Loglineaire, logit- en logistische modellen: een introductie.

P.I.M. Schmitz [1] en Albert Verbeek [2]

Samenvatting

Op een inleidend niveau wordt een overzicht gegeven van loglineaire, logit- en logistische modellen voor de analyse van categorische en gemengd categorisch-continue gegevens. De overeenkomsten tussen deze modellen worden aangegeven: logit- en logistische modellen kunnen worden opgevat als een bijzonder loglineair model. De verschillen tussen de modellen betreft primair de vraagstelling: bij een loglineair model wordt geen a priori onderscheid gemaakt tussen afhankelijke en verklarende variabelen, bij een logitmodel wel. Zodra minstens één van de verklarende variabelen in een logitmodel continu is noemen wij dit een logistisch model. Maximum likelihood schattingen, goodness-of-fit toetsen en modelselektie worden summier behandeld. Tenslotte wordt een toepassing gegeven van het logistische model op het gebied van de medische diagnostiek.

1. Inleiding.

De afgelopen tien jaar hebben de statistische methoden voor categorische gegevens in de vorm van meerdimensionale tabellen een snelle ontwikkeling doorgemaakt. Vooral over het loglineaire model zijn in korte tijd tal van boeken verschenen waarvan met name het boek van Bishop, Fienberg en Holland (1975) voor een zekere doorbraak heeft gezorgd.

Het logitmodel voor categorische gegevens wordt meestal gepresenteerd in de context van een loglineair model. Daarnaast werd het logitmodel voor gemengd discreet-continue gegevens (meestal logistisch model genoemd) ontwikkeld onder andere door Cox (1970). In dit artikel wordt een introductie van deze modellen gegeven, waarbij de nadruk ligt op de uiteenzetting van de diverse verschillen en overeenkomsten tussen het loglineaire, het logit- en het logistische model. Een toepassing in de medische besluitvorming zal kort worden behandeld. In een volgend artikel zal een overzicht worden gegeven van de computerprogrammatuur voor deze modellen, en zal aandacht worden besteed aan een gegeneraliseerd lineair model (GLM) dat alle hier behandelde modellen en ook de regressie- en variantie-analyse omvat.

2. Vraagstelling.

Een voorbeeld van gegevens in de vorm van een meerdimensionale tabel is de 2x4x4-tabel 1. Deze gegevens zijn afkomstig uit een longitudinaal onderzoek naar coronaire hartziekte (CHD) in Framingham (Dawber, Kannel en Lyell, 1963). De eerste variabele (y) is een binaire variabele die de aanwezigheid (y = 1) of afwezigheid (y = 0) van CHD aangeeft. Er wordt nu het verband be-

[1] Erasmus Universiteit Rotterdam, Instituut voor Biostatistica.

[2] Rijksuniversiteit Utrecht, Sociologisch Instituut en Instituut voor Mathematische Statistiek.

Tabel 1: Gegevens van het Framingham longitudinaal onderzoek naar CHD.

Coronary Heart Disease	Serum Cholesterol (mg/100 cc)	Systolic Blood Pressure (mm Hg) <127	127-146	147-166	167+
Present	<200	2	3	3	4
	200-219	3	2	0	3
	220-259	8	11	6	6
	≥260	7	12	11	11
Absent	<200	117	121	47	22
	200-219	85	98	43	20
	220-259	119	209	68	43
	≥260	67	99	46	33

Bron: Fienberg (1978), blz. 6.

studeerd tussen y en een tweede en derde variabele: x_1 = serum cholesterol gehalte (mg/100 cc, vier klassen) en x_2 = systolische bloeddruk (mm Hg, vier klassen). Een dergelijk verband kan worden bestudeerd met behulp van een loglineair model, waarbij de algemene vraagstelling is: bestaat er een samenhang tussen k categorische variabelen $x_1, x_2, \ldots, x_k$?
Er kan een onderscheid worden gemaakt tussen responsvariabelen (in de betekenis van random variabelen, waarvan de waarden niet a-priori door het experiment of door de steekproefopzet worden bepaald) en faktoren (variabelen met vaste niveau's). Dit onderscheid heeft consequenties voor de statistische analyse. Er kan worden opgemerkt dat soms binnen eenzelfde studie een variabele in de ene situatie als random en in een andere situatie als "vast" kan worden beschouwd. Het onderscheid random-vast moet niet worden verward met het onderscheid continu-discreet.

Nog een andere indelingscriterium is het afhankelijke of verklarende karakter van een variabele. Bij loglineaire modellen wordt dit onderscheid niet gemaakt.
Bij logitmodellen echter wordt juist het verband bestudeerd tussen k verklarende (onafhankelijke) variabelen $x_1, x_2, \ldots, x_k$ en een (categorische, meestal binaire) afhankelijke variabele y. Zodra één of meer van de verklarende variabelen continu zijn (zodat het minder zinvol is de gegevens in tabelvorm te representeren: nu blijft de oorspronkelijke datamatrix gehandhaafd) is het gebruikelijk het logitmodel een logistisch model te noemen.
Zoals hierboven aangegeven maken wij dus een duidelijk onderscheid tussen het begrip responsvariabele en het begrip afhankelijke variabele. In sommige studieopzetten, bijv. in case-control onderzoek (Schmitz, 1980) kunnen logistische modellen worden geformuleerd waarin de afhankelijke variabele y een faktor is en geen responsvariabele. In case-control onderzoek wordt een steekproef van individuen uit een ziektepopulatie (y = 1) en een steekproef van individuen uit een controlepopulatie (y = 0) gekozen. De variabele y wordt hier dus vantevoren vastgelegd. Toch kan nu een logistisch model worden geformuleerd waarin y als afhankelijke variabele voorkomt.

3. Voorselektie van variabelen.

Afhankelijk van de vraagstelling en de aard van de variabelen kan worden besloten tot de keuze van een der genoemde drie modelklassen. Binnen zo'n klasse zal vervolgens naar het best passende model worden gezocht, waarbij een gering aantal parameters, significante effecten en eenvoudige interpretatie pluspunten zijn die een rol spelen bij de selektie van het uiteindelijke model. Maar in feite kan verantwoorde modelselektie, waarin ook interaktietermen worden betrokken, slechts plaatsvinden bij een betrekkelijk gering aantal variabelen. In de praktijk, waar vaak sprake is van een relatief groot aantal variabelen, zal dan ook een zekere voorselektie moeten plaatsvinden. Bij voorkeur dient dit te geschieden met behulp van externe informatie (bijv. gegevens uit de literatuur), maar ook dit is lang niet altijd mogelijk. Als een logistisch model wordt toegepast kan eventueel een stapsgewijze selektieprocedure worden gebruikt, zoals bekend uit de lineaire regressie-analyse voor een normaal verdeelde afhankelijke variabele ("A systematic method to make every possible type I error" - J. Finn). Maar hoe te werk te gaan bij een loglineair model, waarbij de gegevens dus in de vorm van een meerdimensionale tabel worden weergegeven? Bij bijv. 10 binaire variabelen ontstaan $2^{10} = 1024$ cellen, zodat bij een steekproefomvang van bijv. 200 of 300 vele lege cellen ontstaan en de asymptotische methoden voor het loglineaire model niet meer geldig zijn.

Freeman en Jekel (1980) geven een methode voor tabelselektie in deze situatie, waarbij geen onderscheid wordt gemaakt tussen verklarende en afhankelijke variabelen. Deze methode bestaat uit een aantal stappen: in de eerste stap worden de significante paarsgewijze associaties gezocht. Hieruit worden samengestelde variabelen gevormd die in de tweede stap worden onderzocht op significante paarsgewijze associaties met de oorspronkelijke variabelen. Zo ontstaat na een aantal stappen een indruk van de variabelen met de sterkste effecten. Deze (meestal niet meer dan 5 of 6) variabelen worden dan gebruikt voor de vorming van de meerdimensionale tabel die als uitgangspunt dient voor verdere analyse met behulp van een loglineair (of logit-) model.
Hoewel tabelselektie (voorselektie van variabelen) en modelselektie (het vinden van een geschikt model voor deze geselekteerde variabelen) in de strategie van Freeman en Jekel gescheiden begrippen zijn, is deze scheiding niet bij alle selektiemethoden (zie ook par. 4.3) zo strikt aan te brengen.

4. Loglineaire modellen.

Beschouw een driedimensionale tabel (zoals tabel 1) die ontstaat door de waarden van drie categorische variabelen te classificeren. Zij x_{ijk} de celfrequentie voor de ie categorie (i = 1, 2, ..., I) van variabele 1, de jde categorie (j = 1, 2, ..., J) van variabele 2 en de kde categorie (k = 1, 2, ..., K) van variabele 3. Naast deze waargenomen celfrequentie wordt ook de verwachte celfrequentie m_{ijk} beschouwd. Een "gesatureerd" (verzadigd) loglineair model kan worden opgevat als de volgende parametrisatie van deze verwachte frequenties:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

met $\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_k u_{3(k)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = \sum_i u_{13(ik)} = \sum_k u_{13(ik)} = \sum_j u_{23(jk)} = \sum_k u_{23(jk)} = \sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0$

De IJK onderling onafhankelijke parameters m_{ijk} zijn op deze wijze uitgedrukt met behulp van IJK onderling onafhankelijke parameters u. Met behulp van de restricties kunnen omgekeerd de u-termen worden uitgedrukt met behulp van m_{ijk}, bijv.:

$$u = \frac{\sum_i \sum_j \sum_k \log m_{ijk}}{IJK} \; ; \; u_{1(i)} = \frac{\sum_{j,k} \log m_{ijk}}{JK} - u$$

De u-termen kunnen worden geïnterpreteerd als "hoofdeffekten" en "interakties", in zekere zin in analogie met variantie-analyse. Zo is $u_{12(ij)}$ de interaktie tussen variabele 1 en variabele 2, en is $u_{123(ijk)}$ de tweede orde interaktie tussen de drie variabelen. Een voordeel van deze herparametrisatie is dat in de praktijk vaak blijkt dat hogere orde-interakties gelijk nul zijn. Hierdoor is het mogelijk om stabielere schattingen van de verwachte celfrequenties m_{ijk} te verkrijgen. Eveneens geven toetsen omtrent de u-parameters een beeld van de onderzochte samenhang van de studievariabelen.

4.1 Maximum likelihood schattingen.

Voor een specifiek loglineair model, bijv. alle interakties nul: $u_{12} = u_{13} = u_{123} = 0$ (onafhankelijkheidsmodel), worden de maximum likelihood schattingen $\hat{m}_{ijk}$ gezocht.
Hiertoe moet eerst de steekproefverdeling van x_{ijk} bekend zijn. De drie meest voorkomende steekproefverdelingen zijn: Produkt-Poisson, Multinomiaal (totale vaste steekproefomvang N) en Produkt-multinomiaal (gestratificeerde steekproef). Deze geven echter dezelfde M.L.-schattingen (binnen de toegestane modellen bij deze verdelingen). Vooral terwille van een eenvoudige interpretatie is het gebruikelijk de diverse te bestuderen modellen te beperken tot de zgn. klasse van hierarchische modellen. Dit impliceert dat als een zekere u-term nul is, bijv. $u_{12} = 0$, ook alle hogere orde u-termen met 12 onder de indices, nul zijn, dus $u_{123} = 0$. Omgekeerd geldt dat als een u-term ongelijk nul is, bijv. $u_{12} \neq 0$, ook alle lagere orde u-termen met indices 1 of 2 ongelijk nul zijn, dus $u_1 \neq 0$ en $u_2 \neq 0$ (en ook $u \neq 0$). De berekening van de M.L. schattingen $\hat{m}_{ijk}$ voor een hiërarchisch loglineair model kan op efficiënte wijze geschieden met behulp van het zg. "iteratief aanpassingsalgoritme" van Deming-Stephan. Een andere bekende methode voor M.L.-schatting voor loglineaire modellen is Newton-Raphson. Haberman (1978) geeft een vergelijkende beschouwing over het Newton-Raphson algorithme versus het Deming-Stephan algorithme. In een volgend artikel zal hier nader op worden ingegaan.

4.2 Goodness-of-fit toetsen.

Nadat voor een specifiek loglineair model de ML-schattingen voor de verwachte celfrequenties zijn verkregen kan de aanpassing van de verkregen schattingen aan het waarnemingsmateriaal worden onderzocht met behulp van een van de volgende twee aanpassingstoetsen. Voor gegevens in drie dimensies:

(1) Pearson's X^2:

$$X^2 = \sum \frac{(\text{observed-expected})^2}{\text{expected}} =$$

$$= \sum_i \sum_j \sum_k \frac{\{x_{ijk} - \hat{m}_{ijk}\}^2}{\hat{m}_{ijk}}$$

(2) G^2 - 2 loglikelihood ratio

$$= 2 \sum (\text{observed}) \log \left(\frac{\text{observed}}{\text{expected}}\right)$$

$$= 2 \sum_i \sum_j \sum_k x_{ijk} \log\left\{\frac{x_{ijk}}{\hat{m}_{ijk}}\right\}$$

Onder de nulhypothese (het gespecificeerde loglineaire model is geldig) en bij voldoend grote steekproefomvang hebben beide toetsen een X^2-verdeling met ν vrijheidsgraden:

ν = (aantal cellen met verwachte celfrequenties ongelijk nul) minus (aantal parameters dat kan worden geschat).

De LR-toetsingsgrootheid G^2 wordt vaker dan X^2 gebruikt in verband met een voor de modelselektie handige eigenschap: conditioneel partitioneren van G^2 is mogelijk. Beschouw twee hierarchische modellen zodat model 1 méér u-termen, zeg $\{u_\nu\}$, bevat dan model 2: model 2 $\subset$ model 1.
De LR-toetsingsgrootheid

$G^2(2|1)$ = -2 loglikelihood ratio

$$= 2 \sum (\text{observed}) \log \left(\frac{\text{expected}_1}{\text{expected}_2}\right)$$

voor de hypothese $H_0 : \{u_\nu\} = 0$ gegeven dat model 1 goed past is:

$$G^2(2|1) = G^2(2) - G^2(1)$$

waarbij $G^2(i)$ de LR-toetsingsgrootheid is voor model i (i = 1, 2).
Zo kan voor een keten van vier hierarchische modellen
model 4 $\subset$ model 3 $\subset$ model 2 $\subset$ model 1
worden geschreven:

$$G^2(4) = G^2(4|3) + G^2(3|2) + \underbrace{G^2(2|1) + G^2(1)}_{G^2(2)}$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{G^2(3)}$$

Het op deze wijze opsplitsen van G^2 heet conditioneel partitioneren. Een mogelijke strategie bij de selektie van een van de vier modellen is te beginnen bij model 1 achtereenvolgens te berekenen: $G^2(1)$, $G^2(2|1)$, $G^2(2)$, $G^2(3|2)$, enz. Zodra een van deze G^2-waarden een overschrijdingskans $< \alpha$ geeft wordt gestopt. I.v.m. het herhaald toetsen is het feitelijk toetsingsniveau natuurlijk $> \alpha$, al is de grootte hiervan moeilijk te bepalen. Op deze wijze wordt een passend model met zo weinig mogelijk parameters gevonden (binnen de gekozen vier modellen).

4.3 Modelselektie.

In de praktijk zal meestal meer dan één goed passend loglineair model worden gevonden. Er zal daarom een strategie moeten worden bepaald om tot een selektie te komen. Wij noemden al eerder diverse overwegingen die hierbij een rol spelen: naast een goed passend model wordt een gering aantal parameters, significante "effecten" (u-termen) en een eenvoudige interpretatie van het verkregen model bij de modelkeuze betrokken. Vele strategieën zijn bekend, maar een overzicht hiervan past niet binnen het inleidende karakter van dit artikel.
Verwezen zij hier naar de literatuur: Conditioneel partitioneren (Fienberg, 1978 blz. 48 e.v., Bishop, Fienberg en Holland, 1975 blz. 158 e.v., stapsgewijze selektie (Goodman, 1971), screenen van effekten (Brown, 1976; Benedetti en Brown, 1978) en simultane toetsingsprocedures (Aitkin, 1979; Aitkin, 1980) zijn slechts enkele van de mogelijkheden.
Bij deze strategieën wordt vrijwel steeds gebruik gemaakt van een asymptotische toets voor de nulhypothese $H_0 : \{\mu_v\} = 0$. Soms wordt hiervoor de al in 4.2 genoemde LR-toets gebruikt, in andere situaties wordt de Wald-toets (Rao, 1973) toegepast, met name waar $\{\mu_v\}$ één u-term betreft (de Waldtoets komt dan overeen met de methode der gestandaardiseerde u-termen). In de literatuur over loglineaire modellen wordt echter een derde, door Rao (1973) genoemde asymptotische toets voor $H_0 : \{u_v\} = 0$ niet genoemd: de efficiente score-toets.
Mede omdat door sommige auteurs (Breslow en Day, 1980; Day en Byar, 1979) de gunstige eigenschappen van deze toets, met name bij kleinere aantallen, worden genoemd (asymptotisch zijn de drie toetsen equivalent) verrichten wij hier momenteel verder onderzoek naar.

5. Logitmodellen.

Beschouw een 2xJ-tabel waarvoor het verzadigde loglineaire model kan worden geschreven als:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

$$i = 1, 2; \; j = 1, 2, \ldots, J;$$

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0.$$

Veronderstel nu dat de eerste (binaire) variabele wordt opgevat als een afhankelijke variabele, en de tweede als een verklarende variabele. Verder worden de marginalen $\{x_{+j}\}$ vast verondersteld. De kansen $P_{1(j)}$ en de logits L_j worden gedefinieerd door:

$$P_{1(j)} = \frac{m_{1j}}{m_{+j}}, \quad L_j = \log \frac{P_{1(j)}}{1-P_{1(j)}} = \log \frac{m_{ij}}{m_{2j}}$$

Het logitmodel voor deze tabel kan nu worden geschreven als:

$$L_j = \log \frac{m_{1j}}{m_{2j}} = 2u_{1(1)} + 2u_{12(ij)} = w + w_{2(j)}$$

$$P_{1(j)} = \frac{1}{1 + \exp\{-(w+w_{2(j)})\}}$$

Met behulp van de toets voor $H_0 : u_{12} = 0$ ofwel $H_0 : w_{2(j)} = 0$ kan nu de invloed van de verklarende variabele op de kans $P_{1(j)}$ (dus op de binaire afhankelijke variabele) worden onderzocht. Analoog kan een logitmodel voor een binaire afhankelijke variabele en meer dan één verklarende variabele (bijv. drie) worden geformuleerd:

$$L_{jkl} = w + w_{2(j)} + w_{3(k)} + w_{4(l)} + w_{23(jk)} + w_{24(jl)} + w_{34(kl)} + w_{234(jkl)}.$$

In het algemeen zijn er meerdere loglineaire modellen te geven die in formulevorm tot hetzelfde logitmodel leiden, maar wel andere schattingen voor de w-termen geven. De verschillen houden verband met de mogelijkheid de verklarende variabelen als vast of als random te beschouwen. Gebruikelijk is alle verklarende variabelen als vast te beschouwen, maar noodzakelijk is dit (meestal) niet.
De interpretatie van een logitmodel heeft veel overeenkomsten met die van multiple regressie voor een normaal verdeelde afhankelijke variabele. Behalve dat de interpretatie veelal eenvoudiger is dan bij een loglineair model, kunnen nu ook continue verklarende variabelen worden toegelaten. De parametrisatie en een aangepaste wijze van maximum likelihood schattingen voor deze logitmodellen worden in par. 6 behandeld.
Een logitmodel voor categorische verklarende variabelen is niets anders dan een bijzonder loglineair model. Alle schattings- en toetsingsmethoden voor het loglineaire model zijn dan ook bruikbaar voor zo'n logitmodel.

6. Logistische modellen.

6.1 Het logistische model als logitmodel met continue verklarende variabelen.

Tot nu toe zijn we ervan uitgegaan dat de categorische variabelen in het loglineaire of het logit-model geen geordende categorieën bezitten. Beschouw echter nu een verzadigd loglineair model voor drie variabelen, waarbij de eerste variabele binair is, de tweede polytoom met J geordende categorieën en de derde polytoom met K geordende categorieën. Veronderstel dat aan de categorieën van de tweede en derde variabele a priori scores kunnen worden toegekend, zeg resp. $\{v_j^{(2)}\}$ en $\{v_k^{(3)}\}$. Een parametrisatie die de ordening verwerkt is:

$$u_{12(ij)} = \{v_j^{(2)} - \bar{v}^{(2)}\}\ u_{1(i)}^{(2)}$$

$$u_{13(ik)} = \{v_k^{(3)} - \bar{v}^{(3)}\}\ u_{1(i)}^{(3)}$$

terwijl de hiërarchische veronderstelling leidt tot de interaktieterm:

$$u_{123(ijk)} = \{v_j^{(2)} - \bar{v}^{(2)}\}\{v_k^{(3)} - \bar{v}^{(3)}\}u_{1(i)}^{(23)}$$

N.B.: $\bar{v}^{(2)} = \sum_j v_j^{(2)}/J$, $\bar{v}^{(3)} = \sum_k v_k^{(3)}/K$.

Met behulp van deze parametrisatie kan het logitmodel voor de binaire afhankelijke variabele worden geschreven als:

$$L_{jk} = \log \frac{m_{1jk}}{m_{2jk}} = \beta_o + \beta_2 v_j^{(2)} + \beta_3 v_k^{(3)} + \beta_{23} v_j^{(2)} v_k^{(3)}$$

waarbij de β-parameters kunnen worden geschreven als een funktie van de u- en v-parameters. Stel vervolgens dat de tweede en derde variabele continu zijn en elk N (totale steekproefomvang) verschillende waarden kunnen aannemen. Deze (geordende) waarden kunnen tevens als "score" worden gehanteerd. In feite ontstaat zo een 2xNxN-tabel met in iedere cel een 1 of een 0.
Het spreekt vanzelf dat de eerder genoemde iteratieve berekening van de M.L.-schattingen voor m_{ijk} (Deming-Stephan, aangepast voor geordende categorieën) op deze wijze uiterst inefficient is, nog afgezien van de (on)-geldigheid van de asymptotische theorie. Het is daarom dan ook gebruikelijk het logitmodel voor verklarende variabelen die ook continu kunnen zijn (dus ook logistisch model genoemd) op een iets andere wijze te formuleren.

6.2 Het algemene logistische model.

Veronderstel dat N onderling onafhankelijke waarnemingen aan een binaire afhankelijke variabele y en k verklarende variabelen z_1, z_2, ..., z_k zijn verricht. De variabelen $z_1, \ldots, z_k$ zijn binair, polytoom geordend of continu. De binaire variabele y kan de waarden 0 of 1 aannemen.

$$E(y_i) = Pr\{y_i = 1\} = p_i,\ i = 1, 2, \ldots, N$$

Het algemene logistische model kan nu worden geschreven als:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_o + \sum_{j=1}^{k} \beta_j z_{ij}$$

waarbij z_{ij} de ide waarneming (i = 1, 2, ..., N) is aan variabele z_j. (j = 1, 2, ..., k). Een andere schrijfwijze met weglating van de index i is:

$$Pr\{y = 1\} = \frac{1}{1 + \exp\{-(\beta_o + \sum_j \beta_j z_j)\}}$$

Op eenvoudige wijze kan nu de likelihood

$$\prod_{i=1}^{N} Pr\{y_i = y_i\}\ ,\ y_i = 0 \text{ of } 1$$

worden opgesteld. Differentiëren naar de β-parameters van de loglikelihood en gelijk stellen aan nul geeft een stelsel niet-lineaire vergelijkingen voor de M.L.-schattingen $\hat{\beta}_j$. Dit stelsel vergelijkingen kan worden opgelost met behulp van het algorithme van Newton-Raphson. Als bij-produkt van deze procedure worden tevens de standard-errors van β_j geleverd, zodat een asymptotisch betrouwbaarheidsinterval voor deze coëfficiënten kan worden opgesteld.
Helaas is voor het logistische model geen goodness-of-fit toets beschikbaar, al bestaan er wel enkele empirische methoden om de aanpassing van het model te onderzoeken (Dirschedl, 1980; Tsiatis, 1980; Stephens, 1979). De selektie van een logistisch model kan plaats vinden met behulp van een LR-toets voor het vergelijken van twee modellen (ofwel: het toetsen van H_0 : $\{\beta_v\} = 0$, waarbij $\{\beta_v\} = 0$ staat voor een deelverzameling van de β-parameters).

7.Een toepassing op het gebied van de medische diagnostiek.

Ter illustratie van de behandelde modellen volgt hier een toepassing van het logistische model op het gebied van de medische diagnostiek. Deze toepassing betreft een deel van het patientenmateriaal zoals behandeld in het proefschrift van V.d. Does en Lubsen (1978). In deze IMIR-studie (Imminent Myocardial Infarction Rotterdam) zijn gegevens verzameld van patienten die hun huisarts bezochten met symptomen verdacht voor cardiale oorsprong. Een aantal faktoren die significant bleken samen te hangen met de aanwezigheid van een acuut hartinfarct zijn geregistreerd. De specifieke details omtrent diverse medische aspecten in dit onderzoek zullen hier niet worden besproken. In tabel 2 is een overzicht gegeven van de variabelen die wij hier zullen beschouwen. De vraagstelling is: hoe groot is de diagnostische kans op de aanwezigheid van een acuut myocardinfarct voor een individu dat zich met genoemde klachten bij de huisarts heeft gemeld, en waarvoor de gegevens x_1 t/m x_{12} zijn vastgesteld.

Tabel 2. Studievariabelen in de IMIR-studie.

y = AMI(Acuut MyocardInfarct) ja(=1) of nee(=0).
Er zijn n_1 = 92 cases en n_2 = 1211 non-cases.
x_1 = sex (0-1)
x_2 = age (in jaren)
x_3 = chestpain < 48 h (0-1)
x_4 = duration chestpain > 30 min. (0-1)
x_5 = stabbing chestpain (0-1)
x_6 = chestpain primary symptom (0-1)
x_7 = cold clammy skin (0-1)
x_8 = heart rate (beats/min)
x_9 = premature beats > 3/min (0-1)
x_{10} = systolic minus diastolic blood pressure (mmHg)
x_{11} = syst.bl.pr. < 110 mmHg (0-1)
x_{12} = chestpain indicator (0-1)

Dit probleem kan precieser worden geformuleerd in termen van een discriminantanalyse. De klassieke Fisher-lineaire discriminantanalyse kan hier ook (eventueel met behulp van een transformatie naar normaliteit) worden toegepast, maar deze aanpak zal hier niet worden beschouwd. Voor de kans op een hartinfarct:

$$Pr\{y = 1\}$$

werd een logistisch model volgens par.6.2 opgesteld, met de 3 continue en 9 binaire verklarende variabelen x_1 t/m x_{12} volgens tabel 2.

$$Pr\{y = 1|x_1, x_2, \ldots, x_{12}\} =$$

$$= \frac{1}{1 + \exp\{-(\beta_0 + \sum_{i=1}^{12} \beta_i x_i)\}}$$

De maximum likelihood schattingen $\hat{\beta}_i$, hun standard errors $SE(\hat{\beta}_i)$ en de standaard-normaal verdeelde grootheden $\beta_i/SE(\hat{\beta}_i)$ zijn berekend met behulp van een programma van Lee (1974) dat is gebaseerd op het Newton-Raphson algorithme. De resultaten zijn vermeld in tabel 3.

Tabel 3. Maximum likelihood-schattingen voor de IMIR-gegevens.

	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}/SE(\hat{\beta})$
x_1	0.51	0.26	1.94
x_2	0.06	0.0098	5.60
x_3	1.59	0.34	4.70
x_4	1.08	0.39	2.77
x_5	-0.87	0.43	-2.01
x_6	1.17	0.38	3.07
x_7	1.61	0.44	3.65
x_8	0.02	0.0066	3.03
x_9	0.62	0.37	1.67
x_{10}	-0.02	0.0073	-2.22
x_{11}	0.80	0.44	1.80
x_{12}	0.25	0.37	0.66
CONST	-8.87		

Met behulp van de aldus verkregen kansen (op te vatten als a posterior kansen in de context van discriminantanalyse) kan een classificatieregel worden opgesteld:
Beslis tot infarct indien

$$Pr\{y = 1\} \geq c.$$

Afhankelijk van de gekozen grenswaarde c kan de sensitiviteit en de specificiteit van deze beslissingsregel worden geevalueerd. Deze aanpak met behulp van een logistisch model is slechts een van de mogelijke benaderingen voor dergelijke gegevens. Een vergelijking van discriminant-analyse methoden voor gemengd discreet-continue (verklarende) variabelen is op dit ogenblik nog onderwerp van nadere studie.

Dankwoord

Wij willen graag de Heer J. Wijnen bedanken voor zijn waardevolle commentaar op een eerste concept van dit artikel en Dr.J. Lubsen voor het bereidwillig afstaan van de IMIR-gegevens.

Literatuur

Aitkin,M.,A simultaneous test procedure for contingency table models, Appl.Statist.,28,nr.3(1979),p.233-242.
Aitkin,M.,A note on the selection of loglinear models,Biometrics, 36(1980),p.173-178.
Benedetti,J.K. en M.B.Brown,Strategies for the selection of loglinear models,Biometrics,34(1978),p.680-686.
Bishop,Y.M.M.,S.E.Fienberg en P.W.Holland,Discrete multivariate analysis: theory and practice,MIT-press,Cambridge(1975).
Breslow,N.E. en N.E.Day,Statistical methods in cancer research,vol.I, The analysis of case-control studies,IARC Scientific Publications no.32, International Agency for Research on Cancer, Lyon (1980).
Brown,M.B.,Screening effects in multi-dimensional contingency tables, Appl.Statist.,25,nr.1(1976),p.37-46.
Cox,D.R.,The analysis of binary data,Methuen,London(1970).
Dawber,T.R.,W.B.Kannel en L.P.Lyell, An approach to longitudinal studies in a community: the Framingham study,Ann.N.Y.Acad,Sci.,107(1963),p.539-556.
Day,N.E. en D.P.Byar,Testing hypothesis in case-control studies--Equivalence of Mantel-Haenszel statistics and logit score tests,Biometrics,35(1979), p.623-630.
Dirschedl,P.,Praktischer erfahrungen mit dem multiplen logistischen modell, Medizinische Informatik und Statistik,Band 17: Biometrie-heute und morgen, Herausgeber:S.Koller,P.L.Reichertz und K.Uberla,Springer Verlag,Berlin(1980), p.278-300.
Does,E.van der en J.Lubsen,The Imminent Myocardial Infarction Rotterdam study, proefschrift Erasmus Universiteit Rotterdam(1978).
Fienberg,S.E.,The analysis of cross-classified categorical data,MIT-press, Cambridge(1978).
Freeman,D.H. en J.F.Jekel, Table selection and loglinear models, J.Chron. Dis.,33(1980),p.513-524.
Goodman,L.A.,The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications,Technometrics,13(1971),p.33-61.
Haberman,S.J.,Analysis of qualitative data.I en II.,Academic Press, New York(1978).
Lee,E.T.,A computer program for linear logistic regression analysis, Comp.progr. in biomed.,4(1974),p.80-92.
Rao,C.R.,Linear statistical inference and its applications, John Wiley, New York(1973).
Schmitz,P.I.M.,Statistische methoden voor case-control onderzoek, Technical Report,Institute of Biostatistics,Erasmus University Rotterdam (april 1980).
Stephens,M.A.,Tests of fit for the logistic distribution based on the empirical distribution function,Biometrika,66,nr.3(1979),p.591-595.
Tsiatis,A.A.,A note on a goodness-of-fit test for the logistic regression model,Biometrika,67,nr.1(1980),p.250-251.