aanpassing van het corresponderende log-lineaire model (zie ook Goodman (1971)).

6.4. Resultaten

Meestal toetst men bij een log-lineaire analyse verschillende modellen tegen het verzadigde model met de G^2-toets uit (6.5). Als een niet-verzadigd model een significante afwijking van het verzadigde model vertoont, zijn de weggelaten parameters significant. In onze toepassing geeft, mede vanwege de grote aantallen, iedere vereenvoudiging van het verzadigde model, zelfs het weglaten van de vier-term-interactie SLBI, een significante G^2-waarde. We gebruiken daarom een andere, niet op significantie-toetsen gebaseerde, methode om een model te selecteren. De methode is gebaseerd op een aan de R^2 van regressieanalyse verwante passingsmaat. Deze passingsmaat is gegeven door

$$R^2 = [G^2(H) - G^2(H^*)] \; / G^2(H) \qquad (6.8)$$

(zie o.a. Goodman (1971) en Zahn en Fein (1979)), en is te interpreteren als de proportie van de samenhang (G^2) in de tabel die het model H^* meer verklaart dan het model H. Deze interpretatie wordt eenvoudiger als we ons beperken tot logitmodellen. We nemen dan voor H het model SLB I. Dit model bevat geen enkele interactie tussen de afhankelijke variabele I en de onafhankelijke variabelen, maar wel de interacties tussen de variabelen S, L en B. De G^2-waarde van dit model geeft de totale variatie van de logits van variabele I aan. Voor H^* nemen we een model dat van model H verschilt doordat er parameters aan zijn toegevoegd. Omdat we logitmodellen beschouwen, bevatten alle toegevoegde parameters variabele I. R^2 is nu de proportie van de totale variatie van de logits van variabele I die wordt verklaard door deze toegevoegde parameters.

Met behulp van deze R^2 kunnen we een voorwaartse selectieprocedure toepassen. Uitgaande van het eenvoudigste logitmodel H worden steeds parameters toegevoegd tot we een voldoende hoeveelheid variatie hebben verklaard of totdat het toevoegen van nieuwe parameters niet veel extra meer verklaart.
De parameters worden toegevoegd in volgorde van belangrijkheid, de belangrijkste eerst.

Tabel 6.1 geeft een samenvatting van de resultaten van de voorwaartse selectieprocedure. Uit deze tabel blijkt het grote belang van de parameters van SI en LI t.o.v. de andere parameters. Dus scholing en vooral leeftijd zijn de variabelen die het inkomen bepalen.

Tabel 6.1. Variatie van de logits van inkomen

M o d e l				df	G^2	R^2
SLB	I			284	114 068	0
SLB	LI			272	48 570	0,57
SLB	LI	SI		264	9 041	0,92
SLB	LI	SI	BI	244	3 932	0,97
SLB	LBI	SI		184	2 660	0,98
SLB	LBI	SLI	SBI	120	1 132	0,99

Als we een model met een R^2 van 0,90 acceptabel vinden, dan is het model SLB LI SI het model met het minste aantal parameters dat een voldoende passing op de data geeft. De geschatte waarde van de parameter u onder dit model is 4,376. De schattingen van de parameters $u^{LI}_{j\ell}$ en $u^{SI}_{i\ell}$ geven het volgende resultaat:

Tabel 6.2. Parameterschattingen voor leeftijd maal inkomen.

$\hat{u}^{LI}_{j\ell}$	inkomen(I) $\ell = 1$	2	3	4	5
leeftijd(L)					
j = 1	3,355	1,343	-0,406	-1,535	-2,756
2	-0,660	0,013	0,282	0,227	0,135
3	-1,399	-0,669	0,164	0,690	1,215
4	-1,295	-0,686	-0,040	0,618	1,403

Tabel 6.3. Parameterschattingen voor scholing maal inkomen

$\hat{u}^{SI}_{i\ell}$	inkomen(I) $\ell = 1$	2	3	4	5
scholing(S)					
i = 1	1,762	0,811	-0,263	-1,075	-1.236
2	0,551	0,224	-0,025	-0,208	-0,542
3	-2,313	-1,035	0,287	1,283	1,777

6.5. Conclusies

Twee variabelen, de variabelen leeftijd en scholing, hebben een belangrijk effect op de inkomensverdeling. Het effect van leeftijd is het grootst. De variabele bedrijfstak heeft ook enige invloed, maar deze is veel kleiner dan die van leeftijd en inkomen. Het effect van interacties tussen meer dan twee verklarende variabelen is te verwaarlozen.

Uit het patroon van de parameters van de interactie tussen de variabelen leeftijd en inkomen blijkt dat in de laagste leeftijdcategorie (L1) het laagste inkomen (I1) relatief het sterkst vertegenwoordigd is, de klasse I2 het op één na sterkst vertegenwoordigd, en dit patroon zet zich voort tot I5, die het minst sterk is vertegenwoordigd. Bij leeftijdcategorie L2 zien we een vrij vlakke inkomensverdeling. Er is een kleine top bij I3 terwijl I1 is ondervertegenwoordigd. Bij L3 en L4 is de trend omgekeerd t.o.v. de trend bij L1; de parameters nemen monotoon toe met de inkomensklassen. De hoogste inkomens zijn dus het sterkst vertegenwoordigd en de laagste inkomens het minst sterk. Omdat de parameters in deze laatste twee groepen niet veel verschillen zouden deze groepen kunnen worden samengenomen zonder veel informatie te verliezen.

Bij het effect van scholing op inkomen zien we eenzelfde beeld. In de laagste scholingscategorie nemen de parameters monotoon af met het inkomen en in de hoogste scholingscategorie nemen de parameters monotoon toe met het inkomen. Ook in de middelste scholingscategorie nemen de parameters af met het inkomen, maar de inkomensverdeling is hier veel vlakker.

7. CLUSTERANALYSE

7.1. Algemeen

Clusteranalyse is een techniek om een verzameling elementen op te delen in een aantal zo homogeen mogelijke deelverzamelingen (clusters). Nadere bestudering van de zo ontstane clusters, met name de gemeenschappelijke kenmerken van de elementen in de clusters, kan aanknopingspunten bieden voor de onthulling van de in het materiaal aanwezige structuur.

Aan ieder element van de te onderzoeken verzameling zijn metingen verricht. Deze metingen worden gebruikt om na te gaan in hoeverre twee elementen op elkaar lijken. Daartoe moet een afstandsmaat gedefinieerd worden. Clusteranalyse schrijft geen afstandsmaat voor. Specificatie wordt overgelaten aan de gebruiker.

Er zijn veel verschillende typen clusteranalysetechnieken in omloop. Voor een eenvoudig overzicht kan verwezen worden naar Everitt (1974). De meest toegepaste technieken worden samengevat onder de verzamelnaam agglomeratieve niërarchische clusteranalyse. (Zie Everitt (1974) en Bethlehem (1976).) Het hiërarchische clusteranalyseproces begint met het definiëren van zoveel clusters als er elementen zijn. Ieder element zit in een aparte cluster. In iedere stap van het proces worden die twee clusters samengevoegd die het meest op elkaar lijken. Het proces stopt tenslotte als er nog maar één grote cluster over is waarin alle elementen zitten. Het is de taak van de gebruiker om uit dit hiërarchische proces op relevante wijze de juiste clustering te halen. Een nuttig hulpmiddel hierbij wordt gevormd door het zgn. dendrogram. Dit is een grafische weergave van het clusteringsproces waarbij horizontaal de elementen zijn afgezet en verticaal de afstand waarbij clusters zijn samengevoegd. (Zie fig. 7.1.)
Ook de afstand tussen de clusters moet door de gebruiker nader gespecificeerd worden. Deze definitie bepaalt het type hiërarchische methode. Bij de "single linkage"-techniek wordt de afstand tussen twee clusters bijvoorbeeld gedefinieerd als de afstand tussen de twee het meest op elkaar lijkende elementen uit de twee clusters. Bij de "complete linkage"-techniek wordt dit de afstand tussen de twee het minst op elkaar lijkende elementen.

Indien de door de gebruiker gedefinieerde afstand tussen de elementen alle eigenschappen van het wiskundige begrip afstand bezit ("metriek") kunnen de elementen vaak zonder veel verlies aan informatie in een twee-dimensionale

grafiek worden weergegeven. Deze methode staat in de literatuur bekend onder de naam principale coördinaten analyse (zie Everitt (1978)). Door bestudering van een dergelijk plaatje met het oog kan dikwijls op zinvolle wijze met de hand clusteranalyse worden gedaan.

7.2. Toepassing

In het onderhavige geval wordt de variabele I geanalyseerd in samenhang met de variabelen S, L en B. Daartoe wordt het bestand tot een 72 x 5-tabel (S x L x B) x I getransformeerd (zie bijlage). De rijen van deze tabel zijn de elementen die geclusterd moeten worden. Een cluster bevat dan alle elementen (combinaties van S, L en B) met ongeveer dezelfde inkomensverdeling.

Voor de afstand tussen twee elementen (rijen) is, gezien het nominaal gestelde karakter van de inkomensvariabele, gekozen voor de chi-kwadraat-afstand. Voor de afstand tussen rij i en rij k van de 72x5-tabel is deze gedefinieerd als

$$d(i,k) = \sqrt{\sum_{j=1}^{5} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{kj}}{n_{k+}} \right)^2 / n_{+j}}. \qquad (7.1)$$

Aangezien de chi-kwadraat-afstand een metriek is, kan principale coördinatenanalyse worden toegepast. Dit leidt tot de puntenwolk uit figuur 7.2., waarin iedere rij door een punt wordt voorgesteld. In deze grafiek is 90% van de informatie behouden gebleven ondanks de projectie op het twee-dimensionale vlak. Opvallend is de regelmatige banaan-vorm. Tevens blijkt duidelijk dat het nauwelijks de moeite waard is om in deze situatie clusteranalyse toe te passen. De punten liggen vrij homogeen verdeeld over de banaan. Er vallen geen duidelijke clusters aan te wijzen. Desalniettemin is toch een hiërarchische clusteranalysemethode op de rijen van de tabel toegepast. Deze methode, die is geënt op de chi-kwadraat-afstand (7.1) is een variant (zie Benzécri (1976) en Israëls (1979)) op de methode van Ward. Iedere cluster wordt voorgesteld door een rij. Het hiërarchisch proces wordt begonnen met 72 clusters (rijen) en per stap worden twee clusters samengevoegd door optelling van de frequentieverdelingen van de bijbehorende rijen. Als afstand tussen twee clusters wordt genomen de afname in Pearson's χ^2-toetsingsgrootheid ten gevolge van het samenvoegen van de twee bijbehorende rijen.
Als stopcriterium in het clusteringsproces is genomen een onevenredig grote toename in de afstand tussen de twee clusters die samengevoegd worden.

Dit leidde tot acht clusters. (Zie het dendrogram in figuur 7.1). In figuur 7.2 zijn de clusters weergegeven in de puntenwolk. Opvallend is dat soms ver verwijderde elementen in dezelfde cluster zitten en dicht bij elkaar liggende elementen in verschillende clusters. Dit is voornamelijk te wijten aan het irreversibele karakter van het hiërarchische clusteringsproces. Wanneer twee elementen eenmaal bij elkaar in dezelfde cluster zitten zullen ze nooit meer gescheiden worden. Dit beïnvloedt het gedrag ten opzichte van andere elementen. Bovendien speelt de weging een rol. De ene rij van de tabel bevat meer individuen dan de andere rij. Daardoor kunnen gelijke afstanden tijdens het clusteringsproces verschillend geïnterpreteerd worden.

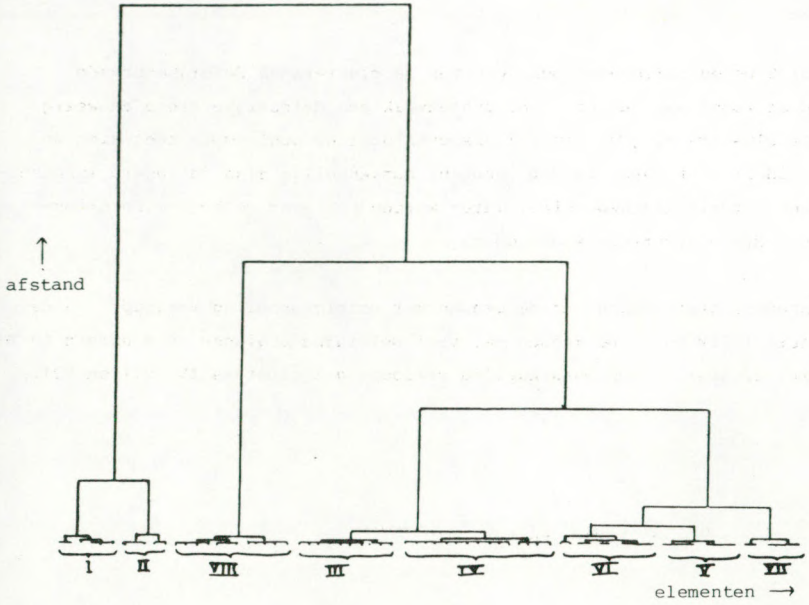
Een volgende stap is het interpreteren van de clusters. Daartoe is voor elke cluster bekeken wat de elementen zijn en hoe de inkomensverdeling van die cluster er uit ziet. Om de verschillen gemakkelijker te kunnen interpreteren zijn een soort residuen uitgerekend. Dit is de afwijking van de inkomensverdeling van de cluster t.o.v. de inkomensverdeling van het hele bestand. Figuur 7.3 bevat deze gegevens.
Gezien het feit dat de diverse klassen van de variabele B willekeurig over de clusters verdeeld lijken te zijn, kan worden geconcludeerd dat deze variabele geen bijdrage levert tot de verklaring van de structuur zoals hij gevonden is. Daarom is deze variabele niet betrokken in de verdere interpretatie.

Cluster I bevat 6 elementen. Dit zijn de jonge mensen met weinig scholing.
Cluster II bevat 5 elementen. Dit zijn ook jonge mensen, maar met iets meer scholing.
Cluster III bevat 10 elementen. Dit zijn iets oudere mensen met wat meer scholing en jonge mensen met veel scholing.
Cluster IV bevat 16 elementen. Dit zijn de oudere mensen met weinig scholing en jongere mensen met iets meer scholing.
Cluster V bevat 7 elementen. Dit zijn oude mensen met weinig scholing en mensen van middelbare leeftijd met wat meer scholing.
Cluster VI bevat 10 elementen. Dit zijn mensen van middelbare leeftijd met een middelbare opleiding.
Cluster VII bevat 6 elementen. Dit zijn de jongere mensen met een hoge opleiding.
Cluster VIII bevat 12 elementen. Dit zijn middelbare en oudere mensen met een hoge opleiding

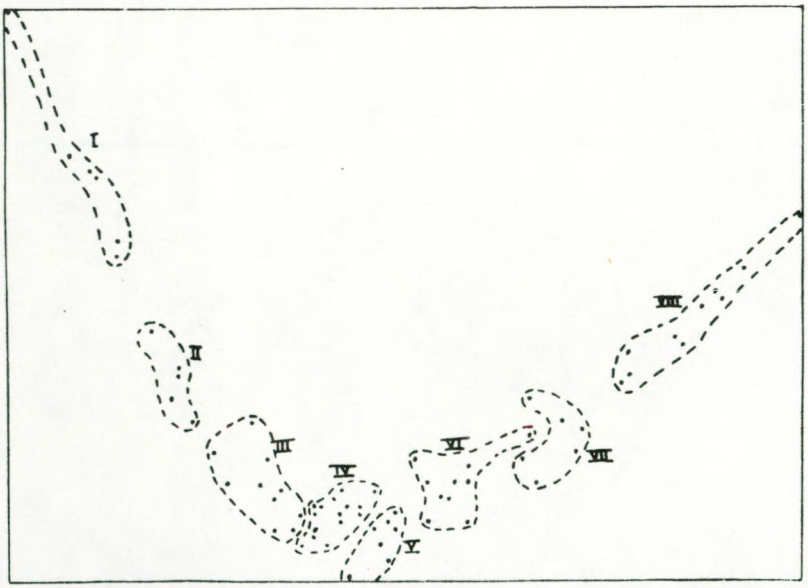
Een blik op de inkomensverdelingen van de clusters en de bijbehorende residuen toont aan dat er in de puntenwolk een duidelijke trend aanwezig is. De clusters volgend vanaf linksboven door de puntenwolk heen zien we een duidelijke toename in het inkomen. Aanvankelijk zijn de lagere inkomensklassen oververtegenwoordigd, later worden het meer de hogere inkomensklassen die oververtegenwoordigd zijn.

De inkomensontwikkeling van de mensen met weinig scholing verloopt via de clusters I, IV en V. De mensen met veel opleiding beginnen en eindigen in een hogere cluster. Hun ontwikkeling verloopt via clusters IV, VII en VIII.

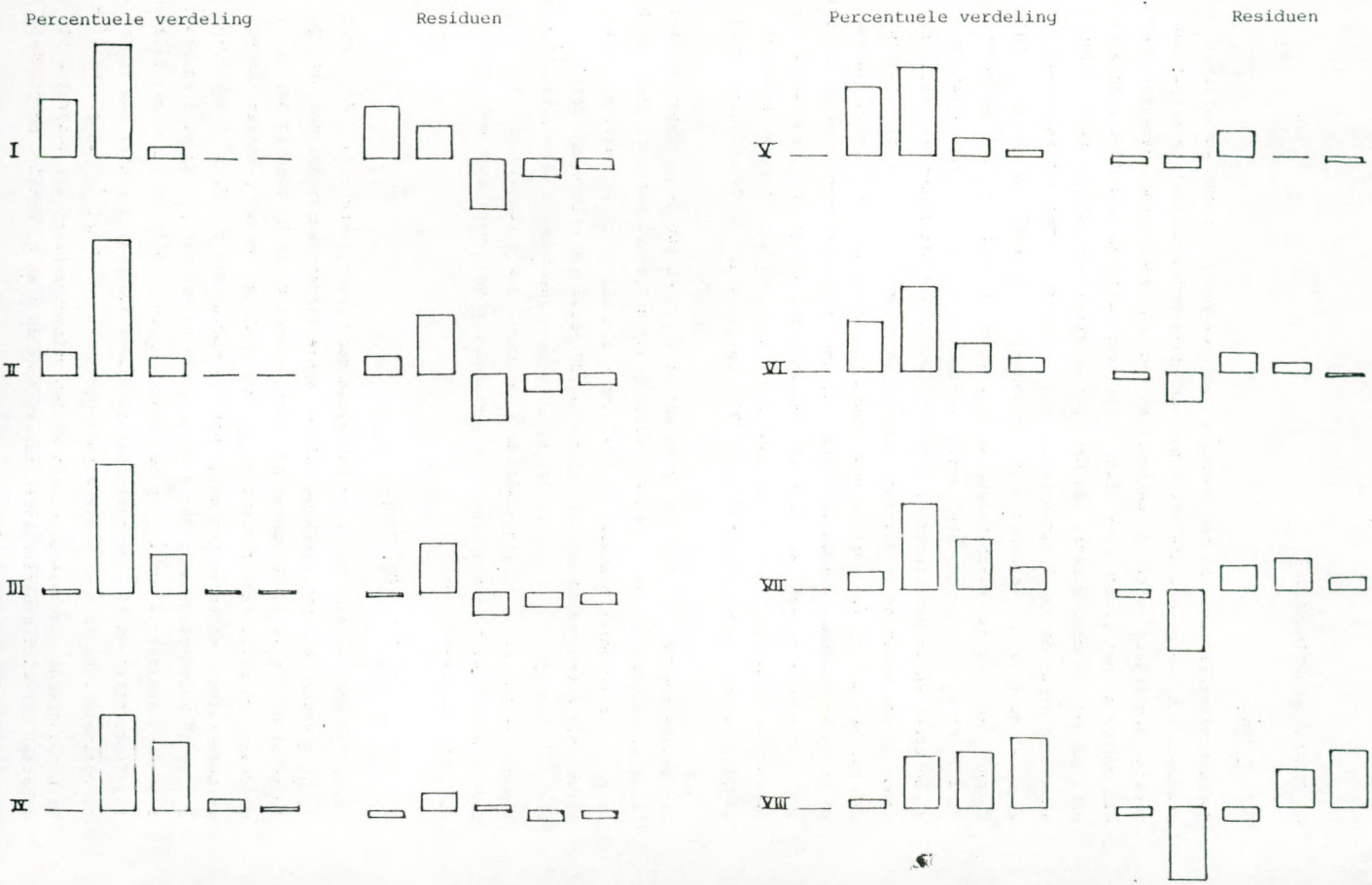
Figuur 7.1. Dendrogram



Figuur 7.2. De puntenwolk en clusters



Figuur 7.3. Inkomensverdelingen van de clusters



8. CORRESPONDENTIEANALYSE

8.1. Techniek

Correspondentieanalyse is een techniek, waarmee men de inhoud van kruistabellen grafisch kan weergeven. Als resultaat van correspondentieanalyse krijgt men meestal een figuur waarin de categorieën van een tabel worden gerepresenteerd als punten in het platte vlak. Een enkele keer heeft men aan twee dimensies niet genoeg. De configuratie van de punten geeft aan welk(e) type(n) van samenhang er in de tabel aanwezig is (zijn). Bekende typen van samenhang zijn: a) de rijen en kolommen zijn splitsbaar in verzamelingen R_1 en R_2 resp. K_1 en K_2 zodanig dat waarnemingen in R_1 samenvallen met die in K_1 en evenzo voor R_2 en K_2 ; b) de tabel wordt voortgebracht door een Guttman-schaal; c) de tabel wordt voortgebracht door een tweedimensionaal normale verdeling (door de waarnemingen in discrete klassen in te delen). Dit soort situaties kan men met behulp van correspondentieanalyse herkennen. De coördinaten van de rijen en kolommen worden berekend door een eigenwaardenprobleem op te lossen. Dit probleem is verwant aan de eigenwaardenvergelijkingen bij principale componenten, discriminantanalyse en canonische correlaties. De techniek wordt uitgebreid beschreven in Benzécri (1976), Lebart e.a. (1977) en Sikkel (1980).

Een benadering van de techniek is om aan de rijen en aan de kolommen van een tabel schaalwaarden toe te kennen, zodanig dat de correlatiecoëfficiënt tussen de rijen en de kolommen maximaal is. We kunnen dit als volgt wat exacter formuleren: zij P een matrix met als elementen de relatieve celfrequenties $\hat{p}_{ij}$ ($i=1, \ldots, M$; $j=1, \ldots, N$); zonder verlies aan algemeenheid veronderstellen we $M \leq N$. Wanneer we met rij i nu schaalwaarde g_i en met kolom j schaalwaarde h_j associëren hebben we twee stochastische variabelen $\underline{g}$ en $\underline{h}$ verkregen met simultane kansverdeling:

$$P\{\underline{g} = g_i \wedge \underline{h} = h_j\} = P_{ij}$$

(hierbij nemen we aan, dat alle schaalwaarden verschillend zijn.) We kunnen $(\underline{g},\underline{h})$ opvatten als één trekking uit een achterliggend kansmodel met kans P_{ij} op resultaat (g_i,h_j). We zoeken nu schaalwaarden g_i en h_j zodanig dat de geschatte correlatiecoëfficiënt $r(\underline{g},\underline{h})$ maximaal is. Om het probleem verder te beschrijven hebben we nog enige notatie nodig. Zij $\hat{p}_{i+}$ de i^e rij-marginaal, $\hat{p}_{+j}$ de j^e kolom-marginaal. Zij R de diagonaalmatrix met $\hat{p}_{i+}$ op de diagonaal, S de diagonaalmatrix met $\hat{p}_{+j}$ op de diagonaal. Dan is $R^{-1}P$ de tabel met relatieve rij-frequenties en PS^{-1} de tabel met relatieve kolom-frequenties. We vinden nu de optimale schalen $\vec{g}$ en $\vec{h}$ als eigenvectoren van $R^{-1}P\,(PS^{-1})'$ resp. $(PS^{-1})'\,R^{-1}P$. De bijbehorende eigenwaarde kunnen worden geïnterpreteerd als kwadraten van (geschatte) correlatiecoëfficiënten, tussen schalen $\underline{g}$ en $\underline{h}$, behalve de grootste eigenwaarde (die gelijk aan 1 is);

deze behoort bij de zgn. triviale schalen: de M-dimensionale vector (1,1,...,1)' en de N-dimensionale (1,1,...,1)' die we verder niet zullen beschouwen. De schalen behorende bij de tweede eigenwaarde zijn de gezochte schalen met de maximale correlatiecoëfficiënt. We vinden echter nog meer schalen. Zo kunnen we de schalen bij de derde eigenwaarde opvatten als die schalen die maximaal gecorreleerd zijn onder voorwaarde dat ze ongecorreleerd zijn met de schalen bij de tweede eigenwaarde.

Wanneer we ons beperken tot de eerste twee paren schalen hebben we per kolom of rij twee coördinaten. Hiervan kan een tweedimensionaal plaatje gemaakt worden dat vaak veel informatie geeft over de inhoud van de bijbehorende tabel. De punten representeren voorwaardelijke verdelingen binnen de rijen of kolommen waar ze bij horen. De oorsprong representeert de marginale verdelingen, dit voor zowel rijen als kolommen.
De afstand van een punt tot de oorsprong kan worden geïnterpreteerd als de mate, waarin de verdeling binnen de bijbehorende categorie afwijkt van de marginale verdeling. De richting vanuit de oorsprong typeert de aard van de afwijking. Wanneer punten dicht bij elkaar liggen lijken de voorwaardelijke verdelingen binnen de rijen of de kolommen op elkaar. Dit geldt uiteraard alleen voor de rijen onderling en de kolommen onderling.
De voorwaardelijke verdelingen van rijen en kolommen zijn uiteraard niet vergelijkbaar. We kunnen met behulp van poolcoördinaten toch een zinvolle interpretatie aan de gecombineerde configuratie van rijen en kolommen geven. Deze interpretatie berust op de volgende principes: a) de oorsprong heeft voor rijen en kolommen dezelfde betekenis (onafhankelijkheid) b) wanneer een rij en kolom in dezelfde richting van de oorsprong afwijken zijn ze positief afhankelijk c) de mate van afhankelijkheid wordt weergegeven door de afstand van de oorsprong.

8.2. Toepassing

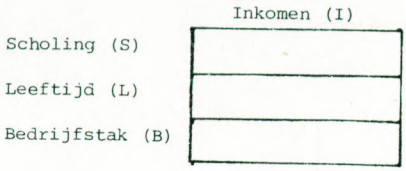
We passen correspondentieanalyse allereerst toe op de 13 x 5 tabel (S + L + B) x I, zie fig. 8.1. Merk op dat ieder individu 3 keer in deze tabel voorkomt, waardoor de 'correlatiecoëfficiënt' anders dan normaal moet worden geïnterpreteerd (zie hoofdstuk 9.4). De structuur van de tabel is bovenaan fig. 8.1. weergegeven. We hebben voldoende theoretische kennis om fig. 8.1. te kunnen interpreteren. De afhankelijke variabele inkomen staat keurig geordend op de eerste, verticale as. Hierbij horen de ordeningen naar leeftijd en scholing. We kunnen dan ook zonder gevaar de eerste as opvatten als een ordening naar inkomen. Opvallend is dat de hoogste twee leeftijdgroepen dicht bij elkaar liggen. Op deze wijze zijn echter ook de bedrijfstakken geordend. Bij kleding/leer (B2) wordt

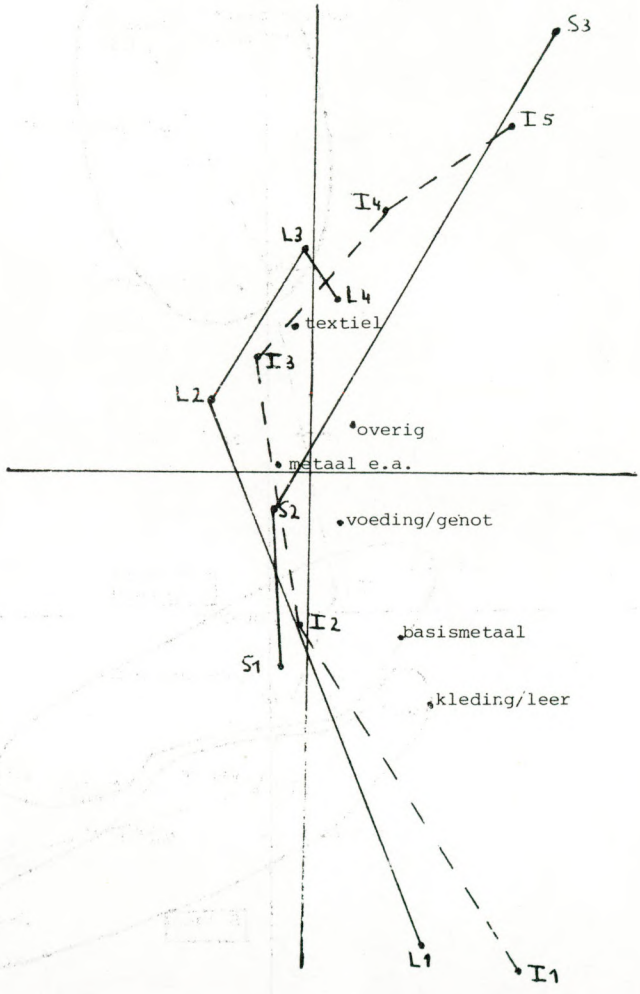
het minst verdiend, bij textiel (B1) het meest. We zien dat de meeste bedrijfstakken in de "holte" van de curve van de ordinale variabelen liggen. Dit kan geïnterpreteerd worden als een hoge variantie binnen die bedrijfstakken. Immers, vanuit de oorsprong "wijzen" ze tegengesteld aan de richting van de middelste inkomensklasse; met andere woorden: ze bestaan voor een relatief groot gedeelte uit de uiterste (lage en hoge) inkomensklassen. Tenslotte zij opgemerkt dat de punten van elk der variabelen een gewogen gemiddelde 0 hebben ofwel, voor alle variabelen ligt het zwaartepunt in de oorsprong. Omdat nu S1 en S2 relatief dicht bij de oorsprong liggen t.o.v. S3 kan men zien dat er relatief weinig hooggeschoolden zijn.

Fig. 8.2. is het resultaat van een correspondentieanalyse op de 18 x 5 tabel (S x B) x I, waarbij dus interacties tussen S en B een rol spelen.
Het is duidelijk dat het verband tussen scholing en inkomen hier domineert. Voor elke bedrijfstak is de volgorde van de scholingscategorieën naar inkomen identiek. Het omgekeerde, dat bij elk opleidingsniveau de volgorde der bedrijfstakken identiek is, gaat niet op. Basismetaal zit bijvoorbeeld bij de hoog opgeleiden relatief hoog, terwijl het bij de lager geschoolden veel lager zit.

Tenslotte is in fig. 8.3 ter vergelijking een zeer populaire vorm van correspondentieanalyse toegepast, namelijk HOMALS (genoemd naar het gelijknamige pakket dat in Leiden is ontwikkeld en is beschreven in Gifi (1980)). Hierbij wordt geen onderscheid gemaakt tussen verklarende en afhankelijke variabelen. In HOMALS worden namelijk alle bimarginalen van de variabelen in de analyse meegenomen. HOMALS is vrijwel identiek aan correspondentieanalyse op de tabel (S+L+B+I) x (S+L+B+I); de beide analyses leveren dezelfde assen, het enige verschil zit in de normalisering van de schalen. Het resultaat staat in fig. 8.3.
Het plaatje blijkt nogal af te wijken van fig. 8.1. Dit komt dan ook door het meenemen van de verbanden tussen de verklarende variabelen.
Zo liggen S3 en L1 in fig. 8.1 ver uit elkaar omdat de eerste categorie een hoog inkomen en de tweede categorie een laag inkomen voorspelt. Echter, beide categorieën hangen positief samen en liggen daarom in fig. 8.3 minder ver uit elkaar. Tweedimensionale figuren van HOMALS kunnen op deze manier voor interpretatieproblemen zorgen, omdat het niet duidelijk is welke samenhangen verantwoordelijk zijn voor de afstanden tussen de verschillende punten.

Figuur 8.1. Correspondentieanalyse op (S + L + B) x I





Figuur 8.2. Correspondentieanalyse op (S x B) x I

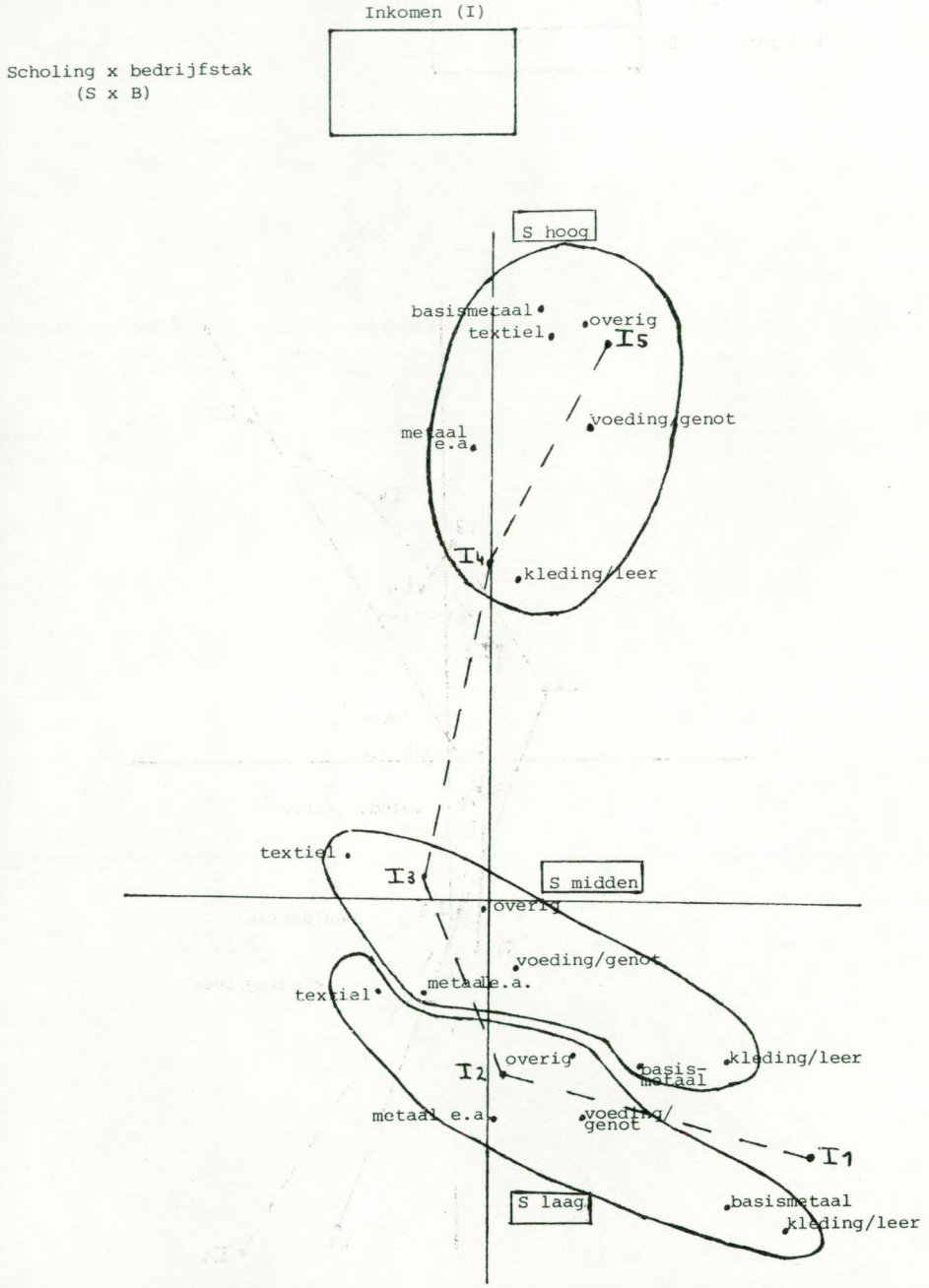
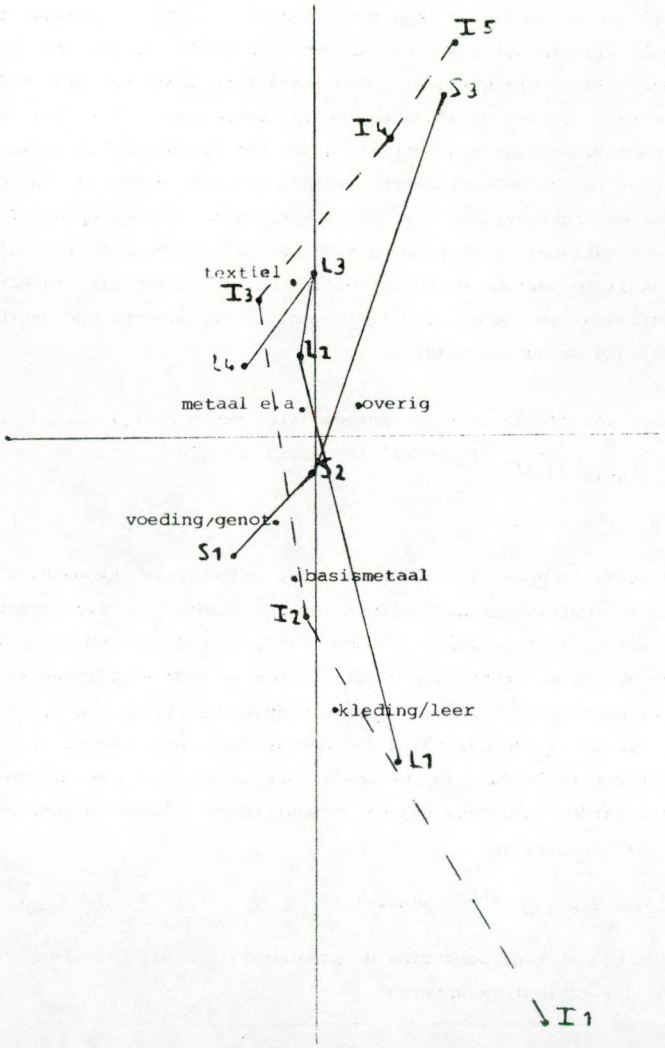


Fig. 8.3. HOMALS op S, L, B en I

Op normalisering na identiek aan correspondentieanalyse op

	Scholing	Leeftijd	Bedrijfstak	Inkomen
Scholing (S)				
Leeftijd (L)				
Bedrijfstak (B)				
Inkomen (I)				



9. REGRESSIEANALYSE

9.1. Inleiding

In dit hoofdstuk willen we de invloed van de onafhankelijke variabelen S, L en B op het inkomen nagaan d.m.v. regressie-analyses. Dit wekt bij sommige lezers misschien verwondering; regressie-analyse wordt immers beschouwd als een techniek waarbij van de afhankelijke variabele wordt verondersteld dat hij minstens op intervalniveau is gemeten, terwijl wij ons in dit rapport beperken tot discrete analyses. Echter, evenals het vaak een goede gewoonte is om de onafhankelijke variabelen als dummy-variabelen te behandelen om op die manier een optimaal verband te krijgen met de numerieke, afhankelijke variabele, zo kan het raadzaam zijn ook de afhankelijke variabele als een set dummy-variabelen te behandelen. Wanneer de afhankelijke variabele van nature nominaal (of ordinaal) is heeft men geen redelijk alternatief; wanneer de afhankelijke variabele numeriek is weet men doorgaans niet (behalve wanneer men over een goede theorie beschikt) welke functionele relatie er tussen de variabelen bestaat. Via discretisering van de variabelen kan men een idee krijgen over de transformatie die men op de afhankelijke, numerieke variabele moet toepassen om het verband tussen afhankelijke en onafhankelijke variabelen te optimaliseren. We zullen dit canonische regressie-analyse noemen, i.t.t. 'gewone' regressie-analyse waarbij de afhankelijke variabele numeriek is.

We kunnen een gewone regressievergelijking van $\underline{y}$ op $x_1, \ldots, x_p$ schrijven als $\underline{y}_\nu = \sum_{j=1}^{p} \beta_j x_{\nu j} + \underline{\varepsilon}_\nu$, ofwel in matrix-notatie

$$\underline{y} = X \beta + \underline{\varepsilon} \quad . \tag{9.1}$$

In ons geval is $\underline{y}$ een 231 833-vector met individuele inkomens, X een 231 833 x p-matrix met als kolomvariabelen dummies (en evt. een constante term), die bepalen in welke klassen van S, L en B een individu zit; verder is β de vector met regressie-coëfficiënten en $\underline{\varepsilon}$ de vector van storingen. Omdat er slechts vijf inkomensscores bestaan schrijven we $\underline{y}$ als $\underline{y} = \underline{Y} w$ met $\underline{Y}$ een 231 833 x 5 dummymatrix, waarbij een 1 in cel (ν,ℓ) aangeeft dat individu ν in inkomensklasse I_ℓ zit en w een vector is met de 5 schaalwaarden behorende bij de verschillende inkomenscategorieën. Formule (9.1) gaat nu over in

$$\underline{Y} w = X \beta + \underline{\varepsilon} \qquad \left(\text{ofwel } \sum_{\ell=1}^{5} w_\ell \underline{Y}_{\nu\ell} = \sum_{j=1}^{p} \beta_j x_{\nu j} + \underline{\varepsilon}_\nu\right) \tag{9.2}$$

We gebruiken voor de schatting de parameters van dit model een variant van de methode der kleinste kwadraten.

Bij een gewone regressie is w een tevoren bepaalde vector. Bij canonische regressie is w een parametervector die moet worden geschat.
Voor de schatting van de regressie-coëfficiënten en de bepaling van R^2 kunnen we op geaggregeerd niveau werken, d.w.z. met de matrices van kruisproducten Y'Y, Y'X en X'X als input. Het is eenvoudig in te zien dat deze matrices direct kunnen worden afgeleid uit frequentietabellen. Ook wanneer we tegelijk optimale schatters voor de scores w en voor de β's willen bepalen, hebben we aan deze matrices voldoende.De kolommen van X dienen lineair onafhankelijk te zijn om de parametervector β eenduidig te kunnen schatten. Hierom laten we één categorie van iedere variabele weg (productvariabelen als S x L worden als één variabele opgevat) en nemen een constante term in de regressievergelijking op, waardoor de eerste kolom van X uit louter enen bestaat.

Als verklarende variabelen worden de volgende combinaties van S, L en B beschouwd:

a. enkele variabelen (S, L of B),
b. meerdere variabelen zonder interacties (S+L, S+B, L+B of S+L+B),
c. meerdere variabelen met interacties (SxL, SxB, LxB of SxLxB).

In het laatste geval spreken we van produktvariabelen; deze worden ieder als één variabele opgevat, omdat ieder individu op slechts één van de klassen scoort. Bij de regressie op de produktvariabele SxLxB hebben we niet de 72 combinaties van klassen van S, L en B als dummies gebruikt, maar bestaan de 72 dummies uit een constante term, univariate marginalen van S, L en B (2+3+5=10 termen) bivariate marginalen van S, L en B (in totaal 6+10+15=31 termen), aangevuld met 2x3x5 = 30 interacties van bijvoorbeeld de vorm S2L1B4.

9.2. Regressie per inkomenscategorie

Als de variabele inkomen als nominale variabele wordt opgevat, kan voor iedere categorie een dummy-variabele worden gecreëerd. Er zijn dan 5 afhankelijke dummy-variabelen y_ℓ (ℓ=1, ..., 5), nl. de kolommen van matrix Y. Voor ieder van deze kolommen kunnen we nu een aparte regressie uitvoeren die overeenkomt met formule (9.1).

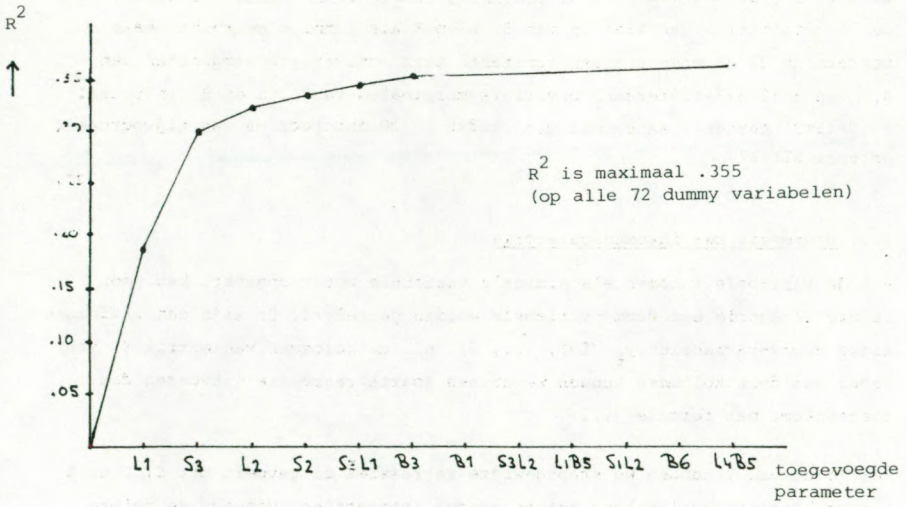
Per kolom van Y hebben we stapsgewijze regressies uitgevoerd met S, L en B als verklarende variabelen, met en zonder interacties. Vanwege de ruimte bespreken we hier slechts enkele resultaten van de 5 regressievergelijkingen met S + L + B als verzameling verklarende variabelen. Het blijkt dat hoe

lager het inkomen is des te 'positiever' is de invloed van L1, terwijl S3 een 'positieve' invloed heeft op de hogere inkomensklassen. De invloed van bedrijfstak is relatief gering.

9.3. Regressie met vaste schaalwaarden

Als de variabele inkomen als numerieke variabele wordt opgevat, kunnen we aan ieder van de vijf inkomenscategorieën scores toekennen. We kiezen hiervoor de bij de klassemiddens aansluitende scores 1 t/m 5. M.a.w.: w' = (1 2 3 4 5). We hebben nu één afhankelijke variabele, die de waarden 1 t/m 5 kan aannemen en kunnen de 'gewone' regressie-analyse uitvoeren volgens (9.2). Het kan eenvoudig aangetoond worden dat de regressiecoëfficienten van deze vergelijking gewogen sommen zijn van de coëfficiënten van de regressies van de vijf afhankelijke dummyvariabelen. De coëfficiënten van de vergelijking voor de eerste inkomenscategorie krijgen gewicht 1, de coëfficiënten van de vergelijking voor de tweede inkomenscategorie krijgen gewicht 2 etc. Voor deze regressie-analyse is een stapsgewijze regressieprocedure met interacties uitgevoerd. De resultaten hiervan zijn weergegeven in figuur 9.1.

Figuur 9.1. Verloop van R^2 bij stapsgewijze regressie met vaste schaalwaarden (incl. interacties)



Hier zien we dat de belangrijkste variabelen te maken hebben met leeftijd en scholing. De belangrijkste variabele is L1. Dus het onderscheid tussen de 20-29 jarigen en de anderen verklaart het meeste van de inkomensverschillen. In de tweede, derde en vierde stap worden resp. S3, L2 en S2 toegevoegd. De eerste interactie is S3L1. Deze interactie heeft een positieve coëfficiënt. Dit betekent dat wanneer het inkomen geschat wordt uit de afzonderlijke variabelen L1, S3, L2 en S2, deze schatting te laag uitkomt voor de personen met een hoge opleiding in de laagste leeftijdscategorie.

9.4. Canonische regressie

In de voorgaande paragraaf zijn we uitgegaan van (9.2) met w een vooraf gespecificeerde vector. In deze paragraaf beschouwen we w als een parametervector. De gewone regressie gaat dan over in canonische regressie. We zoeken nu naar een lineaire combinatie $\underline{Y}w$ van de kolommen van $\underline{Y}$ en een lineaire combinatie $X\beta$ van de kolommen van X, zodanig dat $X\beta$ een optimale predictor is van $\underline{Y}w$. We moeten dus w en β simultaan schatten, en wel zodanig dat R^2 maximaal is. Dit probleem staat beschreven in Gifi(1980) onder de naam CANALS, waarbij uitgegaan wordt van het kleinste kwadraten principe,en in Keller (1980) die uitgaat van het maximum likelihood principe.

Bij canonische regressie op alle categorieën van één variabele (bijv. S,L of B of SxL), is X'X diagonaal. In dit geval zijn w en β op een factor na identiek aan de schalen bij correspondentieanalyse op de bivariate frequentietabel Y'X. In het geval van meer dan één variabele is X'X niet diagonaal, aangezien ieder individu op meer dan één categorie van de x- variabelen scoort. In dit geval leidt canonische regressie tot andere resultaten dan correspondentieanalyse. Zie voor het verband tussen beide technieken Keller (1980) en Sikkel (1980b). Bij correspondentieanalyse moet iedere verklarende variabele (dus zowel S, L als B bij toepassing op S+L+B) zoveel mogelijk 'lijken' op de te schalen inkomensvariabele, terwijl bij canonische regressie naar één lineaire combinatie van de verklarende variabelen wordt gezocht die optimaal is gerelateerd aan de te schalen inkomensvariabele.

In tabel 9.1 staan de waarden van R^2 voor diverse gewone regressies met w' = (1 2 3 4 5) en canonische regressies. Tussen haakjes staan de aantallen vrij te schatten parameters. Dit aantal wordt mede bepaald door de noodzakelijke normalisaties op w en β; zie ook hieronder.

Tabel 9.1. Gekwadrateerde multipele correlatiecoëfficiënten(en aantal vrij te schatten parameters)bij gewone en canonische regressie van I op S, L en B.

Zonder interacties			Met interacties		
Onafhankelijke variabelen	Gewone regressie R^2	Canonische regressie R^2	Onafhankelijke variabelen	Gewone regressie R^2	Canonische regressie R^2
S	0,131 (2)	0,136 (5)			
L	0,204 (3)	0,255 (6)			
B	0,014 (5)	0,019 (8)			
S+L	0,336 (5)	0,355 (8)	SxL	0,345 (11)	0,358 (14)
S+B	0,140 (7)	0,144 (10)	SxB	0,141 (17)	0,144 (20)
L+B	0,216 (8)	0,270 (11)	LxB	0,219 (23)	0,274 (26)
S+L+B	0,344 (10)	0,364 (13)	SxLxB	0,355 (71)	0,371 (74)

We zien in deze tabel dat leeftijd de meest verklarende variabele is en dat leeftijd en scholing samen het meest verklarende paar zijn, welke vorm van regressie men ook verkiest. Bedrijfstak heeft weinig invloed op het inkomen. Deze resultaten komen sterk overeen met bijv. die van de logitanalyse (hoofdstuk 6).

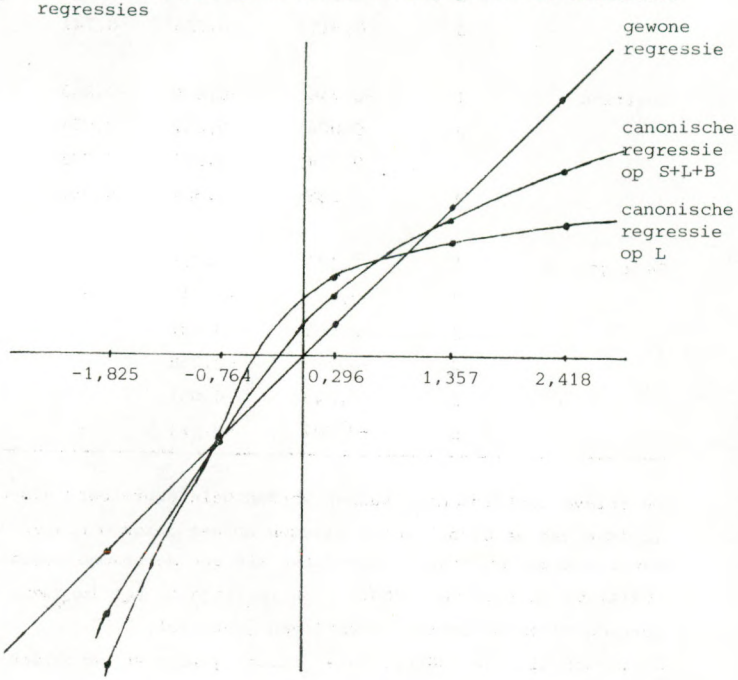
We zien verder in tabel 9.1 dat het toevoegen van interacties weinig oplevert, gegeven de vele extra parameters die dit met zich meebrengt. Het uitvoeren van een canonische i.p.v. een gewone regressie levert wel winst op. Dit wordt nog duidelijker wanneer we de inkomensschalen (w-schalen) met elkaar vergelijken. In tabel 9.2 zijn deze schalen voor verschillende regressies gegeven, nl. voor de 'meest verklarende' combinaties (S,L) en L en de combinatie (S,L,B). We hebben alle w-schalen gestandaardiseerd (gewogen gemiddelde nul en gemiddelde kwadratische afwijking één) teneinde ze vergelijkbaar te maken. De schaal 1 t/m 5 is hierdoor getransformeerd tot w' = (-1,825 -0,764 0,296 1,357 2,418).

Tabel 9.2. Inkomensschalen (gestandaardiseerd)

Inkomensklasse	Gewone regressie	Canonische regressie				
		Zonder interacties			Met interacties	
		S+L+B	S+L	L	SxLxB	SxL
I1	-1,825	-2,440	-2,413	-2,951	-2,371	-2,305
I2	-0,764	-0,797	-0,800	-0,756	-0,801	-0,809
I3	0,296	0,555	0,555	0,736	0,537	0,533
I4	1,357	1,287	1,280	1,049	1,304	1,302
I5	2,418	1,744	1,765	1,208	1,797	1,835

In figuur 9.2 zijn de inkomensschalen van de gewone regressie en van de canonische regressie op L en op S+L+B (de overige schalen vallen hiermee vrijwel samen) uitgezet tegen de inkomensschaal van de gewone regressie.

Figuur 9.2. Vergelijking van inkomensschalen bij gewone en canonische regressies



We zien dat de bij de canonische regressies gevonden inkomensschalen sterk afwijken van die bij de gewone regressie. I1 en I2 blijken bij canonische regressie onderling zeer sterk in schaalwaarden te verschillen, terwijl de verschillen tussen scores van opvolgende inkomensklassen afnemen met toename van het inkomen. Op grond hiervan zien we dat gewone regressie m.b.v. een of andere logaritmische transformatie van de inkomens tot betere resultaten zou hebben geleid dan de hier uitgevoerde gewone regressie. Een fijnere klasse-indeling zou meer informatie geven over het bestaan van een geschikte, weinig parameters bevattende transformatie van de inkomens.

In tabel 9.3 staan de 'regressiecoëfficiënten' van S, L en B bij de regressies zonder interacties. We hebben deze genormeerde regressiecoëfficiënten zodanig bepaald dat per variabele de gemiddelde score over de individuen gelijk is aan nul.

Tabel 9.3. Genormeerde regressiecoëfficiënten bij regressies zonder interacties

Onafhankelijke variabele	Klasse	Gewone regressie op	Canonische regressie op		
		S+L+B	S+L+B	S+L	L
Scholing	1	-0,423	-0,425	-0,437	-
	2	-0,061	-0,040	-0,040	-
	3	0,819	0,726	0,741	-
Leeftijd	1	-0,746	-0,834	-0,833	-0,873
	2	-0,004	0,057	0,055	0,158
	3	0,374	0,394	0,395	0,399
	4	0,386	0,384	0,386	0,295
Bedrijfstak	1	0,143	0,176	-	-
	2	-0,174	-0,243	-	-
	3	0,092	0,081	-	-
	4	-0,217	-0,258	-	-
	5	-0,049	-0,031	-	-
	6	-0,007	-0,021	-	-

Positieve coëfficiënten kunnen worden geïnterpreteerd als een positieve invloed van de bijbehorende klassen op het inkomen t.o.v. het gemiddelde inkomen. Zowel uit de canonische regressies als uit de gewone regressies volgt de positieve invloed van scholing en leeftijd op het inkomen. Vooral de categorieën S3 en L1 spelen hierbij een grote rol.
De verschillen in coëfficiënten tussen gewone en canonische regressie zijn groter dan tussen de verschillende vormen van canonische regressie onderling.

De coëfficiënten voor L3 en L4 (40-49 jr. resp. 50-64 jr.) zijn in bijna alle gevallen van gelijke orde van grootte. Aggregatie van deze twee leeftijdscategorieën kan dus op grond van deze analyses zonder bezwaar gebeuren.
Alhoewel de bijdrage aan de verklaring van de variabele 'bedrijfstak' gering is, lopen de coëfficiënten voor de verschillende bedrijfstakken nog behoorlijk uiteen. De volgorde van de bedrijfstakken, geordend van laag naar hoog inkomen, is basismetaal (B4), kleding/leer (B2), metaal e.a. (B5), voeding/genot (B6), overig (B3), textiel (B1). De in tabel 9.3 genoemde bedrijfstak-coëfficiënten zijn op te vatten als een soort gestandaardiseerde gemiddelde inkomens. Dit aangezien door de regressie op S+L+B als het ware wordt gecorrigeerd voor verschillen in leeftijds- en scholingsverdeling tussen de bedrijfstakken. Een vollediger standaardisatie voor de invloed van L en S kan men bereiken door regressie uit te voeren op S×L+B.

10. VERGELIJKING TUSSEN DE METHODEN

10.1. Inleiding

In dit hoofdstuk worden de behandelde technieken volgens enkele criteria met elkaar vergeleken. We hebben de criteria in 3 groepen verdeeld, nl. "input"(welke structuur wordt aan de data opgelegd?), "throughput" (reken-aspecten) en "output" (wat voor informatie levert de methode op?).

We kunnen bij de vergelijking van de hoofdstukken 3 t/m 9 niet zonder meer spreken van 7 technieken. In de eerste plaats is 'het kijken naar meerdimensionale tabellen' formeel geen techniek. In dat hoofdstuk zijn echter wel standaardiseringstechnieken gebruikt. Daarnaast worden in sommige hoofdstukken verschillende varianten van een techniek getoond. Of een bepaalde techniek aan een bepaald criterium voldoet hangt vaak mede af van de gebruikte variant. Desondanks hebben we summier weergegeven welke hoofdstukken aan de verschillende criteria voldoen. Discussies over sommige uitspraken zijn zeker mogelijk.

Grofweg kunnen we de technieken al bij voorbaat indelen in:

a. Klassificatie-technieken ('selectie m.b.v. G^2' en clusteranalyse).
b. Schaalwaarden opleverende technieken (log-lineaire analyse, correspondentieanalyse en regressieanalyse).
c. Interpretatieve technieken ('het kijken' en ridit-analyse).

10.2. Input

1 of 2 sets variabelen.

De analyses hadden tot doel I 'te verklaren' uit de variabelen S, L en B. Vandaar dat technieken zijn gekozen die uitgaan van 2 sets variabelen, I en (S,L,B). Deze technieken gaan uit van tabellen als (SxLxB)xI of (S+L+B)xI. Sommige van de gebruikte technieken zijn evenwel ook toepasbaar wanneer men over één set beschikt; men gaat dan uit van de vierdimensionale tabel SxLxBxI. In dit rapport genoemde voorbeelden hiervan zijn de HOMALS-versie van correspondentieanalyse en het algemene log-lineaire model.

Symmetrisch vs. asymmetrisch

Omdat de 2 sets variabelen I en (S,L,B) gekenmerkt zijn als afhankelijk resp. onafhankelijk behandelen de meeste technieken de beide sets op verschillende wijze. Zo worden bij clusteranalyse slechts rijen uit de

72x5-tabel samengevoegd. Bij correspondentieanalyse is wel sprake van een symmetrie; immers, toepassing van de techniek op bijv. (S+L+B)xI levert dezelfde resultaten op als toepassing op Ix(S+L+B). Sommige asymmetrische technieken zijn symmetrisch te maken via specifieke definities. Zo kan men bij de klassificatietechnieken naast het groeperen van rijen ook het groeperen van kolommen toelaten.

Meer afhankelijke variabelen

In principe zijn alle technieken toepasbaar wanneer men over meer dan één discrete afhankelijke variabele beschikt. Men kan immers de produktruimte van deze variabelen beschouwen als één variabele. De interpretatie kan echter tot problemen leiden. Het eenvoudigst lijkt wat dit betreft de uitbreiding van canonische regressie.

Meer onafhankelijke variabelen

De technieken die stapsgewijs kunnen worden toegepast zijn uiterst geschikt wanneer men over vele onafhankelijke variabelen beschikt. Een techniek als 'selectie m.b.v. G^2' komt dan zelfs beter tot zijn recht. Het 'kijken' wordt evenwel een nog moeizamer bezigheid. Opgemerkt moet worden dat de stapsgewijze methoden niet altijd in staat zijn een bestaande optimale oplossing te vinden.

Stochastische eigenschappen en achterliggend model

Hoewel de technieken exploratief zijn gebruikt en er van te voren geen veronderstellingen zijn gemaakt over het stochastisch proces dat de data heeft gegenereerd, kan achter iedere methode een stochastisch model worden gedacht. Berekende fracties, ridits en regressiecoëfficiënten zijn op te vatten als schattingen van parameters, en klassificaties als afkomstig van verschillende populaties. Bij alle technieken behalve 'het kijken' en - strict genomen - de ridits bestaan toetsen op onafhankelijkheid.
De methoden zijn echter toegepast als reductietechnieken, waarbij is gestreefd de data zo goed mogelijk weer te geven m.b.v. enkele parameters of klassen. De hierbij ontstane residuen bleven echter zeer significant, wat mede het gevolg is van het grote aantal waarnemingen. Met behulp van verschillende technieken kan men dus modellen specificeren over de mate van afhankelijkheid van inkomen van S, L en B; wanneer men over veel waarnemingen beschikt zullen deze modellen echter worden verworpen.

Meetniveau

In de geheime fase zijn alle variabelen als nominale variabelen beschouwd. Voor alle methoden was dit meetniveau voldoende, behalve voor riditanalyse,

omdat daarbij vooraf kennis omtrent de ordening van de te verklaren variabele noodzakelijk is.

Na de onthulling van het geheime bestand is bij 'het kijken' en bij gewone regressie gebruik gemaakt van het interval-karakter van I door de schaal 1,2,...,5 te gebruiken; deze schaal kwam nl. redelijk overeen met de klassemiddens van de inkomensklassen.

Interacties

Bij alle methoden wordt gekeken naar het aanwezig zijn van interacties. Bij logit-analyse en regressie-analyse kunnen de interacties worden gekwantificeerd; hierbij is overigens sprake van enigszins verschillende definities van interactie (multiplicatief vs. additief).

Steekproefontwerp

Alle methoden zijn toepasbaar bij enkelvoudig aselecte steekproeven. Bij de meeste andere steekproefontwerpen kan men de methoden doorgaans alleen toepassen op herwogen of opgehoogde aantallen. In dergelijke gevallen is het echter bij alle technieken moeilijk of onmogelijk om nog iets te zeggen over de stocnastische eigenschappen van de uitkomsten.

10.3. Throughput

Software

Bij 'het kijken' was de opzet om hoogstens een zakrekenmachine te gebruiken. Ridit-analyse, 'selectie m.b.v. G^2' en clusteranalyse kunnen op een tafelcomputer worden toegepast. De hier toegepaste variant van clusteranalyse is in de meeste clusteranalysepakketten niet geïmplementeerd. Voor log-lineaire-analyse bestaan standaardprogramma's, bijv. ECTA en BMDP/3F. Voor correspondentieanalyse en canonische regressie zijn in Nederland programma's ontwikkeld door de afd. Datatheorie van de Rijksuniversiteit te Leiden en in beperkter mate door het C.B.S.

Besturing

Ingrijpen door de onderzoeker tijdens het rekenproces is bij enkele technieken in meerdere of mindere mate nodig, het meeste bij 'het kijken', dat in principe geheel handwerk is, en bij ridit-analyse. Overigens is bij de meeste technieken een volledig automatische selectieprocedure denkbaar.

10.4. Output

Geheim bestand

'Selectie m.b.v. G^2', log-lineaire analyse, clusteranalyse, correspondentie-analyse en canonische regressie analyse ondergingen geen enkele verandering als gevolg van het bekend worden van de betekenis der variabelen. De verbanden tussen de 4 variabelen, waaronder de ordinale relaties, volgden direct uit de analyse. Bij al deze technieken bleek het onderscheid tussen de oudste 2 leeftijdsklassen zeer klein. Ook bij 'het kijken' bleken de veronderstelde orderelaties gegrond. Voor riditanalyse was inzicht in het ordinale karakter en de ordening van de afhankelijke variabele noodzakelijk. Riditanalyse en de analyses die van de schaal 1,2,...,5 gebruik maken zijn pas op het onthulde bestand toegepast. Er kan met recht worden gezegd dat de technieken vrij objectieve resultaten hebben opgeleverd. Het feit dat achteraf bleek dat we met "harde" variabelen te maken hebben gehad betekende dat meetfouten een niet te grote rol hebben gespeeld. Toch zijn we van mening dat ook in het geval van "zachte" nominale variabelen (bijv. attitude-variabelen) dezelfde analyses kunnen worden gedaan.

Aanpassingsmaat

Bij de meeste technieken bestaat een maat voor de aanpassing (goodness of fit), nl. R^2 bij regressie-analyse, Pearson's χ^2 bij clusteranalyse en correspondentieanalyse, en G^2 bij 'selectie m.b.v. G^2' en log-lineaire analyse. Per techniek kan men in principe de aanpassingsmaat gebruiken om verschillende modellen met elkaar te vergelijken.

Schaalwaarden

Correspondentieanalyse en canonische regressie leveren aparte schaalwaarden voor de onafhankelijke en afhankelijke variabelen. Bij deze technieken kunnen meerdere orthogonale schalen worden bepaald, hetgeen vooral bij een 'strikt nominale' afhankelijke variabele (zonder ordeningsrelatie) zeer zinvol is. Bij gewone regressie is de afhankelijke variabele reeds geschaald en worden slechts schaalwaarden geleverd voor de onafhankelijke variabelen. Logit-analyse levert per onafhankelijke variabele (of een combinatie ervan) 5 inkomensschaalwaarden. Qua interpretatie levert deze techniek daardoor meer moeilijkheden op. Bij de overige technieken kan men achteraf per gevormde klasse van S,L en B een inkomensscore definiëren, omdat het hier een ordinale variabele betreft. Om dit te kunnen doen heeft men echter a priori schaalwaarden voor de inkomensklassen nodig (bijv. ridits of 1,2,...,5).

Standaardfouten

Bij ridit-analyse en bij gewone regressie-analyse zijn eenvoudig standaardfouten te berekenen. Bij sommige andere technieken, bijv. bij canonische regressie, kan men asymptotische standaardfouten en betrouwbaarheidsintervallen bepalen.

10.5. Conclusies

De technieken hebben min of meer tot dezelfde resultaten geleid: Leeftijd is de belangrijkste verklarende variabele, vervolgens scholing, terwijl bedrijfstak relatief weinig doet. Ook interacties zijn relatief onbelangrijk. Bij de klassificatie-technieken is men vooral geneigd de grote verbanden te bekijken; dit geldt ook voor de technieken waarbij men over een aanpassingsmaat beschikt omdat de geringe invloed van bijv. interacties meteen blijkt. Bij 'het kijken' en bij ridit-analyse viel meer de nadruk op de minder triviale verbanden.

Het niet bekend zijn met de betekenis van de variabelen heeft nauwelijks invloed gehad op de conclusies. Slechts enkele aanvullende analyses zijn gedaan na de onthulling van het bestand.

Tenslotte zij nogmaals benadrukt dat de poging tot evaluatie in dit hoofdstuk voor een groot deel een voorlopig karakter draagt en dat de kwaliteiten (of het gebrek daaraan) van de verschillende methoden nog lang niet uitputtend zijn onderzocht. Wel is het duidelijk dat een 'beste' of 'ideale' methode niet bestaat. Meer zicht op de relatieve merites van de verschillende methoden kan worden verkregen door eigenschappen theoretisch te vergelijken. Zo kan men een indeling maken van de discrete multivariate technieken, gebaseerd op Pearson's χ^2 en de likelihoodratio G^2. Alle technieken uit de hoofdstukken 5 t.m. 9 kan men met behulp van beide criteria (of generalisaties ervan) indelen. Vaak bestaat een analogon voor de gepresenteerde techniek door gebruik van het G^2-criterium in plaats van Pearson's χ^2-criterium, of omgekeerd. Men kan de technieken ook beschrijven aan de hand van het'errors-in-variables' model. In deze richting zal nader onderzoek plaatsvinden.

BIJLAGE Verdeling van het inkomen (I) per combinatie van scholing, leeftijd en bedrijfstak (S,L,B)

Indices			Frequentieverdeling						Procentuele verdeling (%)				
S	L	B	I1	I2	I3	I4	I5	Totaal	I1	I2	I3	I4	I5
1	1	1	34	237	8	2	0	281	12	84	3	1	0
1	1	2	174	295	6	0	0	475	37	62	1	0	0
1	1	3	752	1917	120	16	8	2813	27	68	4	1	0
1	1	4	221	212	0	0	0	433	51	49	0	0	0
1	1	5	545	2081	197	0	0	2823	19	74	7	0	0
1	1	6	465	1052	112	12	0	1641	28	64	7	1	0
1	2	1	8	170	134	14	0	326	2	52	41	4	0
1	2	2	18	361	84	31	0	494	4	73	17	6	0
1	2	3	40	1207	647	94	144	2132	2	57	30	4	7
1	2	4	33	338	55	2	0	428	8	79	13	0	0
1	2	5	13	2289	929	106	37	3374	0	68	28	3	1
1	2	6	5	957	320	74	31	1387	0	69	23	5	2
1	3	1	0	242	349	36	22	649	0	37	54	6	3
1	3	2	29	410	104	15	20	578	5	71	18	3	3
1	3	3	10	1293	1345	262	94	3004	0	43	45	9	3
1	3	4	22	403	159	14	7	605	4	67	26	2	1
1	3	5	23	2836	1676	122	113	4770	0	59	35	3	2
1	3	6	28	930	739	115	50	1862	2	50	40	6	3
1	4	1	4	358	403	87	39	891	0	40	45	10	4
1	4	2	41	350	70	12	24	497	8	70	14	2	5
1	4	3	41	1818	1406	325	170	3760	1	48	37	9	5
1	4	4	29	492	326	35	21	903	3	54	36	4	2
1	4	5	20	3314	1889	388	238	5849	0	57	32	7	4
1	4	6	26	1173	770	117	110	2196	1	53	35	5	5
2	1	1	69	1236	210	4	2	1521	5	81	14	0	0
2	1	2	614	1014	115	5	4	1752	35	58	7	0	0
2	1	3	2111	10706	1510	112	30	14469	15	74	10	1	0
2	1	4	584	899	190	0	0	1673	35	54	11	0	0
2	1	5	1783	13838	1507	91	22	17241	10	80	9	1	0
2	1	6	786	3846	504	39	16	5191	15	74	10	1	0
2	2	1	0	567	1010	126	25	1728	0	33	58	7	1
2	2	2	44	826	580	97	61	1608	3	51	36	6	4
2	2	3	38	5055	6905	1541	442	13981	0	36	49	11	3
2	2	4	31	824	488	86	23	1452	2	57	34	6	2
2	2	5	30	9959	7756	781	240	18766	0	53	41	4	1
2	2	6	108	2179	2229	443	143	5102	2	43	44	9	3

BIJLAGE(vervolg)

Indices			Frequentieverdeling						Procentuele verdeling (%)				
S	L	B	I1	I2	I3	I4	I5	Totaal	I1	I2	I3	I4	I5
2	3	1	0	356	1230	417	88	2091	0	17	59	20	4
2	3	2	12	246	517	106	76	957	1	26	54	11	8
2	3	3	38	3485	6296	2384	1283	13486	0	26	47	18	10
2	3	4	8	722	805	145	111	1791	0	40	45	8	6
2	3	5	56	5322	9538	2751	913	18580	0	29	51	15	5
2	3	6	32	1265	2000	731	399	4427	1	29	45	17	9
2	4	1	0	261	612	362	165	1400	0	19	44	26	12
2	4	2	14	372	285	102	97	870	2	43	33	12	11
2	4	3	79	2906	4363	1960	1118	10426	1	28	42	19	11
2	4	4	23	627	478	111	87	1326	2	47	36	8	7
2	4	5	5	3629	6014	1770	1058	12476	0	29	48	14	8
2	4	6	38	1483	1532	594	322	3969	1	37	39	15	8
3	1	1	1	256	155	12	0	424	0	60	37	3	0
3	1	2	0	62	17	1	1	81	0	77	21	1	1
3	1	3	55	1048	640	197	4	1964	3	54	33	10	0
3	1	4	0	40	65	0	0	105	0	38	62	0	0
3	1	5	26	1841	1180	168	9	3224	1	57	37	5	0
3	1	6	25	379	159	33	3	599	4	63	27	6	1
3	2	1	0	45	344	276	87	752	0	6	46	37	12
3	2	2	0	9	50	32	36	127	0	7	39	25	28
3	2	3	11	373	1781	1325	705	4195	0	9	42	32	17
3	2	4	0	21	99	92	65	277	0	8	36	33	23
3	2	5	1	506	3398	1456	643	6004	0	8	57	24	11
3	2	6	0	146	302	304	136	888	0	16	34	34	15
3	3	1	1	6	156	217	232	612	0	1	25	35	38
3	3	2	0	18	31	13	15	77	0	23	40	17	19
3	3	3	0	109	808	1091	1315	3323	0	3	24	33	40
3	3	4	0	6	82	114	77	279	0	2	29	41	28
3	3	5	12	127	1661	1483	1625	4908	0	3	34	30	33
3	3	6	2	34	203	249	303	791	0	4	26	31	38
3	4	1	0	4	37	103	174	318	0	1	12	32	55
3	4	2	0	16	8	15	8	47	0	34	17	32	17
3	4	3	0	76	339	460	767	1642	0	5	21	28	47
3	4	4	0	26	62	36	63	187	0	14	33	19	34
3	4	5	1	98	634	583	885	2201	0	4	29	26	40
3	4	6	0	22	99	69	164	354	0	6	28	19	46

Literatuur

Andersen, E.B., 1980, Discrete statistical models with social science applications (Noord-Holland, Amsterdam).

Benzécri, J.P., 1976, L'analyse des données, tome 1: La taxinomie, tome 2: Correspondances (Dunod, Paris).

Berkson, J., 1938, Some difficulties of interpretation encountered in the application of the chi-square test, Journal of the American Statistical Association, nr. 33, pp. 526-536.

Bethlehem, J.G., 1976, Handleiding voor hiërarchische clusteranalyse, rapport SN 5/76 (Mathematisch Centrum, Amsterdam).

Bishop, Y.M., S.E. Fienberg, P.W. Holland, 1975, Discrete multivariate analysis (M.I.T. Press, Cambridge, Massachusetts).

Bross, I.D.J., 1958, How to use ridit analysis, Biometrics 19, pp. 18-38.

Ende, H.W. van den, 1971, Beschrijvende statistiek voor gedragswetenschappen (Agon Elsevier, Amsterdam).

Everitt, B.S., 1974, Cluster analysis (Heinemann Educational Books, London).

Everitt, B.S., 1978, Graphical techniques for multivariate data (Heinemann Educational Books, London).

Fienberg, S.E., 1977, The analysis of cross-classified categorical data (M.I.T. Press, Cambridge, Massachusetts).

Fleiss, J.L., 1973, Statistical methods for rates and proportions (Wiley, New York).

Gifi, A., 1980, Niet-lineaire multivariate analyse (afd. Datatheorie, Rijksuniversiteit, Leiden).

Goldstein, M. and W.R. Dillon, 1978, Discrete discriminant analysis (Wiley, New York).

Goodman, L.A., 1971, The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications, Technometrics 13, pp. 33-61.

Goodman, L.A., 1978, Analyzing qualitative/categorical data (Abt Books, Cambridge, Massachusetts).

Haberman, S.J., 1974, The analysis of frequency data (The University of Chicago Press).

Israëls, A.Z., 1979, Clusteranalyse met behulp van de chi-kwadraat afstand, interne C.B.S.-nota.

Israëls, A.Z. en J. van Driel, 1980, The use of the chi-square statistic in the analysis of multiway tables, interne C.B.S.-nota.

Jansen, M.E., 1980a, Ridit-analysis, interne CBS-nota.

Jansen, M.E., 1980b, Ridit-analyse: theorie en toepassing, in: C.B.S. Select 1, pp. 399-412 (Staatsuitgeverij, Den Haag).

Keller, W.J., 1980, Het 'errors-in-variables' model (EVM) en andere multivariate technieken, interne C.B.S.-nota.

Kullback, S., 1959, Information theory and statistics (Wiley, New York).

Lebart, A., A. Morineau, N. Tabard, 1977, Techniques de la description statistique (Dunod, Paris).

Leamer, E.E., 1978, Specification searches; ad hoc inference with nonexperimental data (Wiley, New York).

Sikkel, D., 1980a, An application of correspondence analysis to leisure time activities, Statistical Studies nr. 25 (Staatsuitgeverij, Den Haag).

Sikkel, D., 1980b, Verschillen tussen correspondentieanalyse en canonieke correlaties, interne C.B.S.-nota.

Zahn, D.A. and S.B. Fein, 1979, Large contingency tables with large cell frequencies: A model search algorithm and alternative measures of fit, Psychological Bulletin 86, pp. 1189-1201.