KM 40 (1992)
pg.23-46

LOGISTISCHE MODELLEN VOOR POLYTOME AFHANKELIJKE VARIABELEN

Jos Dessens
Wim Jansen[1]

Samenvatting

Sociale wetenschappers houden zich bezig met het verklaren en voorspellen van gedrag. Vaak betekent dit dat een variabele Y als afhankelijk wordt gezien van een aantal predictoren X_1, X_2, ..., X_p. Indien de Y-variabele een 'continue' variabele is, dan ligt, afhankelijk van de aard van de predictoren en eventueel na een transformatie, regressie- en/of variantie-analyse voor de hand. Vaak echter is de Y-variabele nominaal of hoogstens geordend, met slechts weinig categorieën. In deze situaties zullen de in regressie- en/of variantie-analyse gebruikelijke modelveronderstellingen in meer of minder ernstige mate zijn geschonden. Dit kan leiden tot onzuivere parameterschattingen. De aard van de vertekeningen zullen we toelichten aan de hand van een dichotome Y-variabele. Vervolgens worden een aantal modellen voor Y-variabelen met meer dan twee categorieën besproken, waarbij weer een onderscheid gemaakt kan worden tussen geordende en niet geordende categorieën. De variabelen die als predictoren in deze modellen optreden kunnen zowel 'discreet' als 'continu' zijn. Een aantal van de modellen wordt geïllustreerd aan de hand van bestaande of nieuwe data. Telkens zal worden aangegeven met welke software de modellen kunnen worden geanalyseerd.

---

[1] Faculteit Sociale Wetenschappen, Vakgroep Sociologie, Rijksuniversiteit Utrecht, Heidelberglaan 2, 3584 CS Utrecht, 030-532101. EMAIL: <SSLZEJJ@CC.RUU.NL>

De auteurs zijn dank verschuldigd aan Peter van der Heijden, Jeroen Weesie en een anonieme beoordelaar voor hun commentaar op een eerdere versie van dit artikel.

1. Inleiding: een dichotome Y-variabele

Bij het verklaren en voorspellen van gedrag wordt een variabele Y vaak afhankelijk verondersteld van een aantal predictoren X_1, X_2, ..., X_p. Voor een continue Y-variabele ligt, eventueel na een transformatie, regressie- en/of variantie-analyse voor de hand. In de sociale wetenschappen zal de Y-variabele vaak nominaal zijn of hoogstens geordend met een klein aantal categorieën. Voor de situatie dat de Y-variabele dichotoom is, kunnen de waarnemingen worden voorgesteld via $(Y_i, \mathbf{x}_i)$, waarbij $\mathbf{x}_i$ een vector is met waarden van de i-de waarneming op de predictoren X_{1i}, X_{2i}, ..., X_{pi}. Indien de waarnemingen onafhankelijk zijn, dan is $Y_i|\mathbf{x}_i$ binomiaal verdeeld, waarbij

$$Y_i = \begin{cases} 1 & \textit{met kans} \quad \pi(\mathbf{x}_i) \\ 0 & \textit{met kans} \quad 1-\pi(\mathbf{x}_i) \end{cases} \tag{1}$$

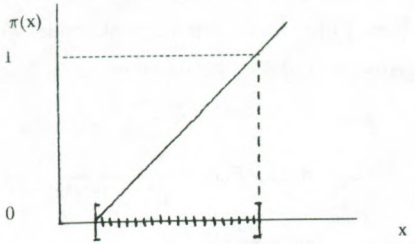
Gewone regressie-analyse in deze situatie veronderstelt een lineaire relatie tussen $\pi(\mathbf{x}_i)$ en de predictoren,

$$E(Y_i) = \pi(\mathbf{x}_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}. \tag{2}$$

Model (2) wordt wel het lineaire kansmodel genoemd. Om een aantal redenen ligt het gewone regressie-analyse model niet zo voor de hand[2]. Ten eerste zijn niet alle waarden van de predictoren toegestaan. Strikt genomen zijn slechts twee waarden voor $E(Y_i)$ toegestaan, namelijk 0 en 1. Aangezien $E(Y_i)$ een kans voorstelt, ligt het voor de hand om alle waarden tussen 0 en 1 toe te staan. Voor de eenvoudige situatie met één predictor variabele X is in figuur 1 direct duidelijk dat X-waarden buiten het gearceerde interval op de horizontale as leiden tot voorspelde Y-waarden buiten het interval [0,1].

---

[2] Indien de doelstelling niet is voorspelling maar het onderscheiden van groepen, dan kan in deze situatie discriminantanalyse worden toegepast (Fienberg, 1980). Voor consistente schatting van de parameters is het echter vereist, dat de predictoren multivariaat normaal zijn verdeeld (Goldstein & Dillon, 1978).



Figuur 1. Toegestane X-waarden bij regressie van een dichotome Y-varabele op een continue X-variabele.

Ten tweede is per definitie niet voldaan aan de binnen het regressiemodel vigerende veronderstelling van homogene varianties van Y_i gegeven $\mathbf{x_i}$. Immers $VAR(Y_i|\mathbf{x_i}) = \pi(\mathbf{x_i})(1-\pi(\mathbf{x_i}))$. Een oplossing is om, in plaats van gewone kleinste kwadraten (OLS) als schattingsmethode, gewogen kleinste kwadraten (WLS) te gebruiken. De weegfactoren zijn dan de inversen van $VAR(Y_i|\mathbf{x_i})$. Aangezien deze varianties schattingen zijn op basis van de data, dienen de weegfactoren in een aantal opeenvolgende iteraties te worden bepaald.
Ten derde geldt dat Y zeker niet normaal is verdeeld, zodat strikt genomen geen statistische toetsen beschikbaar zijn.
Een vierde punt is van meer inhoudelijke aard. Meestal is het onwaarschijnlijk dat de relatie tussen X-variabelen en $\pi(\mathbf{x_i})$ lineair is. Beschouw als voorbeeld de relatie tussen de kans op het kopen van een nieuwe auto en het gezinsinkomen. Het ligt niet voor de hand dat een vaste toename van het inkomen eenzelfde effect heeft op de op het kopen van een nieuwe auto over het gehele bereik van inkomen.
Om deze redenen zijn regressiemodellen ontwikkeld, waarbij de relatie tussen X en $\pi(\mathbf{x_i})$ monotoon wordt verondersteld en waarbij tevens voldaan is aan $0 \leq \pi(\mathbf{x_i}) \leq 1$. Cumulatieve dichtheidsverdelingen voldoen hier bij uitstek aan. Zo kan voor een cumulatieve dichtheidsverdeling F het F-lineaire model, worden geformuleerd, waarin eenvoudigheidshalve slechts één predictor is opgenomen:

$$\pi(x_i) = F(\alpha+\beta x_i) \quad \textit{en dus:} \quad F^{-1}(\pi(x_i)) = \alpha+\beta x_i \tag{3}$$

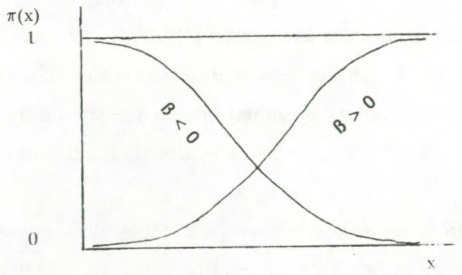
Drie kansdichtheidsverdelingen worden veel gebruikt. De met deze kansdichtheidsverdelingen corresponderende modellen zijn: het logistische regressiemodel, het probit model en het extreme waarden model. In het logistische regressiemodel wordt voor de relatie tussen x en π de cumulatieve logistische verdeling genomen:

$$\pi(x_i) = F(x_i) = \frac{1}{1 + e^{-(\alpha+\beta x_i)}}$$

en dus: (4)

$$\pi(x_i) = \frac{\exp(\alpha+\beta x_i)}{1 + \exp(\alpha+\beta x_i)}$$

De vorm van de relatie tussen X en $\pi(x_i)$ voor respectievelijk positieve en negatieve waarden van ß is in figuur 2 weergegeven.



Figuur 2. De vorm van de relatie tusen X en $\pi(x_i)$ onder het logistische regressiemodel voor ß<0 en ß>0.

De uitdrukking (4) is te schrijven via de 'odds' van $\pi(x_i)$ als:

$$odds(\pi(x_i)) = \frac{\pi(x_i)}{1-\pi(x_i)} = \exp(\alpha + \beta x_i) \quad (5)$$

De natuurlijke logaritme van de odds leidt derhalve tot een lineair model in X. Indien alle predictoren in het model discreet zijn, spreekt men van logitanalyse. Voor deze situatie is

vrij eenvoudig na te gaan dat het logitmodel kan worden verkregen uit een loglineair model. Zij m_{ij} (i=1,2;j=1,2,..., k) de onder een loglineair model verwachte frekwenties voor een 2xk tabel, dan geldt:

$$\begin{aligned} \log(m_{1j}) - \log(m_{2j}) = \log\left(\frac{m_{1j}}{m_{2j}}\right) &= \log\left(\frac{m_{1j}}{m_{+j}} \Big/ \frac{m_{2j}}{m_{+j}}\right) \\ &= \log\left(\frac{\pi_{1j}}{\pi_{2j}}\right) \\ &= \log\left(\frac{\pi_{1j}}{1-\pi_{1j}}\right) \end{aligned} \qquad (6)$$

Indien naast discrete predictoren ook continue predictoren in het model voorkomen, dan spreekt men van logistische-regressie analyse.

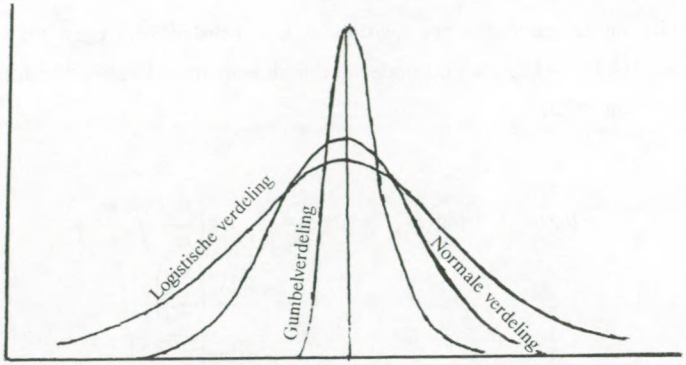
De keuze voor de normale verdeling leidt tot het zogenaamde probit model. Het extreme waarden model wordt verkregen door voor de cumulatieve kansdichtheidsfunctie F ter karakterisering van de relatie tussen x en π de Gumbelverdeling te nemen:

$$\pi(x_i) = F(x_i) = 1 - \exp(-\exp(\alpha + \beta x_i)) \qquad (7)$$

Deze verdeling is niet symmetrisch (zie figuur 3).

De keuze voor een van de genoemde modellen hangt af van de specifieke toepassing. In een groot aantal toepassingen voldoet het logistische regressiemodel. Hoewel de resultaten van het logistische regressiemodel en het probit model in de meeste situaties nauwelijks van elkaar verschillen, vindt men het probit model met name in bio-medische toepassingen op grond van de natuurlijke interpretatie van het model bij bio-assay data. Het extreme waarden model ligt vooral voor de hand bij het modelleren van 'survival' gegevens, zoals geïllustreerd in het voorbeeld naar aanleiding van de gegevens van Bliss (1935) verderop in dit artikel. Voor alle genoemde modellen geldt dat goede schatters voor de parameters kunnen worden verkregen via het maximaliseren van de aannemelijkheidsfunctie van Y.

Een van de methoden hiervoor is het Newton-Raphson algorithme. Dit algorithme is in verschillende computerprogramma's geïmplementeerd, waaronder LOGLINEAR in SPSS/PC+ (Norusis, 1990), BMPD-PR (Dixon, 1990) en GLIM (Payne, 1986).



Figuur 3. De normale verdelingsfunctie, de logistische verdelingsfunctie en de Gumbelverdelingsfunctie. De normale verdeling heeft verwachting 0 en variantie $\pi^2/3$ om de normale verdelingsfunctie te kunnen vergelijken met de logistische verdelingsfunctie (Rothschild & Logothetis, 1985).

Aangetoond kan worden dat maximale aannemelijkheidsschatters asymptotisch normaal zijn verdeeld. Benaderde standaardfouten van de parameters kunnen worden verkregen op basis van Fisher's informatiematrix (= de tweede afgeleiden van de maximale aannemelijkheidsfunctie naar de parameters), zodat op de bekende wijze betrouwbaarheidsintervallen voor de parameters kunen worden opgesteld.

Ter beoordeling van de 'fit' van het model is het begrip 'dimensie' van het model belangrijk. Vaak wordt dit 'losjes' omschreven als 'het aantal onafhankelijke parameters'. Preciezer gezegd is het de rang van de modelmatrix. Het aantal vrijheidsgraden (df) is dan het aantal waarnemingen (N) minus de dimensie van het model. Dit is het aantal dimensies waarin (Y_1, Y_2, ..., Y_N) van het model kan afwijken. We merken hierbij op dat in een discrete situatie N staat voor het aantal cellen in een kruistabel.

De vraag "Hoe goed past het model bij de gegevens?" kan worden beantwoord op basis van de waarde van de ratio van de grootste aannemelijkheid (**L**ikelihood **R**atio) van een model en het verzadigde model (= de data). Voor 'geneste' modellen, dat wil zeggen voor een situatie dat het model M_0 onder H_0 begrepen is in model M_1 onder H_1, geldt dat de ratio van de grootste aannemelijkheid van het model M_0 en het model M_1 een indicatie geeft van het belang van de extra parameters.

(-2 log(LR)) staat bekend onder de term 'deviance' (G^2). Onder milde regulariteitsvoorwaarden is de 'deviance' chi-kwadraat verdeeld met (dim M_1 - dim M_0) vrijheidsgraden. De regulariteitsvoorwaarden worden geschonden indien de predictoren 'echt' continu zijn. In

dat geval heeft in principe iedere waarneming een unieke vector $\mathbf{x}_i$ van scores op de predictorvariabelen, en is LR niet langer bij benadering chi-kwadraat verdeeld. Hoewel de 'fit' van het model in deze situaties niet kan worden beoordeeld op grond van de 'deviance' van het model, blijft het verschil tussen de 'deviances' van twee geneste modellen bij benadering chi-kwadraat verdeeld, waarbij het aantal vrijheidsgraden gelijk is aan het verschil tussen het aantal vrijheidsgraden van de beide modellen. Op deze wijze kan worden getoetst of bepaalde variabelen in het model 'nodig' zijn.

Voorstellen om toch tot een soort 'goodness of fit' maat te komen, zijn gebaseerd op het verschil tussen de LR van het model met alleen de constante en de LR van het model waarin men is geïnteresseerd. McFadden's pseudo R^2 is zo'n maat. Het bezwaar van al deze maten is dat ze afhangen van het aantal waarnemingen. De maximale waarde (=1) wordt aangenomen indien het aantal waarnemingen naar oneindig gaat, hetgeen ze feitelijk ongeschikt maakt als 'goodness of fit' maten. Alternatieven ter verkijging van een indruk van de mate waarin het model past, bestaan uit

(a) het discretiseren van de continue predictorvariabelen om vervolgens het corresponderende loglineaire model op deze tabel te fitten om te zien of dit model redelijk past, en

(b) het rangordenen van de waarnemingen naar de geschatte $\pi(\mathbf{x}_i)$, op basis waarvan de waarnemingen in g even grote groepen worden ingedeeld. Het aantal waarnemingen per groep wordt aangeduid met n_j, j = 1,2,...,g. Een redelijk waarde voor g is 10. Binnen elke groep wordt een onderscheid gemaakt naar de waarde van Y (0 of 1). Er kan vervolgens een 2xg tabel worden gevormd met in de cellen de onder het model verwachte celfrekwenties ($\Sigma\pi(\mathbf{x}_i)$ voor de waarde 1 van Y; (n_j - $\Sigma\pi(\mathbf{x}_i)$) voor de waarde 0 van Y) en de som van de waargenomen celfrekwenties (ΣY_i in de desbetreffende cellen). Pearson's X^2 berekend op deze tabel is bij benadering chi-kwadraat verdeeld bij df = (g-2), en geeft zo eveneens een indruk van de mate waarin het model past (Hosmer & Lemeshow, 1989).

Voor een voorbeeld maken we gebruik van gegevens van Bliss (1935). Deze gegevens hebben betrekking op de gevoeligheid van kevers voor koolstof disulfide (CS_2).

Tabel 1. Gevoeligheid van kevers voor koolstof disulfide.
Gegevens ontleend aan Bliss (1935).

log concentratie CS_2	aantal kevers	aantal dode kevers
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

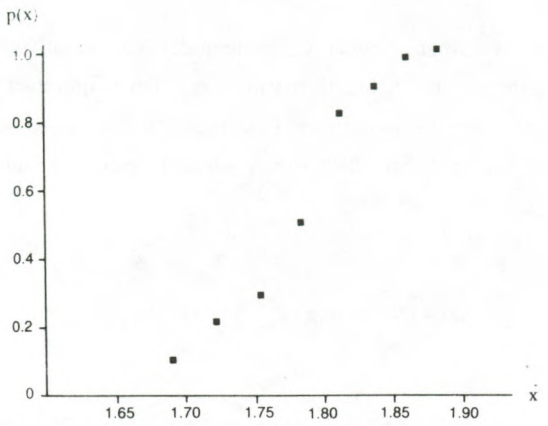
In tabel 2 zijn met behulp van het programma GLIM voor het logistische, het probit en het extreme waarden model de geschatte parameters α en ß en hun standaardfouten, de 'deviance' (G^2) en het aantal vrijheidsgraden (df) weergegeven.

Tabel 2. Geschatte parameters en 'deviances' (G^2) bij toepassing van drie verschillende regressiemodellen op de data van Bliss (1935).

model	logistisch	probit	extreme waarden
α:	-60.72	-34.93	-39.57
s.e.	(5.2)	(2.6)	(3.2)
ß:	34.27	19.73	22.04
s.e.	(2.9)	(1.5)	(1.8)
G^2	11.23	10.12	3.45
df	6	6	6

Zoals op grond van de grote gelijkenis tussen de logistische en de normale verdeling kan worden verwacht, verschillen de 'fit' van het logistische en van het probitmodel nauwelijks.

Aangezien de variantie van de logistische verdeling gelijk is aan $\pi^2/3$ schelen de geschatte parameterwaarden onder het probit model en die onder het logistische model bij benadering een factor $(\pi^2/3)^{1/2}$. Het extreme waarden model past aanzienlijk beter dan het logistische en het probit model. De reden hiervoor is dat bij overlevingsduurgegevens de relatie tussen x en $\pi(x)$ niet symmetrisch is. Dit is duidelijk te zien in een grafiek van de log-concentratie tegen de waargenomen proportie dode kevers.



Figuur 4. Grafiek van de log-concentratie CS_2 tegen de waargenomen proportie dode kevers.

In het navolgende zal worden ingegaan op de situatie dat de Y-variabele meer dan twee waarden kan aannemen.

2.1. Nominale polytome Y-variabele bij discrete predictoren

Voor de situatie dat alle predictoren discreet zijn is de generalisatie naar het logitmodel met een polytome Y-variabele eenvoudig, gezien de relatie van dit model met het loglineaire model. Voor een nominale Y-variabele ligt het model voor de kans op $Y=y_i$ gegeven de predictoren voor de hand. Bij wijze van voorbeeld beschouwen we een IxJ tabel tabel van Y tegen X, waarbij de i-index de Y-variabele indiceert en de j-index de X-variabele.

Teneinde een interessant loglineair model te kunnen beschouwen dat ligt tussen het model van statistische onafhankelijkheid en het verzadigde model, veronderstellen we hier dat de X-variabele geordend is met waarden 1, 2, ... J. Gebruikmakend van de schaalwaarden van X kan het volgende loglineaire model worden gedefinieerd:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + j * \upsilon_{1(i)}. \tag{8}$$

Dit model staat bekend als het rij-effect associatiemodel (Goodman, 1979), waarin u staat voor de 'general mean', $u_{1(i)}$ het hoofdeffect van Y, $u_{2(j)}$ het hoofdeffect van X, en $\upsilon_{1(i)}$ de zogenaamde rij-effect associatie parameter. Onder dit model zijn de $\log(m_{ij}/m_{i+1,j})$ voor $j=1,2, \ldots J$ lineair in j, waarbij het effect van X varieert per combinatie van telkens twee Y-categorieën,

$$\log\left(\frac{m_{ij}}{m_{i+1,j}}\right) = (u_{1(i)} - u_{1(i+1)}) + j * (\upsilon_{1(i)} - \upsilon_{1(i+1)}). \tag{9}$$

Het is eenvoudig te zien dat het rij-effect associatiemodel equivalent is met een model voor π_{ij} (=de kans op $Y=y_i$ gegeven $X=x_i$):

$$\begin{aligned} \pi_{ij} &= \left(\frac{m_{ij}}{m_{+j}}\right) \\ &= \frac{\exp(u+u_{2(j)})\exp(u_{1(i)}+j*\upsilon_{1(i)})}{\exp(u+u_{2(j)})\sum_{t=1}^{I}\exp(u_{1(t)}+j*\upsilon_{1(t)})} \\ &= \frac{\exp(u_{1(i)}+j*\upsilon_{1(i)})}{\sum_{t=1}^{I}\exp(u_{1(t)}+j*\upsilon_{1(t)})} \end{aligned} \tag{10}$$

Om de te schatten parameters te kunnen identificeren, moeten restricties worden opgelegd.

Bijvoorbeeld:

$$u_{1(1)} = v_{1(1)} = 0$$

zodat:

$$\pi_{1j} = \frac{1}{1+\sum_{t=2}^{I} \exp(u_{1(t)} + j * v_{1(t)})} \qquad (11)$$

$$\pi_{ij} = \frac{\exp(u_{1(i)} + j * v_{1(i)})}{1+\sum_{t=2}^{I} \exp(u_{1(t)} + j * v_{1(t)})} \qquad i=2,3,...,I.$$

Uiteraard kunnen meer, geordende zowel als niet geordende, discrete predictoren in het model worden opgenomen. In de volgende paragraaf beschouwen we modellen, waarbij naast discrete predictoren ook continue predictoren voorkomen.

2.2. Nominale polytome Y-variabele bij discrete en continue predictoren

Indien naast discrete predictoren ook continue predictoren voorkomen, is er geen eenvoudige relatie meer met loglineaire modellen. Derhalve kan niet langer gebruik worden gemaakt van de standaard software voor loglineaire analyse. Parameterschattingen voor deze situatie kunnen worden verkregen via maximalisering van de **multinomiale** aannemelijkheid via de Newton-Raphson methode. Verschillende computerprogramma's zijn hiervoor beschikbaar: FREQ (Haberman, 1979), LIMDEP (Greene, 1986) en kalos/c (Rohwer, 1990).

Voor een voorbeeld van multinomiale logistische regressie met de variabele 'beroep' als nominale afhankelijke variabele met vijf categorieën maken we gebruik van gegevens afkomstig uit een internationaal vergelijkend onderzoek: het 'Political Action Survey'[3] van 1973. In dit voorbeeld zijn we met name geïnteresseerd in vragen met betrekking tot de

[3]. De gegevens van het Political Action Survey werden beschikbaar gesteld door het Zentralarchiv für empirische Sozialforschung. De gegevens zijn oorspronkelijk verzameld door onafhankelijke instituten in elk van de deelnemende landen. Noch de oorspronkelijke onderzoeksinstituten noch het Zentralarchiv zijn op enigerlei wijze verantwoordelijk voor de hier gepresenteerde resultaten.

invloed van een aantal variabelen op de kans voor een individu om in een bepaald beroep terecht te komen. We beperken ons tot het effect van opleiding, beschikbaarheid van door de vader op diens kind over te dragen kapitaal, sekse en leeftijd. Daarnaast hebben we nagegaan of er mogelijk sprake is van verschillen tussen landen. Hiertoe hebben we een model gefit, waarbij de parameters van de verschillende predictoren per land kunnen verschillen. De vraag naar het effect van de overdracht van kapitaal is gespecificeerd via twee dummy variabelen: één voor zelfstandige vaders, en één voor vaders met een boerenbedrijf. Voor de kodering van de variabelen, zoals deze in de multinomiale logistische regressie-analyse zijn gebruikt verwijzen we naar tabel 3.

Tabel 3. Gebruikte kodering van de variabelen uit het Political Action Survey 1973.

Land van herkomst	**Huidig beroep respondent**	**Beroep vader respondent**
1. BRD 2. Nederland 3. USA 4. Italië	1: Grote zelfstandigen, hogere leidinggevenden, academici	**referentie categorie** 1: Grote zelfstandigen, hogere leidinggevenden, academici
	2: Lagere leidinggevenden en geschoolde hoofdarbeid	1: Lagere leidinggevenden en geschoolde hoofdarbeid
	2: Routine hoofdarbeid	1: Routine hoofdarbeid
		1: Supervisoren handarbeid, hooggeschoolde handarbeid
		1: Geschoolde handarbeid
		1: Halfgeschoolde en ongeschoolde handarbeid
		1: (Ongeschoolde) landarbeid
Opleiding 0-17 jaar	3: Kleine zelfstandigen met personeel 3: Kleine zelfstandigen zonder personeel	**zelfstandigen** 2: Kleine zelfstandigen met personeel 2: Kleine zelfstandigen zonder personeel
	4: Zelfstandige boeren	
Sekse 1. Vrouw 2. Man	5: Supervisoren handarbeid, hooggeschoolde handarbeid	**boeren** 3: zelfstandige boeren
	5: Geschoolde handarbeid	
	5: Halfgeschoolde en ongeschoolde handarbeid	
	5: (Ongeschoolde) landarbeid	

In de analyse is opleiding als een continue variabelen opgenomen, terwijl de overige variabelen discreet zijn. Het 'beroep van de vader van de respondent' is ontbonden in twee dummy-variabelen ('zelfstandigen' en 'boeren' met de overige categorieën als referentiecategorie). De variabele 'land van herkomst' is in drie dummy-variabelen ontbonden ('NEDERLAND', 'USA' en 'ITALIË' met 'BRD' als referentiecategorie). In de analyse bleek geen van de door ons veronderstelde interacties tussen 'land van herkomst' en andere

in het model opgenomen variabelen van belang. In het uiteindelijke logit model is steeds de eerste beroepscategorie (grote zelfstandigen, hogere leidinggevenden en academici (N=384)) als referentiecategorie gehanteerd. De vier simultane logit vergelijkingen L_i (i= 2, 3, 4, 5) met alleen hoofdeffecten zien er volgt uit:

$$L_i = \beta_0 + \beta_1 NEDERLAND + \beta_2 USA + \beta_3 ITALIE + \beta_4 ZELFSTANDIGEN + \beta_5 BOEREN + \beta_6 OPLEIDING + \beta_7 SEKSE \qquad (i = 2,3,4,5) \tag{12}$$

De via het programma kalos/c (Rohwer, 1990) geschatte parameters zijn in de onderstaande tabel weergegeven.

Tabel 4. Parameterschattingen van het logistische regressiemodel ter voorspelling van het beroep van de respondent. Tussen haakjes de standaardfouten van de parameters. De gegevens zijn afkomstig uit het Political Action Survey 1973.

	Beroep respondent			
Predictor	L2	L3	L4	L5
- constante	5.45	4.59	4.20	7.50
	(.28)	(.36)	(.48)	(.30)
- NEDERLAND	-.38	-.56	-.83	-.49
	(.16)	(.25)	(.31)	(.18)
- USA	.02	.14	-.35	.66
	(.16)	(.25)	(.29)	(.17)
- ITALIE	.93	1.75	.51	.62
	(.23)	(.27)	(.33)	(.24)
- ZELFSTANDIGEN	.12	1.50	-.11	-.72
	(.24)	(.28)	(.76)	(.30)
- BOEREN	.20	.48	3.02	.35
	(.20)	(.25)	(.28)	(.21)
- OPLEIDING	-.22	-.46	-.59	-.58
	(.02)	(.03)	(.04)	(.02)
- SEKSE	-1.76	-.32	.19	-.09
	(.14)	(.18)	(.22)	(.15)
	(N=2487)	(N=324)	(N=198)	(N=1552)

Op basis van tabel 4 valt op te maken dat hoe groter het aantal jaren onderwijs dat men heeft gevolgd, des te groter de kans is om te behoren tot de eerste categorie van beroep

(grote zelfstandigen, hogere leidinggevenden, en academici). De kans om met een groot aantal jaren gevolgd onderwijs te behoren tot de categorieën 4 en 5 van 'beroep van de respondent' is het kleinst. De veronderstelling dat de beschikbaarheid van over te dragen kapitaal bij zelfstandigen en boeren van invloed is op de huidige beroepscategorie van respondenten wordt bevestigd. Respondenten met een 'zelfstandigen' achtergrond hebben een opvallend grote kans om zelf ook te gaan behoren tot de categorie 'zelfstandigen'; respondenten met een 'boeren' achtergrond hebben een opvallend grote kans om zelf ook te gaan behoren tot de categorie 'boeren'. Hetgeen verder opvalt is de grote voorspelde kans voor vrouwen op een beroep in de categorie 'routine hoofdarbeid'.

De 'fit' van het model kan nu niet beoordeeld worden op grond van de waarde van de 'deviance', aangezien de 'deviance' niet langer bij benadering chi-kwadraat is verdeeld. De vraag of een model een goede beschrijving is van de gegevens kan bij benadering worden beantwoord door de continue predictoren te discretiseren en vervolgens het overeenkomstige **loglineaire** model te fitten. Verschillen tussen de 'deviances' van geneste logistische regressiemodellen zijn **wel** bij benadering chi-kwadraat verdeeld, zodat het belang van de predictoren en eventuele interacties tussen predictoren kan worden vastgesteld. Een R^2-achtige maat voor de 'fit' van het model kan worden verkregen door de 'log-aannemelijkheid' van het model met alleen de constante (l_{const}) te vergelijken met de 'log-aannemelijkheid' van het gefitte model (l_{model}):

$$\textit{McFadden's pseudo } R^2 = \frac{l_{const} - l_{model}}{l_{const}} = \frac{-6009.29 + 4736.01}{-6009.29} = .21 \qquad (13)$$

Een bezwaar tegen McFadden's pseudo R^2 is, dat indien N groot wordt, de waarde van de maat naar 1 gaat (Agresti, 1990:111). Alternatieven ter beoordeling van de mate waarin het model past zijn besproken in paragraaf 2. Het verschil is alleen dat we nu een multinomiaal verdeelde Y-variabele hebben. Het eerste alternatief, het discretiseren van continue predictoren om vervolgens het corresponderende loglineaire model te fitten, blijft zonder meer bruikbaar. In het hierboven gegeven voorbeeld is de variabele 'opleiding' in drie categorieën gediscretiseerd: (0 t/m 7 jaren opleiding; 8 t/m 12 jaren opleiding, en 13 en

meer jaren opleiding). Het corresponderende loglineaire model, waarbij alle interacties tussen het huidig beroep van de respondent en de overige variabelen zijn opgenomen, heeft een 'deviance' van 320 bij 252 vrijheidsgraden. Dit doet vermoeden dat het multinomiale logistische model goed is gespecificeerd. Het opnemen van interacties tussen het huidig beroep van de respondent en een of meer predictoren, zoals het land van herkomst en de opleiding is dus niet nodig.

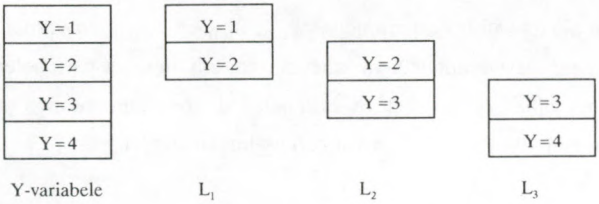
Niet duidelijk is hoe het tweede alternatief, het ordenen van de waarnemingen naar de waarde van de geschatte kansen, gegeneraliseerd kan worden naar een situatie met een multinomiaal verdeelde Y-variabele.

3. Geordende polytome Y-variabelen

Uitgaande van een multinomiaal verdeelde Y-variabele willen we modellen beschouwen die rekening houden met een ordening van de Y-categorieën. We onderscheiden hier drie typen modellen, waarbij het onderscheid tussen de modellen vooral gelegen is in de wijze waarop de te modelleren log-odds (L_i) worden gedefinieerd.

3.1. Het log-odds model voor naastgelegen categorieën

Het model voor de log-odds voor naastgelegen categorieën van Y kan het best geïllustreerd worden aan de hand van figuur 5.



Figuur 5. De log-odds (L_i) in het logitmodel voor naastgelegen categorieën.

Voor een Y-variabele met categorieën 1, 2, ..., I zijn de log-odds (L_i) in het logitmodel voor naastgelegen categorieën als volgt gedefinieerd:

$$L_i = \log\left(\frac{\pi_i(x)}{\pi_{i+1}(x)}\right) \qquad i = 1,...,(I-1). \tag{14}$$

Beschouwen we nu het loglineaire model van uniforme associatie voor de situatie dat naast de Y-variabele er één discrete X-variabele is met respectievelijke schaalwaarden x_1, x_2, ..., x_J:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \beta(i * x_j). \tag{15}$$

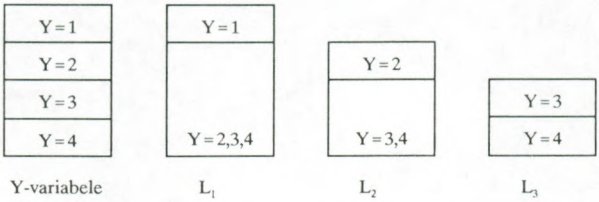
In termen van de log-odds L_i correspondeert het model van uniforme associatie met een model waarbij de relatie tussen alle L_i en X voorgesteld kan worden door rechte lijnen met dezelfde richtingscoëfficiënt ß (=de uniforme associatieparameter) en verschillende constante termen α_i ($= u_{1(i)} - u_{1(i+1)}$). Uit (15) volgt:

$$\begin{aligned} L_i &= \log(m_{ij}) - \log(m_{i+1,j}) \\ &= \alpha_i - \beta x_j \end{aligned} \tag{16}$$

Grafisch kan de relatie tussen L_i en X worden weergegeven via drie evenwijdige lijnen voor respectievelijk L_1, L_2 en L_3. Het hier beschreven model kan worden uitgebreid door bijvoorbeeld te veronderstellen dat voor L_1, L_2 en L_3 de relatie tussen L_i en X niet via evenwijdige lijnen kan worden voorgesteld. Zulks is te modelleren via het zogenaamde rij-effect associatiemodel met in plaats van ß de parameters $ß_{1(1)}$, $ß_{1(2)}$ en $ß_{1(3)}$. Een verdere uitbreiding is uiteraard naar meer verklarende X-variabelen, waarbij deze X-variabelen zowel discreet als continu kunnen zijn. Voorzover de X-variabelen discreet zijn, kunnen de corresponderende logitmodellen zijn alle gefit worden via een loglineair model.

3.2. Het 'continuation-ratio' logitmodel

Behalve log-odds van naastgelegen categorieën zijn andere definities van log-odds gebruikelijk. Beschouwen we steeds de gebeurtenis $Y = y_i$ gegeven dat $Y \geq y_i$, dan kunnen logits L_i worden gevormd op de wijze zoals afgebeeld in figuur 6. Dit logitmodel wordt het 'continuation-ratio' logitmodel genoemd.



Figuur 6. De log-odds (L_i) in het 'continuation-ratio' logitmodel.

Definiëren we de $\tau_i = P(Y_i = y_i | Y_i \geq y_i)$, dan is de logit L_i in termen van deze conditonele kansen τ_i als volgt gedefinieerd:

$$L_i = \log\left(\frac{\tau_i}{1-\tau_i}\right) = \log\left(\frac{\pi_i(x)}{\pi_{i+1}(x)+\pi_{i+2}(x)+\ldots+\pi_I(x)}\right) \qquad i = 1,2,\ldots,(I-1). \qquad (17)$$

Het is eenvoudig aan te tonen dat de multinomiale kansdichtheid van de variabele Y geschreven kan worden als het product van de binomiale kansdichtheden behorende bij de dichotomieën waarop L_1, L_2 en L_3 zijn gebaseerd. Op grond van deze factorisering van de multinomiale kansdichtheid maakt het niet uit of we L_1, L_2 en L_3 simultaan modelleren dan wel afzonderlijk. In het geval dat de log-odds afzonderlijk worden gemodelleerd, komt dat neer op drie aparte logit analyses waarbij de Y-variabele telkens wordt gedichotomiseerd zoals aangegeven in figuur 6.

Indien we de log-odds simultaan willen modelleren, dienen we die delen uit de datamatrix

in figuur 7 die corresponderen met de in figuur 6 gegeven dichotomiseringen samen te voegen.

Y	X_1	X_2	...	X_k
1	5	21		7
1	3	17		5
.	.	.		.
1	2	9		1
2	2	15		4
2	3	7		6
.	.	.		.
2	4	13		2
3	2	6		1
3	3	12		4
.	.	.		.
3	5	19		6
4	2	8		5
4	5	11		4
.	.	.		.
4	3	14		2

Figuur 7. Datamatrix waarbij de waarnemingen geordend zijn op de Y-scores.

De opeenvolgende submatrices worden geïndiceerd door een variabele T (figuur 8). De variabele T wordt vervolgens ontbonden in (I-1) dummy-variabelen, die in de vorm van interacties met de predictoren in het logitmodel worden opgenomen. In het gegeven voorbeeld heeft T de waarden 1, 2 en 3, zodat twee dummy-variabelen T_1 en T_2 dienen te worden gemaakt. Bij slechts één verklarende variabele X ziet het simultane logitmodel er dan als volgt uit:

$$L = \beta_0+\beta_1 T_1+\beta_2 T_2+\beta_3 X+\beta_4(T_1*X)+\beta_5(T_2*X) \qquad (18)$$

L	X_1	X_2	...	X_k	T
1	5	21		7	1
1	3	17		5	1
.	.	.		.	1
1	2	9		1	1
0	2	15		4	1
.	.	.		.	1
0	3	14		2	1
1	2	15		4	2
1	3	7		6	2
.	.	.		.	2
1	4	13		2	2
0	2	6		1	2
.	.	.		.	2
0	3	14		2	2
1	2	6		1	3
1	3	12		4	3
.	.	.		.	3
1	5	19		6	3
0	2	8		5	3
.	.	.		.	3
0	3	14		2	3

Figuur 8. Geëxpandeerde datamatrix voor logit-analyse van het 'continuation-ratio' logitmodel. De verschillende submatrices worden geïndexeerd door de variabele T.

Ter illustratie geven we een voorbeeld op basis van Canadese gegevens afkomstig uit het 'Social Change in Canada Project' van het Institute for Behavioural Research van York University. De Y-variabele heeft betrekking op de arbeidsmarktparticipatie van de vrouw (niet; deeltijd; voltijd). Als predictoren zijn gehanteerd: het inkomen van de echtgenoot INKOMEN (bruto jaarinkomen/1000) en de aanwezigheid van kinderen in het gezin: KIND (0 = geen kinderen; 1 = wel kinderen). De dummy-variabele T indiceert het arbeidsmarktgedrag van de vrouw: T = 0 (werken versus niet werken) en T = 1 (voltijd werk versus deeltijd werk).

Het simultane logitmodel voor deze situatie ziet er als volgt uit:

$$L = \beta_0+\beta_1 T+\beta_2 INKOMEN+\beta_3 KIND+\beta_4(T*INKOMEN)+\beta_5(T*KIND) \qquad (19)$$

Voor T = 0, respectievelijk T = 1 ziet de modelvergelijking er als volgt uit:

$$\begin{aligned} T = 0 &: L = \beta_0+\beta_2 INKOMEN+\beta_3 KIND \\ T = 1 &: L = (\beta_0+\beta_1)+(\beta_2+\beta_4)INKOMEN+(\beta_3+\beta_5)KIND \end{aligned} \qquad (20)$$

De parameterschattingen voor het effect van het inkomen van de echtgenoot en het effect van de aanwezigheid van kinderen voor respectievelijk T = 0 en T = 1 zijn weergegeven in tabel 5.

Tabel 5. Parameterschattingen voor het 'continuation-ratio' logitmodel.
Tussen haakjes de t-waarden van de parameters.

Variabele	T = 0 werken versus niet werken	T = 1 voltijd versus deeltijd
INKOMEN	-.04	-.11
	(-2.14)	(-2.74)
KIND	-1.58	-2.65
	(-5.39)	(-4.90)
constante	1.34	3.48

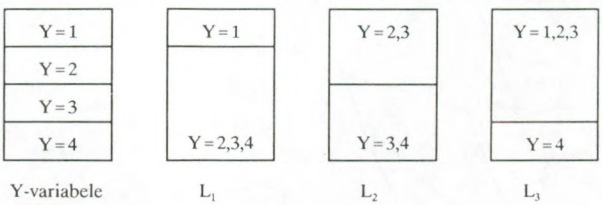
In het algemeen geldt dat hoe hoger het inkomen van de echtgenoot is, des te geringer de bereidheid is om te gaan werken respectievelijk om in voltijd te gaan werken. De aanwezigheid van kinderen in het gezin doet de bereidheid om te gaan werken respectievelijk om in voltijd te gaan werken eveneens afnemen. Zowel het inkomen van de echtgenoot als de aanwezigheid van kinderen in het gezin hebben een groter effect op de beslissing om vanuit deeltijd voltijd te gaan werken dan op de beslissing om vanuit de situatie van niet-werken te gaan werken.

3.3. Het cumulatieve logitmodel

Tenslotte bespreken we het model voor cumulatieve logits. We definiëren:

$F_i(x) = \pi_1(x) + \pi_2(x) + ... + \pi_i(x)$ met $i = 1, 2, ..., I$.

De bij het cumulatieve logitmodel behorende dichotomiseringen zijn weergegeven in figuur 9.



Figuur 9. De log-odds (L_i) in het cumulatieve logitmodel.

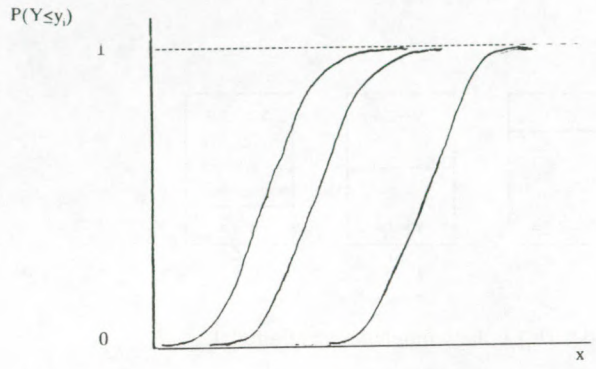
De log-odds voor het cumulatieve logitmodel kunnen als volgt worden gedefinieerd:

$$L_i = logit(F_i(x)) = \log\left(\frac{F_i(x)}{1-F_i(x)}\right) \qquad i = 1,2,...,(I-1). \tag{21}$$

Beschouwen we nu een model voor de cumulatieve logits in termen van één continue verklarende X-variabele: $L_i = \alpha_i + \beta X$. Het verschil tussen L_i voor $X = x_1$ en L_i voor $X = x_2$ is gelijk aan de logaritme van wat de cumulatieve odds-ratio wordt genoemd. Het verschil tussen $L_i(x_1)$ en $L_i(x_2)$ is gelijk aan:

$$\begin{aligned} L_i(x_1) - L_i(x_2) &= \log\left(\frac{P(Y\le i|x_1)/P(Y>i|x_1)}{P(Y\le i|x_2)/P(Y>i|x_2)}\right) \\ &= \beta(x_1 - x_2). \end{aligned} \tag{22}$$

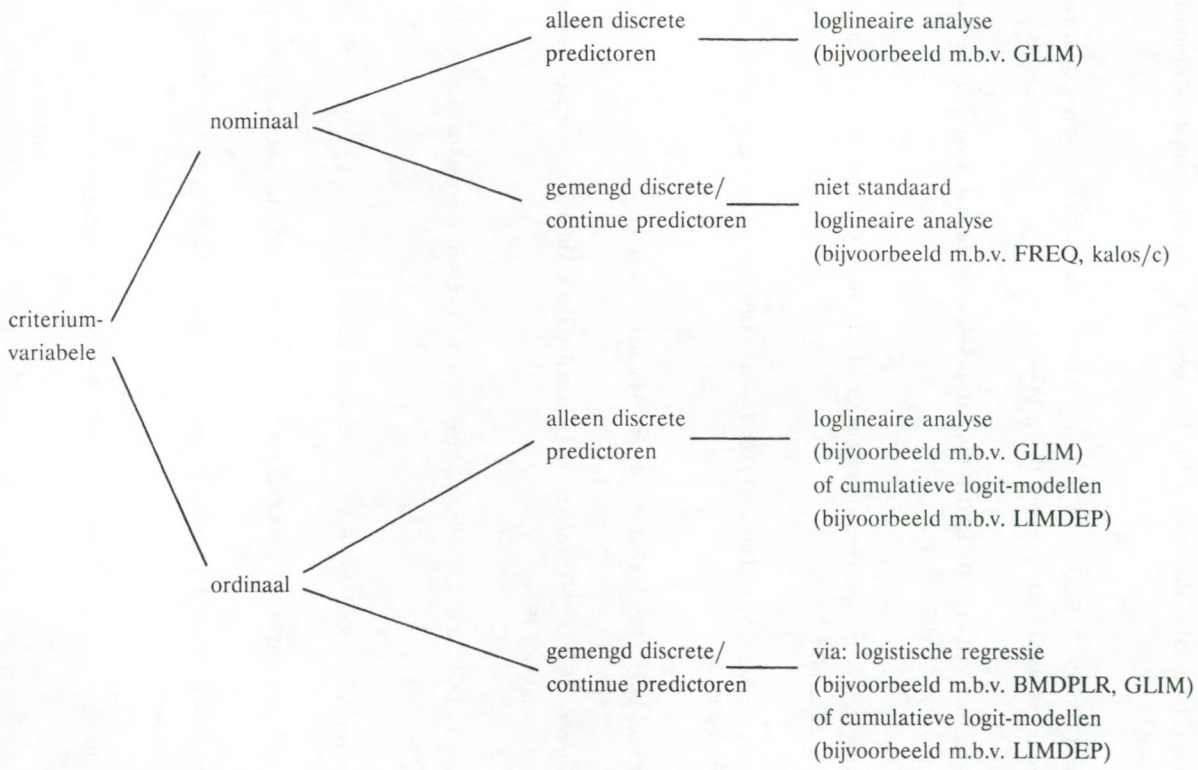
Met andere woorden de logaritme van de cumulatieve odds-ratio is evenredig met het verschil tussen x_1 en x_2. Vanwege deze eigenschap wordt het model ook wel het proportional odds model genoemd. De odds van een response $Y \leq y_i$ voor $X = x_1$ is gelijk aan $\exp(\beta(x_1-x_2))$ vermenigvuldigd met de odds van een respons $Y \leq y_i$ voor $X = x_2$.
De relatie tussen X en $P(Y \leq y_i)$ voor $i=1,2,...I$ zijn alle gelijkvorming en kunnen via horizontale verschuiving uit elkaar worden verkregen (zie figuur 10).



Figuur 10. De relatie tussen X en $P(Y \leq y_i)$ voor het proportional odds model.

Behalve het cumulatieve logitmodel is er ook nog het model met de cumulatieve normale dichtheidsfunctie en het model met de cumulatieve extreme waardenverdeling. Dit laatste model is het bekende proportional hazards model van Cox (Cox, 1972).
Geen van de cumulatieve logitmodellen is met standaard software te fitten. Het programma LIMDEP (Greene, 1986) stelt onderzoekers in staat deze modellen te fitten.

OVERZICHT VAN DE IN DIT ARTIKEL BESPROKEN MODELLEN



Literatuur

Agresti, A. (1990). **Categorical Data Analysis**. New York: John Wiley & Sons.

Bliss, C.I. (1935). Calculation of the Dosage-Mortality Curve. **Annals of Applied Biology**, 22, 134-167.

Cox, D.R. (1972). Regression Models and Life Tables (with discussion). **Journal of the Royal Statistical Society. Series B**, 74, 187-220.

Dixon, W.J. (ed.) (1990). **BMDP Statistical Software Manual. Volume 2**. Berkeley: University of California Press.

Fienberg, S.E. (1980). **The Analysis of Cross-Classified Categorical Data**. Cambridge, Mass.: The MIT Press.

Goldstein, M. & W.R. Dillon (1978). **Discrete Discrimination Analysis**. New York: John Wiley & Sons.

Greene, W.H. (1986). **LIMDEP User's Manual**. New York.

Haberman, S.J. (1979). **Analysis of Qualitative Data. Volume 2: New Developments**. New York: Academic Press.

Hosmer, D.W. & S. Lemeshow (1989). **Applied Logistic Regression**. New York: John Wiley & Sons.

Norusis, M.J. (1990). **SPSS/PC+ Advanced Statistics 4.0**. Chicago, Ill.: SPSS, Inc.

Payne, C.D. (ed.) (1985). **The GLIM System Release 3.77 Manual**. Oxford: NAG.

Rohwer, G. (1990). **Programmbeschreibung kalos/c (V2.2)**. Hamburg: Hamburger Institut für Sozialforschung.

Rothschild, V. & N. Logothetis (1985). **Probability Distributions**. New York: John Wiley & Sons.

ontvangen 20-11-1991
geaccepteerd 19- 3-1992