DISCRIMINANTANALYSE EN INSOLVENTIEVOORSPELLINGEN

Prof. Dr. A. P. J. Abrahamse en Drs. R. A. I. van Frederikslust

Voorwoord

Discriminantanalyse is een statistische techniek die objecten in groepen klassificeert aan de hand van bepaalde informatie. Verschillende schrijvers hebben de techniek gebruikt om ondernemingen te klassificeren in solvente en insolvente ondernemingen. Discriminantanalyse begint een bekende klank te krijgen in de bedrijfseconomie. Men mag verwachten dat deze tendens zich zal voortzetten omdat de belangstelling voor het probleem van de beoordeling van de kredietwaardigheid van ondernemingen toeneemt als gevolg van de economische teruggang.

In verband hiermee ontwikkelt zich een behoefte aan een min of meer volledige uiteenzetting van de techniek en zijn achtergrond (het model dat aan de techniek ten grondslag ligt). Het doel van dit artikel is om in deze behoefte te voorzien. De bestaande overzichten zijn in het algemeen ofwel te onvolledig ofwel te moeilijk toegankelijk voor mensen die geen statistici van huis uit zijn.

De auteurs van dit artikel presenteren hierbij het resultaat van een poging om a) volledig genoeg te zijn om inzicht te geven in de reikwijdte en de beperkingen van de techniek en b) een uiteenzetting te geven in *elementaire* begrippen. Zij koesteren niet de illusie dat het artikel eenvoudig is. Wel hebben zij vermeden gebruik te maken van statistische theorie die geen gemeengoed is en niet strikt noodzakelijk is voor mensen die discriminantanalyse met begrip willen toepassen.

Wel is gebruik gemaakt van matrixalgebra, vooral omdat daardoor gemakkelijk aangetoond kan worden dat de denktrant van het kleinste-kwadratenregressiemodel voor een groot deel kan worden gevolgd.

In het tweede deel van het artikel is het gebruik van de discriminantanalyse volledig geïllustreerd met behulp van een probleem uit de financieringstheorie, te weten de voorspelbaarheid van ernstige betalingsmoeilijkheden (insolventie).

1 Inleiding

Laat S een verzameling objecten zijn, opgedeeld in J uitputtende en elkaar uitsluitende deelgroepen S_j, $j = 1,2,\ldots,J$; dus $S_1 \cup S_2 \cup \ldots \cup S_J = S$ en $S_j \cap S_{j'} = \phi$ voor $j \neq j'$. We geven t-de object van groep S_j aan met s_{jt}; dus $s_{jt} \in S_j$. Verder geven we het aantal elementen in groep S_j aan met N_j en het aantal elementen in S noemen we N, dus

$$N = \sum_{j=1}^{J} N_j.$$

Samengevat:

Tabel 1

Groep	Objecten	Aantal
S_1	$s_{11}\ s_{12}\ \cdots s_{1N_1}$	N_1
S_2	$s_{21}\ s_{22}\ \cdots s_{2N_2}$	N_2
..	.....	..
S_J	$s_{J1}s_{J2}\ \cdots s_{JN_J}$	N_J
S		N

Stel dat de N objecten uit S niet van elkaar te onderscheiden zijn en dat we ze volgens de één of andere regel klassificeren in de J groepen. Telkens trekken we dus een element uit S en klassificeren het in één van de J klassen S_j. Er zullen dan goede en foute klassificaties plaatsvinden en de ene klassificatieregel zal meer en de andere minder foute klassificaties (misklassificaties) opleveren.

Voorbeeld: Beschouw een urn met 20 rode en 40 witte ballen. R is de groep rode en W is de groep witte ballen. De ballen zijn uiterlijk niet van elkaar te onderscheiden, bijvoorbeeld doordat ze alle omwikkeld zijn met papier. We kunnen ons voorstellen dat we de ballen over de groepen R en W klassificeren in de verhouding 10 : 50 door willekeurig een bal uit de urn te trekken, vervolgens met een dobbelsteen te werpen en de bal als Rood te klassificeren wanneer een 6 boven komt en als Wit wanneer dit niet gebeurt.

De kans dat een getrokken bal rood is bedraagt 20/60 en de kans dat hij wit is bedraagt 40/60. De kans dat een getrokken bal als Rood wordt geklassificeerd is 10/60 en de kans dat hij als Wit wordt geklassificeerd is 50/60. De kans dat een getrokken bal rood is en als Rood wordt geklassificeerd is dan

20/60 × 10/60 = 1/18;
de kans dat hij rood is en als Wit wordt geklassificeerd is
20/60 × 50/60 = 5/18;
de kans dat hij wit is en als Wit wordt geklassificeerd is
40/60 × 50/60 = 10/18;
de kans dat hij wit is en als Rood wordt geklassificeerd is
40/60 × 10/60 = 2/18.
Deze kansen zijn samengevoegd in de volgende tabel.

Tabel 2

		Urn		
		R	W	
Beslissing	R	1/18	2/18	1/6
	W	5/18	10/18	5/6
		1/3	2/3	1

In de twee hoofddiagonale cellen staan de kansen op correcte klassificaties; hun som is 11/18. Als we de ballen met gelijke kansen over R en W klassificeren krijgen we de volgende tabel.

Tabel 3

		Urn		
		R	W	
Beslissing	R	1/6	2/6	1/2
	W	1/6	2/6	1/2
		1/3	2/3	1

De som van de kansen op correcte klassificaties is nu 1/2. In dit opzicht is de vorige klassificatieregel dus superieur. De som van beide kansen is maximaal als alle ballen in de grootste groep worden geklassificeerd. Dan geldt de volgende tabel.

Tabel 4

		Urn		
		R	W	
Beslissing	R	0	0	0
	W	1/3	2/3	1
		1/3	2/3	1

met een kans op correcte klassificatie van 2/3.

Als de enige informatie waarover men beschikt de omvang van de subgroepen R en W is en als men een zo groot mogelijk aantal correcte klassificaties zou nastreven dan zou men het beste alle ballen in de grootste groep kunnen klassificeren.

Als men daarentegen de kleuren van de getrokken ballen zou kennen m.a.w. als men over volledige informatie zou beschikken, dan zou men altijd juist klassificeren; de kans op juiste klassificatie zou 1 zijn. In ons ballenvoorbeeld is het verschil tussen deze laatste kans (1) en de kans (2/3) - bij klassificatie in de grootste groep - terug te voeren tot een verschil in beschikbare informatie, in dit geval tussen volkomen informatie en geen informatie (afgezien van de omvang van de groepen).

In praktische gevallen bevindt men zich zelden in een dergelijke extreme situatie; men heeft noch volledige, noch totaal ontbrekende informatie. In ons ballenvoorbeeld betekent dit dat we van een getrokken bal een aantal kenmerken kennen, zoals het gewicht, de diameter, de mate van rondheid, etc. Wij geven in hetgeen volgt zulke kenmerken aan met $x_1, x_2, \ldots, x_p$. Als bijvoorbeeld de rode ballen gemiddeld iets zwaarder zijn dan de witte, is het in beginsel mogelijk een kans op correcte klassificatie groter dan 2/3 te krijgen door van deze informatie gebruik te maken. We kunnen tussen de ballen discrimineren op grond van waarnemingen betreffende de kenmerken $x_1, x_2, \ldots, x_p$. Het is niet zonder meer duidelijk wat de juiste wijze is om dit soort informatie te verwerken. Discriminantanalyse is een techniek die een klassificatieregel geeft, gebaseerd op een optimaal gebruik van de informatie betreffende de x_i. In de eerstvolgende paragrafen wordt deze techniek behandeld voor het geval van twee groepen, dus voor $J = 2$. Een generalisatie naar meer dan twee groepen volgt in Paragraaf 5.

Discriminantanalyse is gebaseerd op een bepaald model dat geacht wordt de waarnemingen gegenereerd te hebben. Een van de mogelijke modellen voor twee groepen is als volgt.

Basismodel voor twee groepen

Beschouw twee groepen objecten S_1 en S_2 en stel dat de p kenmerken $x_1, x_2, \ldots, x_p$ in beide groepen simultaan normaal verdeeld zijn met een p-element vector μ als gemiddelde en met de p x p-covariantiematrix Σ.

De kansdichtheidsfunctie van $x = (x_1, x_2, \ldots, x_p)'$ is dan

$$(1.1) \quad l(x;\mu) = c \exp\{-\tfrac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)\} \quad c = \text{constant}$$

De groepen verschillen slechts in het gemiddelde: voor groep S_1 geldt $\mu = \mu_1$ en voor groep S_2 geldt $\mu = \mu_2$.

In Paragraaf 2 bezien we het geval van bekende parameters μ en Σ. In Paragraaf 3 behandelen we het geval waarin deze parameters moeten worden geschat uit een steekproef.

2 Het geval met bekende parameters

We geven om te beginnen een klassificatieprocedure die berust op het aannemelijkheidsprincipe.

Een redelijke klassificatieprocedure

De aannemelijkheidsfunctie van de kansvariabele x is de simultane dichtheidsfunctie van x. Hij is gedefinieerd in (1.1) en hij wordt opgevat als een functie van de vector μ van gemiddelden.

Voor elke waarde van μ geeft hij de aannemelijkheid dat de kansvariabele de specifieke steekproefwaarde x aanneemt. Stel dat de waarde van μ bekend is en bijv. gelijk aan μ_0, dan is de meest aannemelijke uitkomst die waarde x waarvoor $1(x;\mu_0)$ maximaal is.

Laat de p-element vector x de attributen weergeven van een getrokken object met onbekende herkomst (groep S_1 of groep S_2).

Voor beide mogelijke waarden μ_1 en μ_2 kunnen we de aannemelijkheid van de getrokken steekproef $x = (x_1, x_2, \ldots, x_p)'$ uitrekenen; ze zijn resp. $1(x;\mu_1)$ en $1(x;\mu_2)$.

Als $1(x;\mu_1) > 1(x;\mu_2)$ is de aannemelijkheid van x groter voor μ_1 dan voor μ_2. Een redelijke klassificatiebeslissing is dan om het object waarop x betrekking heeft als tot groep S_1 behorend te bestempelen. Omgekeerd zouden we het element in groep S_2 kunnen klassificeren ingeval $1(x;\mu_1) < 1(x;\mu_2)$[1]).

Deze klassificatieregel kan als volgt worden weergegeven: een exemplaar met attribuutwaarden x klassificeren we in groep S_1 (met $\mu = \mu_1$) als de aannemelijkheidsverhouding

$$(2.1) \quad \lambda = \frac{1\,(x;\mu_1)}{1\,(x;\mu_2)}$$

groter is dan 1 en in groep S_2 (met $\mu = \mu_2$) als $\lambda < 1$.
Schematisch:

$$(2.2) \quad \lambda \begin{cases} >1 : \text{object} \rightarrow S_1 \\ <1 : \text{object} \rightarrow S_2 \end{cases}$$

Vereenvoudiging van λ

We kunnen de aannemelijkheidsverhouding λ als volgt vereenvoudigen.
Substitutie van (1.1) in λ geeft:

$$\lambda = \exp\{-\tfrac{1}{2}[(x-\mu_1)'\Sigma^{-1}(x-\mu_1) - (x-\mu_2)'\Sigma^{-1}(x-\mu_2)]\}$$
$$= \exp\{-\tfrac{1}{2}[(x'\Sigma^{-1}x - x'\Sigma^{-1}\mu_1 - \mu_1'\Sigma^{-1}x + \mu_1'\Sigma^{-1}\mu_1) -$$
$$(x'\Sigma^{-1}x - x'\Sigma^{-1}\mu_2 - \mu_2'\Sigma^{-1}x + \mu_2'\Sigma^{-1}\mu_2)]\}$$

[1]) In het geval dat $\lambda = 1$, is men vrij in zijn keuze. De keuzeregel zou dus ook kunnen luiden in termen van $\geq$ en $\leq$ i.p.v. $>$ en $<$. De kans op een waarde 1 voor λ is echter 0.

Door o.a. herrangschikking van termen en door gebruik te maken van

$$x' \Sigma^{-1} \mu_i = \mu_i' \Sigma^{-1} x \qquad (i = 1,2)$$

wordt dit:

$$\lambda = \exp \{ -\tfrac{1}{2} [-2x'\Sigma^{-1} \mu_1 + 2x' \Sigma^{-1} \mu_2 + \mu_1' \Sigma^{-1} \mu_1 - \mu_2' \Sigma^{-1} \mu_2] \}$$
$$= \exp \{ x'\Sigma^{-1} (\mu_1 - \mu_2) - \tfrac{1}{2} (\mu_1 + \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \}$$

In de laatste uitdrukking is de eerste term in de exponent lineair in x en is de tweede term in de exponent een constante (onafhankelijk van x). Daarom schrijven we λ als

(2.3) $\lambda = \exp \{x' b - b_0\}$ met $b = \Sigma^{-1} (\mu_1 - \mu_2)$ en
$b_0 = \frac{1}{2} (\mu_1 + \mu_2)'b$

Klassificatie aan de hand van de log-aannemelijkheidsverhouding
Als we in (2.3) aan weerszijden van het gelijkteken de natuurlijke logarithme nemen krijgen we

(2.4) $\log \lambda = x' b - b_0$

Omdat x normaal verdeeld is en omdat $\log \lambda$ lineair is in x, is $\log \lambda$ eveneens normaal verdeeld. Zijn verwachting is

(2.5) $E(\log \lambda) = E(x' b - b_0) = \mu' b - b_0$

omdat $E(x) = \mu$, zie (1.1). Zijn variantie is

(2.6) $\text{var} (\log \lambda) = \text{var} (x' b - b_0) = \text{var} (x'b)$
$= \text{var} (b'x)$
$= E (b' x - b' \mu)^2$
$= E [b' (x - \mu) (x - \mu)'b]$
$= b'E [(x - \mu) (x - \mu)']b$
$= b' \Sigma b$

waar $\Sigma = E [(x-\mu) (x-\mu)']$ de covariantiematrix van x voorstelt, zie (1.1). De klassificatieprocedure (2.2) wordt in termen van $\log \lambda$,

(2.7) $\log \lambda \begin{cases} > 0 : \text{object} \rightarrow S_1 \\ < 0 : \text{object} \rightarrow S_2 \end{cases}$

of, door substitutie van (2.4),

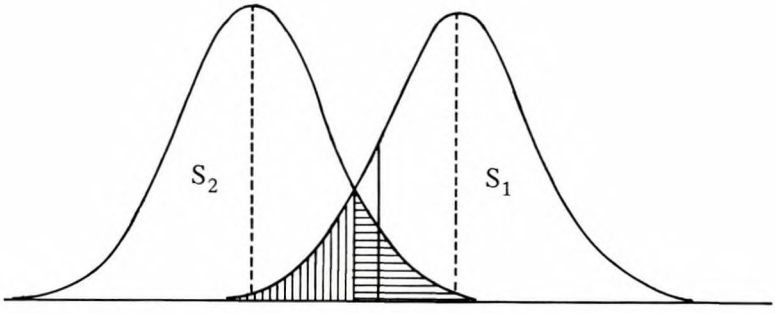
(2.8) $x'b \begin{cases} > b_0 : \text{object} \to S_1 \\ < b_0 : \text{object} \to S_2 \end{cases}$

waar b en b_0 gedefinieerd zijn in (2.3).
De scalair $y = x'b$ wordt *discriminantfunctie* genoemd.

De dichtheidsfuncties van $\log \lambda$ voor resp. $\mu = \mu_2$ en $\mu = \mu_1$ zijn in Figuur 1 geschetst.

Figuur 1

Dichtheden van $\log \lambda$ $(\mu_1 > \mu_2)$



$\mu_2'b - b_0$ o $\mu_1'b - b_0$ $\log \lambda = x'b - b_0$

Dat het punt midden tussen de gemiddelden $\mu_2'\, b - b_0$ en $\mu_1'\, b - b_0$ precies de klassificatiegrens 0 uit regel (2.7) is, volgt uit

$$\tfrac{1}{2}\,(\mu_2' b - b_0 + \mu_1' b - b_0) = \tfrac{1}{2}(\mu_1 + \mu_2)' b - b_0$$

waar

$$b_0 = \tfrac{1}{2}\,(\mu_1 + \mu_2)' b$$

volgens (2.3).

De gearceerde gebieden geven de kansen op foute klassificaties weer, het gebied links van 0 voor onterecht in groep S_2 en het gebied rechts van 0 voor onterecht in groep S_1 klassificeren. Deze kansen zijn resp.

$$(2.9)\quad P[\log \lambda < 0/\mu = \mu_1] = P\left[\frac{\log \lambda - E(\log \lambda)}{\sqrt{\text{var}(\log \lambda)}} < \frac{-\mu_1' b + b_0}{\sqrt{b'\,\Sigma b}}\right]$$

$$= N\left[\frac{-\mu_1' b + b_0}{\sqrt{b'\,\Sigma b}}\right]$$

m a b blz. 238

$$P[\log \lambda > 0/\mu = \mu_2] = 1 - P[\log \lambda < 0/\mu = \mu_2]$$

$$= 1 - N\left[\frac{-\mu_2' b + b_0}{\sqrt{b' \Sigma b}}\right]$$

waar N [x] de standaard normale verdelingsfunctie voorstelt (de kans op een waarde kleiner dan - of gelijk aan - x).

Als we voor b en b_0 de uitdrukkingen gedefinieerd in (2.3) substitueren, worden de kansen op foute klassificaties:

$$(2.10) \quad P[\log \lambda < 0/\mu = \mu_1] = N[-\sqrt{+\tfrac{1}{4}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)}] =$$

$$= N[-\tfrac{1}{2}\delta]$$

$$P[\log \lambda > 0/\mu = \mu_2] = 1 - N[\sqrt{\tfrac{1}{4}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)}]$$

$$= N[-\sqrt{+\tfrac{1}{4}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)}]$$

$$= N[-\tfrac{1}{2}\delta]$$

waar $\delta^2 \equiv (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$.

We zien dat beide kansen gelijk zijn en dat ze kleiner zijn naarmate μ_1 en μ_2 verder uiteen liggen.

Weging van de beide soorten fouten[2])

Zojuist constateerden we dat de kansen op de twee soorten klassificatiefouten even groot zijn bij de besproken klassificatieprocedure. We krijgen verschillende kansen als we de klassificatiegrens in Figuur 1 naar rechts of naar links verschuiven. Verschuiving naar rechts betekent dat de kans op onterechte klassificatie in groep S_2 die op onterechte klassificatie in groep S_1 gaat overtreffen; verschuiving naar links heeft het omgekeerde tot gevolg.

De keuze voor de aanvankelijk beschreven klassificatieprocedure impliceert in zekere zin gelijke weging van de twee soorten fouten. Het is duidelijk dat men dit laatste lang niet altijd juist acht; in de meeste gevallen vindt men de ene soort fout ernstiger dan de andere.

Laten we aannemen dat de kosten van onterecht een object in S_1 klassificeren C_{12} bedragen en dat de kosten van onterecht een object in S_2 klassificeren C_{21} zijn en dat correcte klassificatie geen kosten meebrengt. We zullen nu de klassificatie gaan bezien als een beslissingstheoretisch probleem om licht te werpen op de rol van C_{12} en C_{21} in de klassificatieprocedure.

---

[2]) Kennis van dit onderdeel is niet noodzakelijk voor het kunnen volgen van wat in de rest van dit artikel is geschreven.

Als we een willekeurig object uit $S = S_1 \cup S_2$ trekken is de kans $[P(\mu_1)]$ dat het tot S_1 behoort (waar geldt $\mu = \mu_1$) gelijk aan de proportie objecten van groep S_1 ten opzichte van het totaal aantal exemplaren:

$$(2.11) \quad P(\mu_1) = \frac{N_1}{N}$$

De kans dat een willekeurig getrokken object tot groep S_2 behoort is analoog:

$$(2.12) \quad P(\mu_2) = \frac{N_2}{N}$$

De kansen $P(\mu_1)$ en $P(\mu_2)$ worden *a priori kansen* genoemd; zij geven onze a priori opinies weer. Als we nu waarnemingen x aan het getrokken exemplaar gaan verrichten, wijzigen in het algemeen onze opinies; de a priori kansen gaan over in zgn. *a posteriori kansen:* $P(\mu_1/x)$ en $P(\mu_2/x)$. De a posteriori kansen zijn via de informatie x gerelateerd aan de a priori kansen. Al deze kansen moeten een coherent geheel vormen en voldoen aan de basiswetten van de kansrekening. Coherentie is gewaarborgd als we de a priori en a posteriori kansen relateren volgens het Theorema van Bayes, dat luidt:

$$(2.13) \quad P(\mu_i/x) = 1(x;\mu_i)\, P(\mu_i)/P(x) \qquad (i = 1,2)$$

waarvan $1(x;\mu_i)$ de aannemelijkheidsfunctie is en $P(x)$ de dichtheidsfunctie van x.

We willen de steekproefruimte van x partitioneren in twee uitputtende en elkaar uitsluitende gebieden R_1 en R_2, zodat we een object met waarneming $x = x_0$ klassificeren in S_1 als $x_0 \epsilon R_1$ en in S_2 als $x_0 \epsilon R_2$.

Als we de voorwaardelijke kans op het onterecht in S_1 klassificeren aangeven met $P(1/2)$ en die op het onterecht in S_2 klassificeren met $P(2/1)$, dan zijn de verwachte kosten van de klassificatieregel:

$$r = P(\mu_1)\, P(2/1) C_{21} + P(\mu_2)\, P(1/2) C_{12}$$

waar

$$P(2/1) = \int_{R_2} 1(x;\mu_1)\, dx \text{ en } P(1/2) = \int_{R_1} 1(x;\mu_2)\, dx$$

Als we $C_{21} = C_{12} = 1$ stellen is r gelijk aan de kans op misklassificatie $P(\mu_1)P(2/1) + P(\mu_2)P(1/2)$.

Het minimeren kan worden opgevat als het klassificeren van elke mogelijke waarneming x in R_1 of in R_2 zodanig dat de kans op misklassificatie $P(\mu_1)P(2/1) + P(\mu_2)P(1/2)$ minimaal is.

De laatstgenoemde kans kan worden geschreven als

m a b blz. 240

$$P(\mu_1) \int_{R_2} 1(x;\mu_1)dx + P(\mu_2) \int_{R_1} 1(x;\mu_2)\,dx$$

waaruit blijkt dat de bijdrage van klassificatie in R_1 van een bepaalde waarneming x_o gelijk is aan $P(\mu_2)\ 1\ (x_o;\mu_2)$; als x_o in R_2 wordt geklassificeerd is de bijdrage $P(\mu_1)\ 1\ (x_o;\mu_1)$. Het criterium waarmee R_1 en R_2 worden samengesteld is dus het quotiënt

$$\frac{P(\mu_2)1(x;\mu_2)}{P(\mu_1)1(x;\mu_1)}$$

dat krachtens Bayes' theorema gelijk is aan

$$\frac{P(\mu_2/x)P(x)}{P(\mu_1/x)P(x)} = \frac{P(\mu_2/x)}{P(\mu_1/x)}$$

Als een bepaald object waarneming x_o geeft, klassificeren we het object in S_1 als $P(\mu_1/x_o) > P(\mu_2/x_o)$ en in S_2 als $P(\mu_1/x) < P(\mu_2/x)$, m.a.w. we klassificeren het object in de meest waarschijnlijke groep gegeven x_0. R_1 en R_2 zijn dus als volgt gedefinieerd:

$$R_1 : P(\mu_1/x) > P(\mu_2/x)$$
$$R_2 : P(\mu_1/x) < P(\mu_2/x)$$

en de klassificatieregel luidt:

$$\frac{P(\mu_1/x)}{P(\mu_2/x)} \begin{cases} > 1 : \text{object} \rightarrow S_1 \\ < 1 : \text{object} \rightarrow S_2 \end{cases}$$

Als C_{12} en C_{21} willekeurig zijn wordt de procedure:

$$(2.14) \quad \frac{P(\mu_1/x)C_{21}}{P(\mu_2/x)C_{12}} \begin{cases} > 1 : \text{object} \rightarrow S_1 \\ < 1 : \text{object} \rightarrow S_2 \end{cases}$$

Substitutie van (2.13) geeft:

$$(2.15) \quad \frac{1(x;\mu_1)\ P(\mu_1)C_{21}}{1(x;\mu_2)\ P(\mu_2)C_{12}} \begin{cases} > 1 : \text{object} \rightarrow S_1 \\ < 1 : \text{object} \rightarrow S_2 \end{cases}$$

en we zien direct dat klassificatieregel (2.15) equivalent is met de regel (2.2) als de a priori kansen (en dus de groepen) even groot zijn: $(P(\mu_1) = P(\mu_2))$, en als tevens de fouten even zwaar worden gewogen: $C_{21} = C_{12}$.

Regel (2.15) kan ook worden geschreven als:

$$(2.16)\quad \lambda = \frac{1(x;\mu_1)}{1(x;\mu_2)} \begin{cases} > \dfrac{P(\mu_2)}{P(\mu_1)}\dfrac{C_{12}}{C_{21}} & : \text{object} \rightarrow S_1 \\ < \dfrac{P(\mu_2)}{P(\mu_1)}\dfrac{C_{12}}{C_{21}} & : \text{object} \rightarrow S_2 \end{cases}$$

Uit deze uitdrukking concluderen we dat de klassificatiegrens in Figuur 1 meer naar rechts schuift naarmate $P(\mu_2)/P(\mu_1)$ (de relatieve omvang van groep S_2) groter is en naarmate de relatieve kosten C_{12}/C_{21}, van onterecht in groep S_1 klassificeren, hoger zijn. Dit is in overeenstemming met de intuïtie omdat het naar rechts schuiven van de grens betekent dat minder in groep S_1 wordt geklassificeerd en meer in groep S_2. Merk op dat men, om regel (2.16) te kunnen gebruiken a priori-uitspraken moet doen omtrent de verhouding tussen de grootten van de groepen en omtrent de weging der beide mogelijke soorten fouten.

3 De discriminantfunctie ingeval μ en Σ geschat moeten worden

In de praktijk kennen we in 't algemeen de gemiddelden μ_1 en μ_2 en de covariantiematrix Σ niet, maar beschikken we hoogstens over een steekproef van waarnemingen betreffende de variabelen $x_1, x_2, \ldots, x_p$ voor een aantal, zeg N, objecten. Stel dat er van deze N objecten N_1 tot groep S_1 behoren en N_2 tot groep S_2 ($N_1 + N_2 = N$) en laten we de waarneming van kenmerk i betreffende het t-de object van groep S_j aangeven met x_{ijt} ($i = 1,2,\ldots,p$; $j = 1,2$; $t = 1, 2, \ldots, N_j$) en laten we vervolgens alle waarnemingen in de volgende (Nxp)-matrix weergeven:

$$(3.1)\quad X = \begin{bmatrix} x_{111} & x_{211} & \cdots & x_{p11} \\ x_{112} & x_{212} & \cdots & x_{p12} \\ \vdots & \vdots & & \vdots \\ x_{11N_1} & x_{21N_1} & \cdots & x_{p1N_1} \\ \hline x_{121} & x_{221} & \cdots & x_{p21} \\ x_{122} & x_{222} & \cdots & x_{p22} \\ \vdots & \vdots & & \vdots \\ x_{12N_2} & x_{22N_2} & \cdots & x_{p2N_2} \end{bmatrix} = \begin{bmatrix} x'_{11} \\ x'_{12} \\ \vdots \\ x'_{1N_1} \\ \hline x'_{21} \\ x'_{22} \\ \vdots \\ x'_{2N_2} \end{bmatrix} = \begin{bmatrix} X_1 \\ \hline X_2 \end{bmatrix}$$

Het ligt nu voor de hand om uit X schatters $\hat{\mu}_1$, $\hat{\mu}_2$ en $\hat{\Sigma}$ te berekenen en de klassificatieregel (2.8) te gebruiken met vervanging van b en b_0 door de schatters

m a b blz. 242

$$\hat{b} = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$$

(3.2)

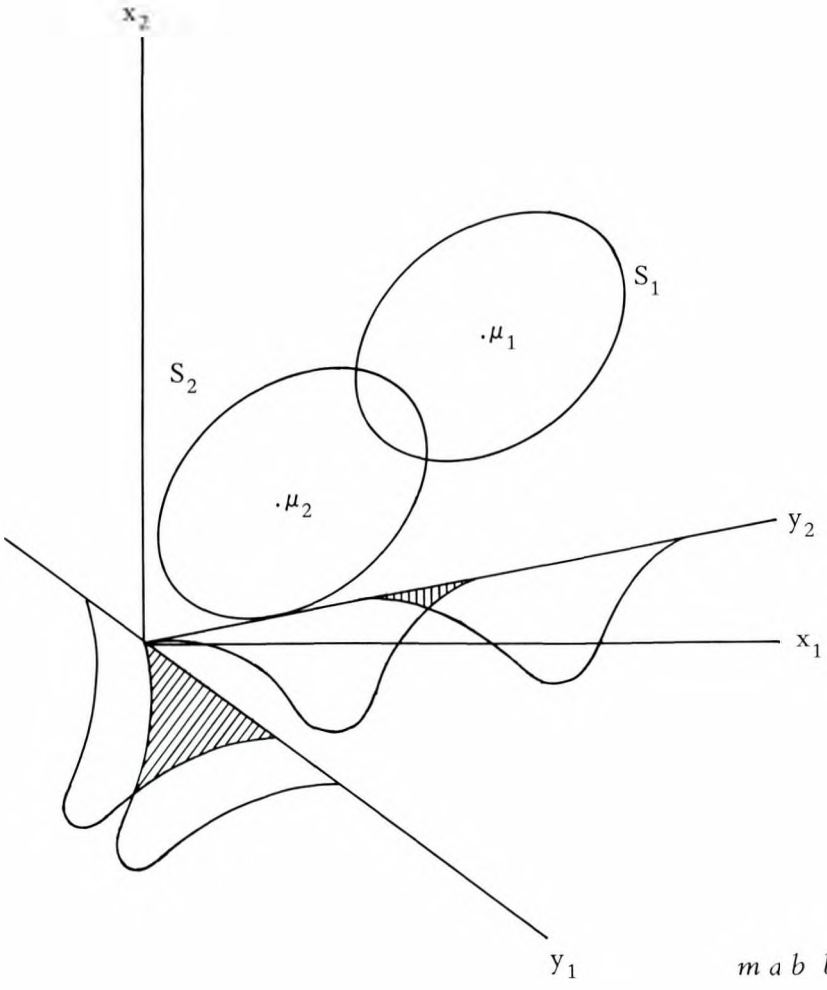
$$\hat{b}_0 = \tfrac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)b$$

In deze paragraaf laten we zien dat deze procedure, die klassificeert volgens de geschatte lineaire discriminantfunctie $y = x'\hat{b}$, optimaal is, in die zin dat $\hat{b}$ de waarde van b is die de verdeling van y voor groep S_1 en de verdeling van y voor groep S_2 zover mogelijk uit elkaar legt of, anders gezegd, die de overlapping van beide verdelingen zo klein mogelijk maakt.

Het is van belang op te merken dat deze klassificatieprocedure niet meer uitgaat van de veronderstelling van normaal verdeelde x_i. Er wordt eenvoudigweg geklassificeerd op grond van een lineaire functie die van alle mogelijke lineaire functies de kleinste overlapping van verdelingen geeft.

We illustreren het probleem met onderstaande figuur voor het geval dat $p = 2$.

Figuur 2



m a b blz. 243

De twee ellipsen zijn contourlijnen van de bivariate verdelingen, resp. voor groep S_1 en voor groep S_2; y_1 en y_2 zijn twee verschillende lineaire discriminantfuncties $y = x'b$. De ellipsen kunnen worden beschouwd als grondvlakken van twee „bergen" waarschijnlijkheidsmassa. Projectie van die „bergen" op het vlak dat door y gaat en loodrecht op het $x_1 x_2$ - vlak staat geeft de verdeling over y voor elk van beide groepen S_1 en S_2. We zien direct dat de verdelingen over y_2 elkaar minder overlappen dan die over y_1, d.w.z. dat y_2 beter discrimineert dan y_1. Omdat μ en Σ onbekend zijn kennen we de verdelingen over y niet, maar voor een gegeven $b = (b_1\ b_2)'$ hebben we een waarde $y_{jt} = b_1 x_{1jt} + b_2 x_{2jt}$ voor elk object, dus voor $j = 1,2$ en $t = 1,2,\ldots,N_j$. De N_1 waarden y_{1t} vormen de empirische y-verdeling voor S_1 en de N_2 waarden y_{2t} vormen de empirische y-verdeling voor S_2.

Een redelijk doel lijkt nu om b zó te kiezen dat de empirische y-verdelingen elkaar minimaal overlappen. Merk op dat de mate van overlapping geringer is naarmate de gemiddelden van de empirische y-verdelingen verder uiteenliggen en naarmate de variantie binnen die verdelingen kleiner is. We zouden nu b zó kunnen kiezen dat het quotiënt met als teller het gekwadrateerde verschil tussen de beide gemiddelden en met als noemer de variantie binnen de verdelingen, maximaal is.

We zullen nu een uitdrukking voor een doelstellingsfunctie formuleren. In de empirische y-verdelingen

$(y_{11}, y_{12}, \ldots, y_{1N_1})$ voor groep S_1
$(y_{21}, y_{22}, \ldots, y_{2N_2})$ voor groep S_2 $\quad (N = N_1 + N_2)$

geven we het rekenkundig gemiddelde voor S_1 aan met $\bar{y}_1$, dat voor S_2 met $\bar{y}_2$ en het totale gemiddelde met $\bar{y}_{..}$.

Er geldt dan

$$(3.3)\quad \sum_{j=1}^{2}\sum_{t=1}^{N_j}(y_{jt} - \bar{y}_{..})^2 = \sum_{j=1}^{2}\sum_{t=1}^{N_j}[(y_{jt} - \bar{y}_{j.}) + (\bar{y}_{j.} - \bar{y}_{..})]^2$$

$$= \sum_{j=1}^{2}\sum_{t=1}^{N_j}(y_{jt} - \bar{y}_{j.})^2 + \sum_{j=1}^{2}\sum_{t=1}^{N_j}(\bar{y}_{j.} - \bar{y}_{..})^2$$

omdat het kruiselingse product nul is:

$$\sum_j \sum_t (y_{jt} - \bar{y}_{j.})(\bar{y}_{j.} - \bar{y}_{..}) = \sum_j (\bar{y}_{j.} - \bar{y}_{..}) \sum_t (y_{jt} - \bar{y}_{j.})$$

$$= \sum_j (\bar{y}_{j.} - \bar{y}_{..})(N_j\bar{y}_{j.} - N_j\bar{y}_{j.}) = 0$$

Het linkerlid van (3.3) is de totale variatie binnen de y_{jt}, de eerste term in het rechterlid is de „binnengroeps-variatie" en de tweede term in het rechterlid is de „tussengroeps-variatie"[3]). In de volgende pargaraaf wordt bewezen dat de „tussengroepsvariatie" op een constante na gelijk is aan het gekwadrateerde verschil tussen de twee groepsgemiddelden:

[3]) We spreken over *variaties* en niet over *varianties* omdat we de sommen niet gedeeld hebben door de aantallen waarnemingen.

$$(3.4)\quad \sum_j \sum_t (\bar{y}_{j.} - \bar{y}_{..})^2 = \frac{N_1 N_2}{N} (\bar{y}_{1.} - \bar{y}_{2.})^2$$

Ons probleem wordt dus: vindt een waarde voor b zodat

$$(3.5)\quad \psi = \frac{\Sigma\Sigma (\bar{y}_{j.} - \bar{y}_{..})^2}{\Sigma\Sigma (y_{jt} - \bar{y}_{j.})^2}$$

maximaal is.

Matrixnotatie
We introduceren de volgende vectoren

$$(3.6)\quad y = \begin{bmatrix} y_{11} \\ y_{12} \\ \cdot \\ \cdot \\ y_{1N_1} \\ y_{21} \\ y_{22} \\ \cdot \\ \cdot \\ y_{2N_2} \end{bmatrix}, \; h_1 = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \; h_2 = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ 1 \end{bmatrix}, \; \iota = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$$

waar de eerste N_1 elementen telkens betrekking hebben op groep S_1 en de laatste N_2 op groep S_2. We kunnen de drie gemiddelden dan achtereenvolgens weergeven als

$$(3.7)\quad \bar{y}_{1.} = \frac{1}{N_1} h_1' y, \; \bar{y}_{2.} = \frac{1}{N_2} h_2' y, \; \bar{y}_{..} = \frac{1}{N} \iota' y$$

Verder definiëren we de matrices

$$(3.8)\quad E = I - \frac{1}{N} \iota\iota', \; F = I - \frac{1}{N_1} h_1 h_1' - \frac{1}{N_2} h_2 h_2'$$

Deze matrices worden ingevoerd om waarnemingen in afwijking van bepaalde gemiddelden te nemen. Vóórvermenigvuldiging van de vector y met E geeft de elementen van y in afwijking van hun totale gemiddelde $\bar{y}_{..}$. Vóórvermenigvuldiging van y met F geeft de elementen van y in afwijking van hun eigen groepsgemiddelden, de eerste N_1 elementen in afwijking van $\bar{y}_{1.}$ en de laatste N_2 in afwijking van $\bar{y}_{2.}$.

Het is gemakkelijk na te gaan dat de matrices E en F symmetrisch en idempotent zijn, d.w.z. dat geldt

m a b blz. 245

(3.9) $E' = E$, $EE = E^2 = E$, $F' = F$ en $FF = F^2 = F$

Tenslotte definiëren we een vector h die, vermenigvuldigd met y, het verschil $\bar{y}_{1.} - \bar{y}_{2.}$, tussen beide groepsgemiddelden oplevert. Dit is

(3.10) $h = \frac{1}{N_1} h_1 - \frac{1}{N_2} h_2$

want

(3.11) $h'y = \frac{1}{N_1} h_1'y - \frac{1}{N_2} h_2'y = \bar{y}_{1.} - \bar{y}_{2.}$

Tussen E, F en h bestaat het volgende verband

(3.12) $$\begin{aligned} E - F &= I - \tfrac{1}{N}\, \iota\iota' - I + \tfrac{1}{N_1} h_1 h_1' + \tfrac{1}{N_2} h_2 h_2' \\ &= -\tfrac{1}{N}(h_1 h_1' + h_2 h_2') + \tfrac{1}{N_1} h_1 h_1' + \tfrac{1}{N_2} h_2 h_2' - \tfrac{1}{N} h_1 h_2' - \tfrac{1}{N} h_2 h_1' \\ &= \frac{N_2}{NN_1} h_1 h_1' + \frac{N_1}{NN_2} h_2 h_2' - \tfrac{1}{N} h_1 h_2' - \tfrac{1}{N} h_2 h_1' \\ &= \frac{N_1 N_2}{N} \left[\left(\tfrac{1}{N_1} h_1 - \tfrac{1}{N_2} h_2\right) \left(\tfrac{1}{N_1} h_1 - \tfrac{1}{N_2} h_2\right)' \right] \\ &= \frac{N_1 N_2}{N} hh' = nhh' \quad \left(n = \frac{N_1 N_2}{N}\right) \end{aligned}$$

waar gebruik gemaakt is van $\iota = h_1 + h_2$.
We kunnen nu (3.3) in matrixnotatie schrijven als

(3.13) $$\begin{aligned} y'Ey &= y'Fy + y'(E - F)y \\ &= y'Fy + y'hh'y \,.\, n \end{aligned}$$

en (3.5) als

(3.14) $\psi = \dfrac{y'hh'y}{y'Fy} \,.\, n$

Merk op dat de tussengroepsvariatie geschreven kan worden als

$y'hh'y \,.\, n = (\bar{y}_{1.} - \bar{y}_{2.})^2 n.$

Hiermee is dus (3.4) bewezen.

De elementen van de vector y zijn gedefinieerd als

$y_{jt} = b_1 x_{1jt} + b_2 x_{2jt} + \ldots + b_p x_{pjt}$

hetgeen in matrixalgebra neerkomt op

m a b blz. 246

(3.15) $y = Xb$

waar y de N-element kolomvector uit (3.6) is, waar X de matrix is uit (3.1) en waar b de p-element vector $(b_1\ b_2 \ldots b_p)'$ is.
Substitutie van (3.15) in (3.14) leidt tot

$$(3.16) \quad \psi = \frac{b'X'hh'Xb}{b'X'FXb}\, n$$

$$= \frac{b'dd'b}{b'Cb}\, n$$

waar $C \equiv X'FX$ en waar $d \equiv X'h$.

Het verband tussen de „binnengroepscovariantiematrix" $\frac{1}{N-2}\,C$ en de gebruikelijke matrix van varianties en covarianties

$$\{\tfrac{1}{N} \sum_j \sum_t (x_{ijt} - \bar{x}_{i..})(x_{kjt} - \bar{x}_{k..})\} = \tfrac{1}{N} X'EX$$

kan worden weergegeven als volgt:

$$(3.17) \quad X'EX = X'[F + nhh']X$$

$$= X'FX + n\,X'hh'X$$

$$= C + ndd'$$

waar gebruik gemaakt is van (3.12).

In de volgende paragraaf maximeren we $\psi' \equiv \psi/n$ hetgeen op hetzelfde neerkomt als het maximeren van ψ omdat n een constante is.

Maximering van ψ'
We zoeken dus een vector $\hat{b}$ zodat de uitdrukking

$$(3.18) \quad \psi' = \frac{b'dd'b}{b'Cb}$$

maximaal is voor $b = \hat{b}$. We stellen voor het gemak de noemer b'Cb gelijk aan 1. Dit kan zonder verlies van algemeenheid, want als hij ongelijk is aan 1, laten we zeggen gelijk aan k^2, dan kunnen we hem altijd gelijk aan 1 maken door de vector b te delen door k, hetgeen de waarde van het qiotiënt niet verandert, omdat zowel de teller als de noemer dan door k^2 worden gedeeld. De vector $\hat{b}$ is dus niet uniek; als $\hat{b}$ een maximerende vector is, is $\gamma\hat{b}$ met γ een willekeurige scalair, dit ook. De verhoudingen tussen de elementen van $\hat{b}$ zijn wel vast.

We maximeren dus b'dd'b onder de restrictie b'Cb = 1. Na invoering van een Langrange-multiplicator π wordt de te maximeren functie

(3.19) $\psi = b'dd'b - \pi\,(b'Cb-1)$

De eerste-orde voorwaarden zijn

$$\frac{\partial \psi'}{\partial b} = 2\,dd'b - 2\,\pi\,Cb = 0,$$

of

(3.20) $dd'b - \pi\,Cb = 0.$

Vóórvermenigvuldiging met b' geeft $b'dd'b - \pi\,b'Cb = 0$ waaruit, vanwege $b'Cb = 1$, volgt $\pi = b'dd'b = (d'b)^2$. Vullen we dit laatste in (3.20) in, dan krijgen we

$$d(d'b) - (d'b)^2\,Cb = 0,$$

welke uitdrukking we ook kunnen schrijven als

(3.21) $$b = \frac{1}{d'b}\,C^{-1}\,d = \alpha\,C^{-1}d$$

waar $\alpha \equiv 1/d'b$.
Als we (3.21) invullen in (3.18) krijgen we

(3.22) $$\psi' = \frac{d'C^{-1}d\;d'C^{-1}d}{d'C^{-1}d} \cdot \frac{\alpha^2}{\alpha^2} = d'C^{-1}d$$

ongeacht de waarde van de multiplicatieve scalair α.
We concluderen dat elke vector

(3.23) $\hat{b} = \gamma\,C^{-1}d$ (γ een willekeurige scalair)

een oplossing is van de eerste-orde voorwaarden (3.20), die ons quotiënt de waarde $d'C^{-1}d$ geeft. Het kan worden bewezen dat deze waarde inderdaad een maximum is.

Voor het discrimineren met behulp van de discriminantfunctie $y = x'\hat{b}$ hebben we nog een schatter $\hat{b}_0$ van b_0 nodig, waarmee $x'\hat{b}$ moet worden vergeleken, zie regel (2.8). In regel (2.8) was b_0 het gemiddelde van de gemiddelden der beide y-verdelingen. De beide kansen op misklassificatie zijn dan even groot, zie Figuur 1.

We kiezen voor $\hat{b}_0$ het gemiddelde van de beide empirische y-verdelingen:

(3.24)
$$\begin{aligned}\hat{b}_0 &= \tfrac{1}{2}\,(\tfrac{1}{N_1}\,h_1'y + \tfrac{1}{N_2}\,h_2'y)\\ &= \tfrac{1}{2}\,(\tfrac{1}{N_1}\,h_1'X\,\hat{b} + \tfrac{1}{N_2}\,h_2'X\,\hat{b})\\ &= \tfrac{1}{2}\,(\bar{X}_1 + \bar{X}_2)'\,\hat{b}\end{aligned}$$

waar $\bar{X}_j$ voorstelt het gemiddelde van de N_j waarnemingsvectoren x_{jt} ($t = 1,2,\ldots,N_j$) van groep j, zie voor notatie (3.1).

m a b blz. 248

4 Kleinste-kwadraten regressie van een dummy-variabele op X

In deze paragraaf wordt aangetoond dat maximering van het quotiënt (3.18) ook kan worden uitgevoerd met behulp van kleinste-kwadraten regressie.

Laat z een dummy-variabele zijn die de waarde 1 aanneemt als een object tot groep S_1 behoort en de waarde nul als een object tot groep S_2 behoort. De vector $z = (z_1, z_2, \ldots, z_N)'$ bestaat dus uit N_1 enen en N_2 nullen en is, eventueel na rangschikking van de z_i gelijk aan h_1 uit (3.6).

De normaalvergelijkingen van de regressie van z op de variabelen $x_1, \ldots, x_p$ zijn

$$(4.1) \quad X'EX\beta = X'Ez \; (= X'Eh_1)$$

waar X gedefinieerd is in (3.1).
Volgens (3.12) kan het rechterlid worden herschreven als

$$(4.2) \quad X'Eh_1 = X'[F + nhh']\,h_1 = X'Fh_1 + nX'hh'h_1$$

$$= 0 + nX'h\,(\tfrac{1}{N_1}\,h_1'h_1 - \tfrac{1}{N_2}\,h_2'h_1)$$

$$= nX'h = nd$$

waar o.a. gebruik gemaakt is van $h_1'h_2 = 0$ en van (3.10).
We kunnen (4.1) dus schrijven als

$$(4.3) \quad X'EX\beta = nd$$

De kleinste-kwadraten schatter is

$$(4.4) \quad \hat{\beta} = (X'EX)^{-1}\,nd$$

Substitutie van (3.17) in (4.3) geeft

$$(4.5) \quad (C + ndd')\,\beta = nd$$

of

$$C\beta = (1 - d'\beta)nd,$$

zodat de kleinste-kwadratenschatter van β geschreven kan worden als

$$(4.6) \quad \hat{\beta} = (1 - d'\hat{\beta})n\,C^{-1}d$$

$$= \gamma C^{-1}d$$

waar $\gamma = (1 - d'\hat{\beta})n$. We zien dat (4.6) op de scalaire factor γ na gelijk is aan (3.21) en we weten dat de waarde van de scalaire factor irrelevant is omdat hij geen invloed heeft op de waarde van het te maximeren quotiënt (3.18).

m a b blz. 249

Significantietoetsen

Zojuist is aangetoond dat de coëfficiënten van de discriminantfunctie kunnen worden bepaald via de kleinste-kwadratenprocedure, toegepast op een speciaal geval van het algemene lineaire model. Voor het laatste model bestaan procedures om verschillende veronderstellingen te toetsen betreffende de coëfficiëntenvector β.

We zullen laten zien dat een aantal van deze procedures ook geldig zijn voor het bijzondere lineaire model dat de coëfficiënten van de discriminantfunctie oplevert. Dit is niet automatisch het geval, omdat in het discriminantanalysemodel de basisveronderstellingen van het algemene lineaire model, waarvan de toetsen op β afhangen, niet alle opgaan. Bijvoorbeeld: de onafhankelijke variabelen $x_1, x_2, \ldots, x_p$ kunnen in het lineaire model als niet-stochastisch worden beschouwd en in het discriminantanalysemodel kan dit niet.

De te behandelen significantietoetsen zijn wel gebaseerd op de veronderstelling van normaal verdeelde kenmerken x_i.

In het algemene lineaire model wordt voor het toetsen van veronderstellingen betreffende β gebruik gemaakt van toetsgrootheden die functies zijn van de gekwadrateerde correlatiecoëfficiënt R^2, welke de proportie verklaarde variatie in de z-waarnemingen meet. De totale variatie in de z-en kan namelijk worden opgesplitst in een deel dat kan worden toegekend aan de variatie in de onafhankelijke variabelen $x_1, x_2, \ldots, x_p$, de zgn. „verklaarde variatie", en een „onverklaarde" rest:

$$(4.7) \quad z'Ez = \hat{\beta}'X'EX\hat{\beta} + \text{rest}$$

R^2 is gedefinieerd als

$$(4.8) \quad R^2 = \frac{\hat{\beta}'X'EX\hat{\beta}}{z'Ez} = \frac{\hat{\beta}'X'Ez}{z'Ez}$$

waar gebruik gemaakt is van $X'EX\hat{\beta} = X'Ez$, zie (4.1).
De grootheid

$$(4.9) \quad F = \frac{R^2/p}{(1-R^2)/(N-p-1)} = \frac{N-p-1}{p} \frac{\hat{\beta}'X'Ez}{z'Ez-\hat{\beta}'X'Ez}$$

meet de verklaarde variatie per eenheid onverklaarde variatie en heeft een F-verdeling met p en $N-p-1$ vrijheidsgraden. Hij wordt gebruikt om de hypothese $H_0: \beta = 0$ te toetsen. Als de berekende F voldoende groot is, wordt H_0 verworpen.

In het discriminantanalyse-model kan (4.7) vereenvoudigd worden, omdat

(4.10) $z'Ez = z'\,[I - \frac{1}{N}\,\iota\iota'\,]\,z = z'z - \frac{1}{N}\,z'\,\iota\iota'\,z$

$$= N_1 - \frac{1}{N}\,N_1^2 = N_1\,(1 - \frac{N_1}{N})$$

$$= \frac{N_1 N_2}{N} = n$$

en omdat volgens (4.1) en (4.2) geldt

$X'EX\hat{\beta} = X'Ez = nd.$

De gelijkheid (4.7) wordt dan

(4.11) $1 = d'\hat{\beta} + \text{rest}/n$

De uitdrukking in (4.9) wordt dus

(4.12) $F = \frac{N-p-1}{p}\,\frac{d'\hat{\beta}}{1-d'\hat{\beta}}$

Door substitutie van $\hat{\beta} = \gamma\,C^{-1}d$ en $\gamma = (1-d'\hat{\beta})n$ (zie 4.6) wordt dit

(4.13) $F = \frac{R^2/p}{(1-R^2)/(N-p-1)} = \frac{N-p-1}{p}\,n\,d'C^{-1}d$

waar C voorstelt de „binnengroepscovariantiematrix" vermenigvuldigd met N−2. We laten nu zien dat de uitdrukking (4.12) een F-verdeling heeft met p en N−p−1 vrijheidsgraden.

Beschouw twee onafhankelijke, normale populaties

(4.14) $x_1 \approx N(\mu_1, \sigma^2)$ en $x_2 \approx N(\mu_2, \sigma^2)$

en stel dat we uit elk een aselecte steekproef hebben:
$x_{11}, \ldots, x_{1N_1}$ en $x_{21}, \ldots, x_{2N_2}$
De gebruikelijke grootheid om de hypothese $H_0: \mu_1 = \mu_2$ te toetsen is de volgende, die een t-verdeling met N−2 vrijheidsgraden heeft:

(4.15) $t = \frac{\bar{x}_1 - \bar{x}_2}{S_p}\,.\,\sqrt{n}$

waar $\bar{x}_i$ (i = 1,2) voorstelt het gemiddelde van de i-de steekproef:

$$\bar{x}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$$

waar S_p^2 de „gepoolde" variantie is,

$$S_p^2 = \frac{1}{N-2}\left[(N_1 - 1) \left\{ \sum_{j=1}^{N_1} (x_{1j} - \bar{x}_1)^2/(N_1 - 1) \right\} + \right.$$
$$\left. +(N_2 - 1)\left\{ \sum_{j=1}^{N_2} (x_{2j} - \bar{x}_2)^2/(N_2 - 1) \right\} \right]$$

en waar $n = N_1N_2/N$, $N = N_1 + N_2$.

Men kan ook het kwadraat van (4.15),

$$(4.16) \quad t^2 = \frac{n(\bar{x}_1 - \bar{x}_2)^2}{S_p^2},$$

gebruiken. Deze grootheid heeft een F(1,N−2)-verdeling (een F-verdeling met 1 en N−2 vrijheidsgraden), zie bijvoorbeeld Mood & Graybill[4])

Stel nu dat we twee multivariate verdelingen hebben: $N_p(\mu_1,\Sigma)$ en $N_p(\mu_2,\Sigma)$, waar p het aantal dimensies voorstelt. Uit elk van beide hebben we een aselecte steekproef, resp. de N_1 x p matrix X_1 en de N_2 x p matrix X_2. De „gepoolde" covariantiematrix is

$$\frac{1}{N-2} C = \frac{1}{N-2} X'FX$$

waar

$$X = \begin{bmatrix} X_1 \\ \hdashline X_2 \end{bmatrix} \text{ zoals in (3.1)}$$

De multivariate analogon van (4.16) wordt

$$(4.17) \quad T^2 = n(\bar{X}_1 - \bar{X}_2)'C^{-1}(\bar{X}_1 - \bar{X}_2) \cdot (N-2) =$$
$$= n\, h'XC^{-1}X'h\,(N-2)$$
$$= d'C^{-1}d\;\; n(N-2)$$

waar h gedefinieerd is in (3.10) en d = X'h (zie 3.16).
T is Hotelling's „T-statistic" met N−2 vrijheidsgraden en het is bekend dat

$$\frac{T^2}{N-2} \cdot \frac{N-p-1}{p}$$

een F-verdeling heeft met p en N−p−1 vrijheidsgraden (zie Anderson[5]):

[4]) A. M. Mood and F. A. Graybill, *Introduction to the Theory of Statistics*, (1963) McGraw-Hill, p. 233.

[5]) T. W. Anderson, *An Introduction to Multivariate Analysis* (1966), John Wiley, p. 105-106.

$$(4.18)\quad \frac{T^2}{N-2}\,\frac{N-p-1}{p} = \frac{N-p-1}{p}\, n\, d' C^{-1} d \approx F(p, N-p-1)$$

en hiermee is aangetoond dat (4.13) en dus (4.12) een F-verdeling hebben.

Toets op een verschil tussen beide groepen S_1 en S_2
Als er geen verschil tussen beide groepen is, is $\mu_1 = \mu_2$ en is de discriminantfunctie zinloos. In dat geval heeft de verhouding verklaarde variatie/onverklaarde variatie (4.13) de neiging klein te zijn, omdat er niets te verklaren valt. Wil de discriminantfunctie zinvol zijn dan moeten μ_1 en μ_2 verschillen waardoor we voor (4.13) grotere waarden mogen verwachten. We kunnen de zinvolheid van de discriminantfunctie toetsen door de hypothese $H_0 : \mu_1 = \mu_2$ te toetsen m.b.v. de grootheid

$$(4.13)\quad F = \frac{R^2}{1-R^2} \cdot \frac{N-p-1}{p}$$

Bij een α-procent toets wordt H_0 verworpen als $F > F_{1-\alpha}(p,N-p-1)$ waar $F_{1-\alpha}(p,N-p-1)$ het $(1-\alpha)$de percentiel van de $F(p,N-p-1)$-verdeling voorstelt.

Toets op een significante bijdrage van additionele onafhankelijke variabelen
We kunnen de F-verdeling ook gebruiken om te toetsen of een discriminantfunctie gebaseerd op $x_1, \ldots, x_p, x_{p+1}, \ldots, x_{p+q}$ beter is dan één gebaseerd op $x_1, \ldots, x_p$. In dit geval gebruiken we

$$(4.19)\quad F' = \frac{R^2_{p+q} - R^2_p}{1 - R^2_{p+q}} \cdot \frac{N-p-q-1}{q}$$

R^2_{p+q} is de gekwadrateerde correlatiecoëfficiënt (de proportie verklaarde variatie) in de regressie op $x_1, x_2, \ldots, x_{p+q}$ en R^2_p is de analoge grootheid in de regressie op $x_1, x_2 \ldots, x_p$. De grootheid in (4.19) meet de relatieve toename van de proportie verklaarde variatie. De toename is significant met een niveau van α% als

$$(4.20)\quad F' > F_{1-\alpha}(q, N-p-q-1)$$

Toets op gelijkheid van de discriminantfunctie aan een tevoren gespecificeerde relatie.
We berekenen de proportie verklaarde variatie R^2 voor de geschatte discriminantfunctie en de proportie verklaarde variatie R^2_c voor de tevoren gespecificeerde relatie en verwerpen de gelijkheid met een niveau van α procent als

$$(4.21) \quad F'' = \frac{R_p^2 - R_c^2}{1 - R_p^2} \cdot \frac{N-p-1}{p} > F_{1-\alpha}(p, N-p-1)$$

Als $c' = (c_1, \ldots, c_p)$ de coëfficiënten van de gespecificeerde relatie zijn, dan is

$$R_c^2 = \frac{d'c}{1 - d'c}$$

Merk op dat (4.21) identiek is aan (4.13) als $c = 0$.

Geschatte kansen op misklassificatie

Die vinden we door in (2.10) voor μ_1, μ_2 en Σ de zuivere schatters $\bar{X}_1$, $\bar{X}_2$ en $\frac{1}{N-2} C$ in te vullen. De geschatte kansen zijn beide gelijk aan $N(-\frac{1}{2} D)$, waar

$$D^2 = (\bar{X}_1 - \bar{X}_2)' C^{-1} (\bar{X}_1 - \bar{X}_2)(N-2) = d'C^{-1} d (N-2) =$$

$$= \frac{R^2}{1-R^2} \cdot \frac{N-2}{n}$$

zie (4.13).

De grootheid D^2 is „Mahalanobis' generalized distance". Men mag verwachten dat de schattingen goed zijn ingeval van een groot aantal vrijheidsgraden want D^2 is een consistente schatter van $\delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$.

Een bezwaar van deze methode voor het schatten van kansen op misklassificaties is dat hij uitgaat van de normaliteitsveronderstelling. Afgezien hiervan is er nog een ander bezwaar. Het is bekend dat in het algemeen D^2 een overschatting van δ^2 oplevert, waaruit volgt dat $N(-\frac{1}{2}\delta)$ door $N(-\frac{1}{2}D)$ wordt onderschat. Vooral om dit laatstgenoemde bezwaar op te heffen is een nogal groot aantal alternatieve schattingsmethoden ontwikkeld.

Lachenbruch en Mickey[6]) hebben de kwaliteit van een aantal methoden om de kansen op misklassificatie te schatten, waaronder de zojuist genoemde, onderzocht d.m.v. simulatie uit multivariate normale verdelingen. Voor kleine steekproeven bevelen zij de zgn. U-methode aan, die als volgt te werk gaat. Men schat een discriminantfunctie uit $N-1$ waarnemingen en klassificeert de resterende waarneming met die functie en deze procedure wordt uitgevoerd voor alle N combinaties van $N-1$ waarnemingen. De kansen op misklassificatie worden geschat door het aantal fout geklasseerde gevallen voor elke groep te delen door het aantal elementen van de groep. We noemen deze schatters resp. P_1^* en P_2^*. Zoals hij hier is beschreven zou de U-methode N maal een matrixinvertering eisen, voor elke discriminantfunctie eenmaal. Lachenbruch[7]) heeft een rekenwijze aan de hand gedaan waarbij met één invertering kan worden volstaan.

6) P. A. Lachenbruch and M. Ray Mickey, Estimation of Error Rates in Discriminant Analysis, Technometrics, Vol. 10, No. 1 (1968) pp. 1-11.

7) P. A. Lachenbruch, An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis, Biometrics, December (1967) pp. 639-645.

Noem de i-de rij van X in (3.1) x_i' en laat $\bar{X}_1' = h_1' X/N_1$ voorstellen de gemiddelde rij van de N_1 met groep S_1 corresponderende rijen en laat $\bar{X}_2'$ analoog zijn gedefinieerd. Definiëren we verder de i-de rij minus de gemiddelde rij van zijn groep

$$U_i' = x_i' - \bar{X}_k' \quad (i = 1,2,\ldots, N;\ k = 1,2)$$

dan kan de discriminantfunctie, die gebruikt wordt voor de klassificatie van x_i $(i = 1,\ldots,N)$, als volgt worden berekend

$$(N-3)\left[x_i - \tfrac{1}{2}(\bar{X}_1 + \bar{X}_2) + \frac{U_i}{2(N_k - 1)}\right]' S_i^{-1}\left[\bar{X}_1 - \bar{X}_2 + (-1)^k \frac{U_i}{N_k - 1}\right]$$

waar

$$S_i^{-1} = C^{-1} + \frac{N_k C^{-1} U_i U_i' C^{-1}}{N_k - 1 - N_k U_i' C^{-1} U_i}$$

waar C gelijk is aan $(N-2)$maal de „gepoolde" covariantiematrix en waar k de index van de groep is waartoe x_i behoort.

Lachenbruch heeft ook een methode gesuggereerd voor het benaderen van betrouwbaarheidsintervallen voor de beide kansen op misklassificatie.

Laat P_1 voorstellen de kans op het ten onrechte in groep S_2 klassificeren van een element uit S_1 en laat P_1^* de zojuist beschreven schatter van P_1 zijn. Een 100 $(1-\alpha)$-procent betrouwbaarheidsinterval voor P_1 kan dan worden benaderd door de volgende vergelijking op te lossen.

$$(4.22) \quad N_1 (P_1^* - P_1)^2 / P_1 (1 - P_1) = z_{\alpha/2}^2$$

waar $z_{\alpha/2}$ voorstelt het 100 $\alpha/2$-de percentiel van de standaard normale verdeling en waar N_1 het aantal elementen van groep S_1 in de steekproef is.

Deze methode berust op de volgende redenering. Als Z_j een dummy-variabele is die de waarde 1 krijgt als een element uit S_1 en S_2 wordt geklassificeerd en anders de waarde 0 aanneemt, kunnen we de schatter P_1^* schrijven als

$$(4.23) \quad P_1^* = \frac{1}{N_1} \sum_{j=1}^{N_1} Z_j$$

De kansvariabele Z_j neemt met een kans P_1 de waarde 1 aan en met een kans $1 - P_1$ de waarde 0. Zijn verwachting en zijn variantie zijn dus respectievelijk $E(Z_j) = P_1$ en $\text{var}(Z_j) = P_1(1 - P_1)$. Dus

$$E(P_1^*) = \frac{1}{N_1} \sum_{j=1}^{N_1} E(Z_j) = P_1$$

$$\text{var}(P_1^*) = [\sum_j \text{var}(Z_j) + \sum_{i \neq j} \sum \text{cov}(Z_i Z_j)]/N_1^2 =$$

$$= \frac{P_1(1-P_1)}{N_1}[1 + (N_1 - 1)\rho]$$

waar ρ de correlatiecoëfficiënt tussen Z_i en Z_j is. Lachenbruch laat zien dat verwacht mag worden dat ρ kleine waarden aanneemt, verwaarlozing ervan geeft slechts geringe fouten. De gestandaardizeerde grootheid

$$\frac{P_1^* - P_1}{\sqrt{P_1(1-P_1)/N_1}}$$

heeft dan bij benadering een standaard normale verdeling; op dit feit is (4.22) gebaseerd.

Voor P_2 kan men op analoge wijze een betrouwbaarheidsinterval benaderen.

Merk op dat de zojuist beschreven methode geen normaliteit van de x_i veronderstelt, hij is nonparametrisch van karakter.

5 Het geval van meer dan twee groepen

In deze paragraaf volgen we voor de afleiding van een klassificatieprocedure voor meer dan twee groepen een minder intuïtieve weg dan we in Paragraaf 1 voor twee groepen deden. Daardoor zal blijken dat de procedure (2.2) en dus procedure (2.8) een procedure is die het verwachte verlies minimeert.

We beschouwen een verzameling objecten S die gepartitioneerd is in de deelverzamelingen $S_1, S_2, \ldots, S_J$ (zie Paragraaf 1). We willen de waarnemingsruimte van x opdelen in J onderling uitsluitende en uitputtende deelruimten $R_1, R_2, \ldots, R_J$, zodat een object in S_j wordt geklassificeerd als de met dat object corresponderende waarde van x in R_j valt.

We noemen de kosten van het klassificeren van een element uit S_i in S_j c(j/i). De kans op deze misklassificatie is

$$(5.1) \quad P(j/i) = \int_{R_j} l_i(x) \, dx$$

waar $l_i(x)$ de aannemelijkheidsfunctie van x in groep S_i voorstelt. Stel dat $P_1, P_2, \ldots, P_J$ de a priori kansen zijn dat een getrokken element resp. tot groep S_1, groep $S_2, \ldots,$ groep S_j behoort. Dan is het verwachte verlies

$$(5.2) \quad \sum_{i=1}^{J} P_i \sum_{\substack{j=1 \\ j \neq i}}^{J} P(j/i) \; c(j/i)$$

Het probleem is de verschillende R_j zo te kiezen dat deze uitdrukking minimaal wordt.

De bijdrage van een steekproefpunt x aan het verwachte verlies is

m a b blz. 256

$$(5.3) \quad h_j(x) = \sum_{\substack{i=1 \\ i \neq j}}^{J} P_i l_i(x)\ c(j/i)$$

bij klassificatie in groep S_j. Om het verwachte verlies te minimeren, klassificeren we het element in die groep waarvan $h_j(x)$ minimaal is; als min $h_j(x) = h_k(x)$, dan klassificeren we in S_k en het element x kennen we aan de deelruimte R_k toe. Deze procedure herhalen we voor alle waarden van x en zo vormen we de deelruimten $R_1, R_2, \ldots, R_J$.

Dus, we kennen een punt x aan R_k toe als

$$(5.4) \quad h_k(x) < h_j(x) \quad \text{voor } j = 1,2,\ldots,J \ (j \neq k)$$

of als

$$(5.5) \quad \sum_{\substack{i=1 \\ i \neq k}}^{J} P_i l_i(x)\ c(k/i) < \sum_{\substack{i=1 \\ i \neq j}}^{J} P_i l_i(x)\ c(j/i)$$

$$\text{voor } j = 1, 2, \ldots, J \quad (j \neq k)$$

Stel $c(i/j) = 1$ en trek in (5.5) aan weerskanten af

$\sum_{\substack{i=1 \\ i \neq j,\,k}}^{J} P_i l_i(x)$. Dan wordt (5.5)

$$(5.6) \quad P_j l_j(x) < P_k l_k(x)$$

Beide termen in deze ongelijkheid zijn, op een constante na, a posteriori kansen; elk is het product van een a priori kans P en een aannemelijkheid l(x). Als (5.6) geldt voor $j = 1,2,\ldots,J$ $(j \neq k)$ kennen we dus x aan R_k toe, waar k de index is die de a posteriori kans $P_i l_i(x)$ maximeert. We klassificeren het corresponderende object in S_k, dus in de *meest waarschijnlijke groep*.

We generalizeren nu het basismodel voor twee groepen dat in Paragraaf 1 behandeld is, d.w.z. dat de vector x p-dimensionaal normaal verdeeld is:

$$(5.7) \quad l(x;\mu_j) = c \exp\{-\tfrac{1}{2}(x-\mu_j)'\Sigma^{-1}(x-\mu_j)\} \ (j = 1,\ldots,J).$$

R_k is dan gedefinieerd door die x-en waarvoor

$$P_j l(x;\mu_j) < P_k(x;\mu_k)$$

of

$$(5.8) \quad R_k : \lambda \equiv \frac{l(x;\mu_k)}{l(x;\mu_j)} > \frac{P_j}{P_k} \quad \text{voor } j = 1,\ldots,J \ (j \neq k)$$

m a b blz. 257

Door vereenvoudiging van λ volgens (2.3) en door het nemen van de logarithme wordt (5.8)

$$(5.9) \quad R_k : x'b - b_0 > \log \frac{P_j}{P_k} \text{ voor } j = 1, \ldots, J \ (j \neq k)$$

waar

$$(5.10) \quad b = \Sigma^{-1}(\mu_k - \mu_j) \text{ en } b_0 = \tfrac{1}{2}(\mu_k + \mu_j)'b$$

We zien direct dat voor $J = 2$ en gelijke a priori kansen de klassificatieregel (5.9) gelijk is aan regel (2.8).

Als de parameters μ_j en Σ niet bekend zijn en als van elk van de J populaties S_j een steekproef ter grootte van N_j beschikbaar is, vullen we in (5.10) in de schatters

$$(5.11) \quad \hat{\mu}_j = \frac{1}{N_j} \sum_{t=1}^{N_j} (x_{1jt}\ x_{2jt} \ldots x_{pjt})' \equiv \frac{1}{N_j} \sum_{t=1}^{N_j} x_{jt}$$

$$\hat{\Sigma} = \sum_{j=1}^{J} \sum_{t=1}^{N_j} (x_{jt} - \hat{\mu}_j)(x_{jt} - \hat{\mu}_j)' / \left(\sum_{j=1}^{J} N_j - J \right)$$

waar x_{ijt} de waarneming van kenmerk i betreffende het t-de object van groep S_j voorstelt. Zie (3.1) voor notatie.

6 Illustratie voor twee groepen

We zullen nu het gebruik van de discriminantanalyse illustreren met een probleem uit de financieringstheorie. Het probleem betreft de voorspelbaarheid van ernstige financiële moeilijkheden (insolventie) van ondernemingen[8]).

Het begrip insolventie

Onder insolventie verstaan we de situatie waarin een onderneming niet op tijd kan voldoen aan haar financiële verplichtingen. Sommige insolvente ondernemingen krijgen surséance van betaling en worden gereorganiseerd; voor andere komt de hulp niet of te laat zodat ze failliet gaan. De mogelijkheid om insolventie te voorspellen is belangrijk zowel voor de onderneming als voor haar financiers. Een tijdige signalering van dit probleem stelt hen in staat om preventieve maatregelen te nemen bijvoorbeeld door het beleid te veranderen, de financiële structuur te reorganiseren of zelfs te besluiten tot liquidatie om verdere verliezen te voorkomen.

[8]) Deze paragraaf is een bewerking van een onderdeel van het artikel „Voorspelbaarheid van insolventie" van R. A. I. van Frederikslust. M.A.B. no. 4, 1976.

Een model

We zullen nu een model ontwikkelen waarmee insolventie van ondernemingen voorspeld kan worden. Daarna zal het model worden getoetst aan een steekproef van 20 solvente en 20 insolvente beursondernemingen betrekking hebbend op de periode 1954-1974. Bij het ontwikkelen van het model zal gebruik worden gemaakt van de volgende symbolen:

Symbolen[9])

NO_t = interne middelen (inclusief het beginsaldo van de liquide middelen) aan het eind van de periode t

KV_t = omvang van de korte schulden aan het eind van periode t (inclusief de aflossing van lange leningen)

EL_t = omvang van de in periode t aangetrokken lange leningen

LI_t = saldo van de liquide middelen aan het eind van periode t

Met behulp van deze symbolen kan voor het kassaldo LI_t al meteen de volgende vergelijking worden opgesteld.

$$(6.1) \quad LI_t = NO_t + KV_t - KV_{t-1} + EL_t$$

Het kassaldo LI_t is gelijk aan de zelf verdiende middelen NO_t plus de mutatie van de korte schulden $KV_t - KV_{t-1}$ plus de in periode t aangetrokken lange leningen EL_t.[10]) Een onderneming is insolvent als haar liquide middelen $LI_t < 0$ of als:

$$(6.2) \quad NO_t + KV_t - KV_{t-1} + EL_t < 0$$

We kunnen deze ongelijkheid ook schrijven als

$$(6.3.) \quad \frac{NO_t + KV_t + EL_t}{KV_{t-1}} < 1 <=> \frac{NO_t}{KV_{t-1}} + \frac{KV_t + EL_t}{KV_{t-1}} < 1$$

De term NO_t/KV_{t-1} is de „eigen dekking" van de korte-termijn schulden, KV_{t-1} en de term $(KV_t + EL_t)/KV_{t-1}$ is de „vreemde dekking" ervan.

De grootte van de „vreemde dekking" hangt af van de beslissing van de bedrijfsleiding en van de financiers. De bedrijfsleiding zal een beslissing nemen op basis van een afweging van de potentiële opbrengsten (rentabiliteit van het eigen vermogen) en de liquiditeitsrisico's van schuldfinanciering[11]).

---

9) Zie voor een (nadere) toelichting van deze grootheden R. A. I. van Frederikslust, Voorspelbaarheid van insolventie, paragraaf 3.3.

10) Investeringen in duurzame activa en dividenduitgaven enz. worden als uitstelbare uitgaven beschouwd en worden derhalve niet in het insolventiemodel opgenomen.

Voorts wordt verondersteld dat in een crisissituatie aandelen niet geplaatst kunnen worden, zodat ook deze grootheid niet in LI_t wordt opgenomen.

11) Zie ook A. I. Diepenhorst en H. Willems, De optimale financiële structuur. Kernproblemen der Bedrijfseconomie, C. F. Scheffer e.a. Agon Elsevier, Amsterdam/Brussel 1966, p. 195 en F. W. C. Blom, Leencapaciteit van ondernemingen p. 214. Bedrijfskunde 1974/4.

De rentabiliteit van het eigen vermogen in periode t is gelijk aan WN_t/EV_t waar WN_t = netto winst na belasting in periode t en EV_t het eigen vermogen aan het eind van periode t. Bij het beoordelen van de kredietwaardigheid van een onderneming zal de financier ook de nadruk leggen op de criteria die de bedrijfsleiding hanteert[12]).

De grootte van de „vreemde dekking" $(KV_t + EL_t)/KV_{t-1}$ zal daarom in hoofdzaak bepaald worden door de liquiditeitsratio NO_t/KV_{t-1} en de rentabiliteitsratio WN_t/EV_t. We kunnen aannemen dat $(KV_t + EL_t)/KV_{t-1}$ een stijgende functie, zeg f(.), is van de criteria $x_1 \equiv NO_t/KV_{t-1}$ en $x_2 \equiv WN_t/EV_t$; dan geldt

$$(6.4) \quad (KV_t + EL_t)/KV_{t-1} = f(x_1, x_2)$$

Door substitutie in (6.3) kunnen we zeggen dat er sprake is van insolventie wanneer

$$(6.5) \quad x_1 + f(x_1, x_2) < 1$$

De liquiditeitsratio x_1 en de rentabiliteitsratio x_2 zullen dus in dit model de insolventie van een onderneming kunnen bepalen.[13]) Er zijn geen a priori overwegingen om te verwachten dat x_1 en x_2 sterk gecorreleerd zijn. Immers de rentabiliteit x_2 is een variabele die de ontwikkeling van de liquiditeit op de lange termijn mede bepaalt. We zullen hieronder trachten met het model insolventie van ondernemingen één jaar tevoren te voorspellen. Voor we tot het bespreken van de voorspelfunctie en de voorspelresultaten overgaan zullen we eerst de steekproefdata bespreken.

Data

Het onderzoek betreft 20 beursondernemingen die in de periode 1954-1974 insolvent zijn geworden. De range van de totale activa van 19 van de 20 ondernemingen was 1.02–30.5 mln gulden (gemiddeld 15.2 mln gulden) één jaar vóór insolventie. Naast deze kleine en middelgrote was er nog een grote onderneming die in de beschouwde periode insolvent werd. De totale activa van deze onderneming beliepen één jaar voor insolventie 153,2 mln gulden. Voor elke insolvente onderneming[14]) werd een solvente partner gekozen uit dezelfde bedrijfstak die in de desbetreffende periode qua grootte (totale activa) zoveel mogelijk overeenkwam met zijn insolvente partner („paired sample design").[15]) De range van de totale activa van 19 van de 20 solvente bedrijven was in het eerste jaar 1.7–33.5 mln gulden (gemiddeld 16.3 mln

---

12) Zie O. Vogelenzang, Informatiebehoefte bij de verschaffers van leenvermogen. Maandblad voor Accountancy en Bedrijfshuishoudkunde no. 7, pp. 281-289, 1974.

13) We gaan hier niet in op de vorm van $f(x_1,x_2)$. Een nader onderzoek volgt hierna.

14) Het is niet strikt nodig de beide groepen bedrijven even groot te kiezen. Men zou bijvoorbeeld de solvente groep groter kunnen kiezen dan de insolvente groep, indien men over relatief weinig insolvente bedrijven beschikt. Voor ons probleem is een steekproef van 40 bedrijven groot genoeg.

15) Deze steekproefsamenstelling heeft tot doel de invloed van de interveniërende variabelen („grootte" en „bedrijfstak") uit te schakelen, omdat deze variabelen storend kunnen werken op de relatie tussen de ratio's en de kans op insolventie.

gulden). Er was geen grote solvente onderneming te vinden van ongeveer gelijke grootte in het eerste jaar als de grote insolvente onderneming. De totale activa van de onderneming die we gekozen hebben waren in het eerste jaar 245.3 mln gulden. Het verschil in omvang tussen beide grote ondernemingen over een periode van zes jaar was evenwel kleiner. De gemiddelde waarde van de totale activa van de grote insolvente onderneming was 186.3 mln en van de grote solvente onderneming 185.0 mln gulden.

Resultaten

Voor elk van de 40 beursondernemingen hebben we de liquiditeitsratio x_1 en de rentabiliteitsratio x_2 berekend die hiervoor zijn afgeleid, één jaar voor „insolventie". Op basis daarvan is een functie geschat waarmee de bedrijven kunnen worden geklassificeerd naar een van de beide groepen. Bij het schatten van die functie is gebruik gemaakt van een regressie-computerprogramma omdat dat in het algemeen beter ontwikkeld is dan discriminantprogramma's. In Paragraaf 4 is aangetoond dat in geval van 2 groepen de regressiecoëfficiënten proportioneel zijn met de coëfficiënten van de discriminantfunctie indien aan de afhankelijke variabele y de waarde 1 (solvent) en 0 (insolvent) wordt toegekend.[16]) We zullen nu de geschatte functie bespreken. Deze heeft de volgende gedaante.

$$(6.6) \quad y_{jt} = .5293 + .4488\, x_{1jt} + 2863\, x_{2jt} \qquad j = 1,2$$
$$(.1129) \quad (.0729) \qquad t = 1,2,\ldots 20$$

waarbij x_{1jt} = de liquiditeitsratio van bedrijf t van groep j
x_{2jt} = de rentabiliteitsratio van bedrijf t van groep j
y_{jt} = de regressie-score van bedrijf t van groep j.

De tekens van de regressie-coëfficiënten zijn in overeenstemming met de theorie[17]). De gemiddelde waarde van de liquiditeitsratio van de solvente groep $\bar{x}_{11}$ is groter dan die ($\bar{x}_{12}$) van de insolvente groep. Hetzelfde geldt voor de gemiddelden van de rentabiliteitsratio van de beide groepen: $\bar{x}_{21} > \bar{x}_{22}$. Omdat bij het schatten van de functie aan y_{jt} de waarde 1 is toegekend voor solvente bedrijven en 0 voor insolvente bedrijven zal bij positieve regressie-coëfficiënten de gemiddelde score $\bar{y}_1$, van solvente bedrijven groter zijn dan die van insolvente bedrijven $\bar{y}_2$.

Het spreekt vanzelf dat discriminatie m.b.v. (6.6) slechts zinvol is als de vector μ_1, van de gemiddelde waarden van de onafhankelijke variabelen van de solvente groep verschilt van de corresponderende vector μ_2 voor de insolvente groep.

---

16) Een regressie op basis van een onafhankelijke dummy variabele is hier gerechtvaardigd omdat deze regressie slechts gebruikt wordt om de bedrijven in groepen te klassificeren. Voor andere doeleinden is zo'n regressie minder geschikt omdat daarbij niet voldaan wordt aan de klassieke kleinstekwadraten veronderstelling van homoscedasticiteit. (Zie ook A. S. Goldberger, Econometric Theory. New York. John Wiley and Sons 1964, pp. 248-247).

17) Merk op dat de regressie-coëfficiënten groter zijn dan twee maal de standaardfout die daarbij tussen haken is opgenomen; hoewel hier niet dezelfde betekenis mag worden toegekend aan de standaardfouten als in het lineaire model. Zie ook voetnoot 16.

Significantietoets[18])
De zinvolheid van de geschatte functie hebben we dan ook getoetst door de hypothese $H_o : \mu_1 = \mu_2$ te toetsen met de grootheid F gedefinieerd in vergelijking (4.13). Het resultaat van deze toets is in de eerste rij van onderstaande tabel opgenomen.

Tabel 6.1 Significantietoetsen

	R^2	F-waarde
y_{jt}	.517	19.87[a]
y'_{jt}	.316	11.24[a]
y''_{jt}	.311	9.78[a]

a)
$\alpha_{min} < .005$ $F_{.995}(2,37) = 6.15$
$F_{.995}(1,37) = 8,94$

Daaruit blijkt dat de gekwadrateerde correlatie-coëfficiënt R^2 van de geschatte functie y_{jt} gelijk is aan .517 en de berekende F-waarde 19.87.

Deze is groter dan $F_{.995}$ (2,37) = 6.15 zodat de hypothese H_o kan worden verworpen.

De (rechter) overschrijdingskans α_{min} (de kans waarmee de gevonden uitkomst van F nog juist tot verwerping van H_o zou leiden) is dus kleiner dan .005. We mogen derhalve verwachten dat met behulp van y_{jt} met succes zal kunnen worden gediscrimineerd tussen de bedrijven van de twee groepen.

Toets op een significante bijdrage van additionele onafhankelijke variabelen
We hebben ook de F-verdeling gebruikt om te toetsen of de regressie y_{jt} gebaseerd op x_1 én x_2 beter is dan y'_{jt} en y''_{jt}, waar y'_{jt} de regressie is gebaseerd op x_1 en y''_t op x_2. In Tabel 6.1 zijn de gekwadrateerde correlatie-coëfficiënten R^2 van deze functies opgenomen. Op basis van vergelijking (4.19) hebben we de F-waarde in geval van y'_{jt} en y''_{jt} berekend die ook in de tabel zijn opgenomen. Daaruit blijkt dat de berekende F-waarde i.g.v. y'_{jt} (11.24) groter is dan $F_{.995}(1,37) = 8.94$; $\alpha_{min} < .005$. Hetzelfde geldt t.a.v. de F-waarde i.g.v. y''_{jt}. We kunnen derhalve stellen dat zowel de liquiditeitsratio x_1 als de rentabiliteitsratio x_2 een significante bijdrage leveren tot het discriminerend vermogen van de geschatte functie y_{jt}.

Geschatte kansen op misklassificatie[19])
We hebben de voorspelresultaten van de geschatte functie y_{jt} geëvalueerd met behulp van de „methode-Lachenbruch" (zie paragraaf 4). Er zijn 40

18) De toetsen die hieronder zullen worden uitgevoerd berusten op de veronderstelling van normaal verdeelde variabelen x_i. In dat geval is log λ (zie paragraaf 2) lineair.

19) De procedure die nu volgt gaat weer niet uit van de veronderstelling van normaal verdeelde x_i. Zie ook paragraaf 4.

regressies geschat op basis van de waarnemingen (ratio's) voor elke combinatie van 39 uit de groep van 40 bedrijven. Voor elke regressie hebben we het weggelaten bedrijf geklassificeerd op basis van een kritieke regressie-score y* die gelijk is aan de gemiddelde score van y_{jt} van de bedrijven; de klassificatie-procedure is als volgt[20]).

$$(6.8) \quad y_{jt} \begin{cases} > y^* \text{ bedrijf} \rightarrow \text{solvente groep } S_1 \\ < y^* \text{ bedrijf} \rightarrow \text{insolvente groep } S_2 \end{cases}$$

Voorts hebben we op basis van vergelijking (4.22) een 100 (1−α-procent betrouwbaarheidsinterval voor elk van de kansen P_1 en P_2 op misklassificatie berekend.

De geschatte kansen op misklassificatie van y_{jt} zijn als volgt:

1. voor de solvente groep S_1
   a. geschatte kans op misklassificatie $P_1^* = .05$
   b. een 95 procent betrouwbaarheidsinterval
      $.009 < P_1 < .236$

2. voor de insolvente groep S_2
   a. geschatte kans op misklassificatie $P_2^* = .10$
   b. een 95 procent betrouwbaarheidsinterval
      $.028 < P_2 < .301$

De functie voorspelt iets beter voor de solvente groep. In totaal zijn 92.5 procent van alle bedrijven correct geklassificeerd.

De gemiddelde waarde van de kritieke score y* die daarbij gehanteerd werd is gelijk aan .50.

Men zou de vraag kunnen stellen of de proportie correcte klassificatie die verkregen is met het discriminantanalysemodel significant verschilt van die van een aselecte klassificatie. In ons geval zijn de groepen even groot. Bij een aselecte klassificatie wordt gemiddeld de helft van de bedrijven correct geklassificeerd. We hebben de volgende grootheid gebruikt om te toetsen of de resultaten van de discriminantanalyse beter zijn dan die van een aselecte klassificatie[21]):

$$(6.11) \quad z = \frac{Q - P}{\sigma_p}$$

waarbij

---

[20]) Vergelijk deze klassificatie-procedure met die van vergelijking (2.8). De waarde van y* is, afgezien van een te verwaarlozen verschil tussen de grootte van de groepen, gelijk aan $\hat{b}_0$ gedefinieerd in (3.24).

[21]) Zie R. E. Frank, W. F. Massy en D. G. Morrison, Bias in multiple discriminant analysis. Journal of Marketing Research, vol. 11, aug. 1965 p. 253 e.v.

Q is de proportie van de steekproef die correct is geklassificeerd door de discriminantfunctie

P is de proportie correcte klassificaties ingeval men over beide groepen klassificeert met kansen P en 1–P.
Als de groepen even groot zijn is P = .5

$$\sigma_p \text{ is } \frac{\sqrt{P\,(1-P)}}{\sqrt{N}}$$

De grootheid z heeft een standaard normale verdeling. Hiermee kan de hypothese Ho : E(Q) = P worden getoetst. Naarmate Q groter is dan P zal z *ceteris paribus* groter zijn. Bij een α-procent toets wordt de hypothese H_o verworpen als $z > z_{1-\alpha}$ waar $z_{1-\alpha}$ het 100 $(1-\alpha)$de percentiel van de standaard normale verdeling voorstelt.

Bij empirische onderzoekingen is het beter α niet a priori te kiezen maar het kleinste percentage α_{min} te vermelden waarvoor de gevonden uitkomst van z nog juist tot verwerping van de hypothese H_o zou leiden (de zgn. overschrijdingskans) omdat die meer informatie verschaft.

In ons geval is Q = .925, P = .5, en N = 40 zodat z = 5.38, hetgeen een overschrijdingskans α_{min} van nul impliceert.

Recapitulatie

Hiervoor is een functie geconstrueerd op basis van de liquiditeitsratio x_1 en de rentabiliteitsratio x_2 van beursondernemingen één jaar voor insolventie. De gemiddelde waarden van de corresponderende ratio's van de groepen verschillen voldoende van elkaar zodat met de functie effectief kan worden gediscrimineerd tussen de bedrijven van de beide groepen. Elk van de ratio's draagt wezenlijk bij tot het discriminerend vermogen van de functie. De relatieve bijdrage van de liquiditeitsratio is evenwel iets groter dan die van de rentabiliteitsratio. De kans op een onterechte klassificatie in de insolvente groep is 5 procent en in de solvente groep 10 procent. In totaal werden 92.5 procent van alle bedrijven correct geklassificeerd. De voorspelresultaten van ons model zijn significant beter dan die van een puur kansmodel.

Literatuur

Anderson, T. W., *An Introduction to Multivariate Analysis* (1966), John Wiley.

Blom, F. W. C., Leencapaciteit van ondernemingen pp. 213-220, *Bedrijfskunde* 1974/4.

Diepenhorst, A. I. en Willems H., De optimale financiële structuur. *Kernproblemen der Bedrijfseconomie.* C. F. Scheffer e.a. Agon Elsevier, Amsterdam, Brussel 1966 pp. 184-197.

Frank, R. E. Massy, W. F. and Morrison, D. G., Bias in Multiple Discriminant Analysis. *Journal of Marketing Research,* Vol. 11 Aug., 1965 pp. 250-258.

Frederikslust, R. A. I. van, Voorspelbaarheid van insolventie. *Maandblad voor Accountancy en Bedrijfshuishoudkunde* No. 4, 1976.

Goldberger, A. S., Econometric Theory (1964) pp. 248-257. John Wiley and Sons. New York.
of misclassification in discriminant analysis, *Biometrics,* December (1967) pp. 639-645.

Lachenbruch, P. A. and Mickey M. R., Estimation of Error Rates in Discriminant Analysis, *Technometrics* Vol. 10 No. 1 (1968) pp. 1-11.

Mood, A. M. and Graybill, F. A., *Introduction to the Theory of Statistics* (1963) McGraw-Hill.

Vogelenzang, O., Informatiebehoefte bij de verschaffers van leenvermogen. *Maandblad voor Accountancy en Bedrijfshuishoudkunde* no. 7 1974, pp. 281-289.

m a b blz. 264