INSTITUUT VOOR CULTUURTECHNIEK EN WATERHUISHOUDING
NOTA No.140 d.d. 25 juli 1962

Selectie van variabelen bij meervoudig regressieonderzoek

L.P. Kamil

1. Inleiding

Verondersteld wordt, dat men bij een onderzoek een variabele in statistische zin, volgens een bepaalde norm, goed kan verklaren uit een aantal andere variabelen. Men kan zich dan afvragen, of men met een combinatie van een kleiner aantal verklarende variabelen tot een bijna even goede verklaring kan komen.

Voor een onderzoek hiernaar kunnen de volgende redenen gelden:

1e. een of enige variabelen, welke bij de opzet van het onderzoek zijn opgenomen, dragen niet bij tot de verklaring;
2e. een of enige variabelen zijn (bijna) afhankelijk van enige andere basisvariabelen;
3e. indien men zo'n kleinere combinatie kan aanwijzen, kan men in het vervolg kosten en tijd sparen, omdat metingen aan uitgeselecteerde variabelen niet behoeven te worden gedaan.

Stelt men zich tot doel zo'n kleinere combinatie te vinden, dan zal men in het 1e geval tot één oplossing kunnen komen, terwijl in het 2e geval verscheidene oplossingen mogelijk moeten zijn. Zijn alle variabelen nodig voor de verklaring, dan zal men geen kleinere combinatie kunnen vinden.

In het geval, dat verscheidene oplossingen mogelijk zijn, welke alle een ongeveer even goede verklaring geven, is het zinvol, om alle combinaties te kennen. Immers, dan kan men op bijvoorbeeld economische of technische gronden een verantwoorde keuze doen.

In de literatuur worden twee voorstellen gedaan voor het oplossen van bovengenoemd probleem, die beide, zowel in het 1e als in het 2e geval tot één oplossing komen, terwijl alle mogelijke andere combinaties van het 2e geval worden verduisterd. Verder blijkt, dat de combinatie, welke gevonden wordt met de ene methode, niet gelijk behoeft te zijn aan die, welke met de andere methode tot stand komt.

Hamaker [3]geeft van deze methoden - namelijk, forward selection en

backward elimination - enige voorbeelden en een literatuuroverzicht. Corsten [1] geeft een methode, die in gewijzigde vorm, tot verscheidene oplossingen kan komen.

In het volgende zal een en ander worden uitgewerkt en onderling vergeleken, met tot slot ter toelichting een voorbeeld uit Hald [2] pag. 647-650.

2. Symbolen en begrippen

Om storende onderbrekingen in de tekst te voorkomen, worden eerst enige veel gebruikte symbolen en begrippen gedefinieerd.

Stochastische grootheden worden aangegeven door een onderstreepte letter, bv. $\underline{x}$. Een vector wordt als zodanig in de tekst gedefinieerd. Twee stochastische variabelen $\underline{x}$ en $\underline{y}$ zijn isomoor ($\underline{x} \cong \underline{y}$) als ze dezelfde kansfunctie hebben.

In formule :

$$\underline{x} \cong \underline{y} \text{ , indien } P(\underline{x} \leqslant c) = P(\underline{y} \leqslant c) \quad \text{voor iedere } c.$$

De verwachtingswaarde van een stochastische variabele (vector) ($\underline{x}$) wordt aangegeven door het symbool $E(\underline{x})$.

Het symbool $\underline{\chi}$ wordt gebruikt voor de stochastische variabele met kansdichtheid :

$$f(x) = (2\pi)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}x^2).$$

Een stochastische vector bestaande uit n onderling onafhankelijke elementen $\underline{x}$, zodat iedere $\underline{x} \cong \underline{\chi}$, wordt aangeduid met het symbool $\underline{\chi}_n$.

3. Meervoudige lineaire regressie

Ter verduidelijking van de wijze, waarop het in de inleiding genoemde probleem behandeld wordt, volgt in het kort de oplossing van het regressieprobleem door gebruikmaking van vectoren.

Men heeft een omschreven verzameling individuen, elk met k eigenschappen. Uit deze verzameling trekt men een steekproef van n individuen en van elk individu worden de k eigenschappen gemeten.

Er zijn dus n groepen van naar k geordende waarnemingsuitkomsten :

$$y_{io}, y_{i1}, \cdots, y_{ij}, \cdots, y_{ik} \qquad i = 1, 2, \cdots, n \quad ,$$

waarin het getal y_{ij} de uitkomst is van de meting van de j^e eigenschap in de i^e groep.

Schrijft men de verkregen waarnemingsuitkomsten overzichtelijk in kolommen, dan kan men de eigenschappen aanduiden als kolomvectoren in de n-dimensionale ruimte.

Men gaat nu uit van het model :

$$\begin{bmatrix} \underline{Y}_{10} \\ \cdot \\ \cdot \\ \cdot \\ \underline{Y}_{io} \\ \cdot \\ \cdot \\ \cdot \\ \underline{Y}_{no} \end{bmatrix} \cong \alpha_1 \begin{bmatrix} y_{11} \\ \cdot \\ \cdot \\ \cdot \\ y_{i1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n1} \end{bmatrix} + \ldots + \alpha_j \begin{bmatrix} y_{1j} \\ \cdot \\ \cdot \\ \cdot \\ y_{ij} \\ \cdot \\ \cdot \\ \cdot \\ y_{nj} \end{bmatrix} + \ldots + \alpha_k \begin{bmatrix} y_{1k} \\ \cdot \\ \cdot \\ \cdot \\ y_{ik} \\ \cdot \\ \cdot \\ \cdot \\ y_{nk} \end{bmatrix} + \begin{bmatrix} \underline{e}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{e}_i \\ \cdot \\ \cdot \\ \cdot \\ \underline{e}_n \end{bmatrix} \qquad (1)$$

waarin de afwijking : $\underline{e}_i = \underline{Y}_{i0} - \alpha_1 y_{i1} - \ldots - \alpha_k y_{ik}$ als effect van onbekende invloeden kan worden gezien.

Verondersteld wordt, dat de stochastische variabelen $\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_n$ onderling onafhankelijk zijn, terwijl tevens wordt aangenomen dat :

$$\underline{e}_1 \cong \underline{e}_2 \cong \ldots \cong \underline{e}_n \cong \sigma \underline{\chi} \qquad (2)$$

In vectornotatie wordt (1) en (2) :

$$\underline{Y}_o = E(\underline{Y}_o) + \sigma \underline{\chi}_n \quad ,$$

waarin

$$E(\underline{Y}_o) = \sum_{j=1}^{k} \alpha_j y_j$$

In het algemeen wordt in het stel vectoren $y_1, \ldots, y_k$ een vector $(1, 1, \ldots, 1)$ opgenomen ter verantwoording van een constante (niveau). Stel dit is y_k. Aangezien de belangstelling niet in de eerste plaats uit zal gaan naar de constante, beschouwt men de componenten van de vectoren y_j in de (n-1)-dimensionale deelruimte loodrecht op de ruimte van het niveau. Deze componenten verkrijgt men door alle waarnemingen in een kolom te verminderen met het gemiddelde van die kolom.

Vervolgens wordt door schaalverandering de lengte der vectoren op de eenheid herleid.

Men beschouwt dus de gestandaardiseerde vectoren :

$$x_j = \frac{y_j - \bar{y}_j}{\sqrt{(y_j - \bar{y}_j)^2}}$$

De verwachtingswaarde van de gestandaardiseerde $\underline{y}_o$ wordt nu met nieuwe parameters:

$$E(\underline{x}_o) = \sum_{j=1}^{k-1} \beta_j x_j \qquad (3)$$

Is X de matrix van de gestandaardiseerde vectoren $(x_1, x_2, .. x_{k-1})$ en β de kolomvector $(\beta_1, \beta_2, .., \beta_{k-1})$, dan wordt (3) in matrixnotatie:

$$E(\underline{x}_o) = X\beta$$

Het verband tussen α_j en β_j wordt gegeven door de betrekking:

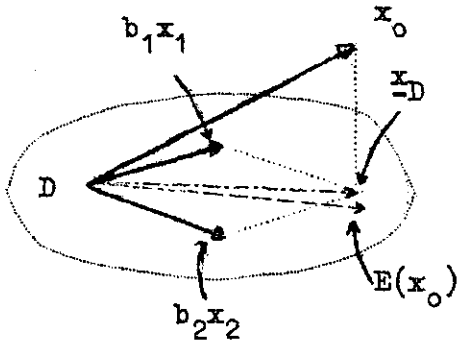
$$\alpha_j = \frac{\sqrt{(y_o - \bar{y}_o)^2}}{\sqrt{(y_j - \bar{y}_j)^2}} \beta_j \qquad j = 1, 2, \ldots, k-1$$

en

$$\alpha_k = \bar{y}_o - \alpha_1 \bar{y}_1 - \ldots - \alpha_{k-1} \bar{y}_{k-1}$$

De verwachtingswaarde van $\underline{x}_o$ is dus volgens (3) een vector in de (k-1) dimensionale deelruimte van de n-dimensionale ruimte, opgespannen door de vectoren $x_1, x_2, \ldots, x_{k-1}$. Schattingen van de grootheden β(regressiecoëfficiënten), σ^2 (restvariantie), r (meervoudige correlatie coëfficiënt) worden verkregen door loodrechte projectie (zie fig.1).

Figuur 1



D is een hypervlak opgespannen door de basisvectoren $x_1, x_2, \ldots, x_{k-1}$. Verondersteld wordt dat $E(\underline{x}_o)$ in D ligt. De beste schatter van $E(\underline{x}_o)$ verkrijgt men door loodrechte projectie van $\underline{x}_o$ op D. Is $\underline{x}_D$ die projectie, dan is de afstand van $\underline{x}_o$ tot D, gegeven door $\underline{x}_o - \underline{x}_D$ minimaal. Wegens standaardisatie komt de correlatiecoëfficiënt overeen met de cosinus van de hoek tussen $\underline{x}_o$ en $\underline{x}_D$ en aangezien de lengte van x_o is : $|\underline{x}_o| = 1$ is $|\underline{x}_D|$ een schatter van de meervoudige correlatiecoëfficiënt. Worden de parameters $b_1, b_2 \ldots$ opgenomen in een vector b, dan is Xb een lineaire combinatie van de basisvectoren, zoals bijvoorbeeld x_D. Nu moet $x_o - x_D$

of $\underline{x}_o$ - Xb loodrecht op D zijn en dus loodrecht op alle basisvectoren van D. Dan is, indien ${}^tX$ de getransponeerde is van X, de voorwaarde :

$$ {}^tX(\underline{x}_o - X\underline{b}) = 0 \qquad (4) $$

De gevraagde schatters volgen uit :

$$ \underline{b} = ({}^tXX)^{-1}\ {}^tX\underline{x}_o $$

$$ \text{cov}\ (\underline{b}) = ({}^tXX)^{-1}\ \frac{1-r^2}{n-k} $$

$$ \underline{S}(\sigma^2) = \frac{1-r^2}{n-k} $$

$$ r^2 = {}^t\underline{b}\ {}^tX\underline{x}_o $$

4. Selectie en eliminatie

Van beide in de inleiding genoemde methoden is het principe nu eenvoudig te zien.

4.1. Selectie

Bij selectie gaat men uit van die basisvector, waarop de projectie van $\underline{x}_o$ het langst is. Bij elke volgende stap kiest men dan die basisvector erbij, die de projectie van $\underline{x}_o$ op de deelruimte, opgespannen door de eerder gekozen vectoren en de nieuw erbij te kiezen vector, het meest doet toenemen.

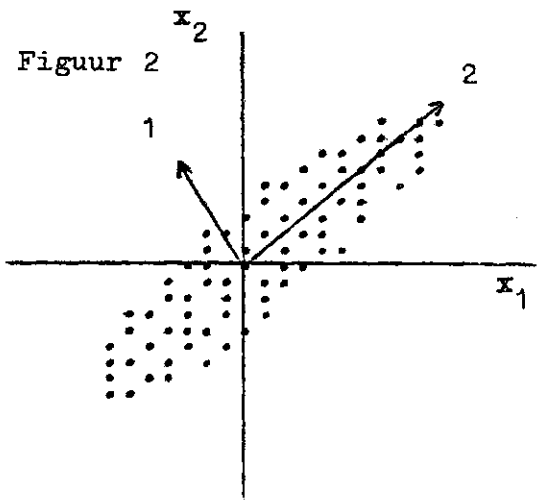
4.2. Eliminatie

Bij eliminatie gaat men uit van de projectie van $\underline{x}_o$ op de deelruimte opgespannen door alle verklarende vectoren. Men elimineert bij iedere stap die basisvector, die de projectie van $\underline{x}_o$ op de deelruimte, opgespannen door de resterende vectoren, het minst doet afnemen.

5. Richtingen met grote variatie

Stelt men zich een puntenwolk voor van de n punten in de (k-1)-dimensionale deelruimte opgespannen door de basisvectoren : (1, 0, 0, ...), (0, 1, 0, ...), ... , corresponderend met de (k-1) eigenschappen, dan zal bij een richting in die ruimte, waarin al die (k-1) eigenschappen niet of weinig gevarieerd hebben, geen effect op $\underline{x}_o$ waar te nemen zijn. In een richting, waarin de variatie groot is, is de mogelijkheid aanwezig een effect op $\underline{x}_o$ waar te nemen (zie fig. 2).

Figuur 2



In de figuur wordt de tweedimensionale deelruimte opgespannen door de basisvectoren (1, 0, 0), (0, 1, 0). De punten in het vlak zijn de projecties van de punten $(x_o, x_1, x_2)_i$ in de driedimensionale ruimte. In richting 1 zal weinig of geen effect waar te nemen zijn op x_o. In richting 2 is de mogelijkheid aanwezig.

Er wordt nu gezocht naar richtingen, waarin de variatie groot is. Een willekeurige richting in de (k-1)-dimensionale deelruimte wordt gegeven door de vector Xz. Men kan zo'n richting opvatten als een nieuwe eigenschap, die een lineaire combinatie is van de oude eigenschappen. Is de variatie in een richting groot, dan komt dit overeen met de uitspraak, dat de lengte van de vector Xz groot is. Eenvoudigheidshalve wordt het kwadraat van de lengte van de vector Xz de variantie van Xz (var(Xz)) genoemd.

De variantie van Xz is:

$$\text{var}(Xz) = {}^t z\, {}^t X X z = {}^t z A z$$

Daar de kolommen in X eenheidsvectoren zijn., is A de correlatiematrix; A is symmetrisch.

Aangezien de kentallen van de vector z in wezen verhoudingsgetallen zijn, is het maximaliseren van var (Xz) alleen zinvol, als aan z een beperking wordt opgelegd, waarvoor gekozen wordt de nevenvoorwaarde , dat z ook een eenheidsvector zal zijn, in matrixnotatie:

$${}^t z z - 1 = 0$$

Wordt een parameter λ als Lagrange vermenigvuldiger ingevoerd, dan wordt gevraagd naar de stationnaire punten van de functie:

$$\phi = {}^t z A z - \lambda({}^t z z - 1)$$

welke worden gevonden uit:

$$\frac{\partial \phi}{\partial z} = 2Az - 2\lambda z = 0$$

en dus

$$Az = \lambda z$$

met

$$^{t}zz = 1$$

Hieruit volgt dat z een eigen vector is van A met bijbehorende eigenwaarde λ. De eigenwaarden zijn de wortels uit de vergelijking van de graad (k-1):

$$|A - \lambda I| = 0$$

met hoogstens k-1 reële wortels.

6. Eigenschappen van eigen vectoren en eigenwaarden

6.1. Bij een symmetrische matrix A zijn alle eigenwaarden reëel.
Bewijs: Stelt $\bar{z}$ de complex toegevoegde van een vector (getal) $z = x + iy$ voor, dan geldt voor alle z:

$$\overline{\bar{z}.Az} = \bar{\bar{z}}.\overline{Az} = z.A\bar{z} = \bar{z}.Az$$

of wel

$\bar{z}.Az$ is reëel.

Indien z een eigenvector van A is, dan is:

$$\bar{z}.Az = \bar{z}.\ \lambda z = \lambda\bar{z}.z$$

met zowel $\bar{z}.Az$ als $\bar{z}.z$ reëel, dus λ is reëel.
Gevolg: de vergelijking van de graad k-1: $|A - \lambda I| = 0$ heeft k-1 reële wortels.

6.2. De functiewaarde in een stationnair punt is, uitgaande van ^{t}zAz en met $^{t}zz = 1$:

$$\text{var}(Xz) = {}^{t}zAz = z.\lambda z = \lambda z.z = \lambda \qquad (5)$$

Gevolg: alle wortels zijn groter dan of gelijk aan nul.

6.3. De eigen vectoren van een symmetrische matrix A zijn bij verschillende eigenwaarden onderling loodrecht.
Bewijs: Stel $\lambda_i \neq \lambda_j$ zijn twee verschillende eigenwaarden, dan volgt:

$$z_i.Az_j = z_i.\lambda_j z_j = \lambda_j\ z_i.z_j$$

$$z_j.Az_i = z_j.\lambda_i z_i = \lambda_i\ z_j.z_i$$

en wegens identiteit der linkerleden geldt dan :

$$(\lambda_i - \lambda_j)\, z_i \cdot z_j = 0$$

dus

$$z_i \cdot z_j = 0 \qquad (6)$$

6.4. De som der wortels is het spoor van de matrix A.

Bewijs : Uitwerken en rangschikken van de producten van de elementen van de determinant $|A - \lambda I| = 0$ geeft :

$$(-1)^{k-1} \lambda^{k-1} + (a_{11} + a_{22} + \dots + a_{(k-1)(k-1)})\, (-1)^{k-2} \lambda^{k-2} + \dots = 0$$

Hieruit volgt voor de som der wortels :

$$\sum_{i=1}^{k-1} \lambda_i = \sum_{i=1}^{k-1} a_{ii} \qquad \text{(het spoor van de matrix).}$$

Aangezien $a_{11} = a_{22} = \dots = a_{(k-1)(k-1)} = 1$ is :

$$\sum_{i=1}^{k-1} \lambda_i = k-1 \qquad (7)$$

6.5. Indien Xz_i en Xz_j richtingen zijn met z_i en z_j eigenvectoren bij verschillende eigenwaarden λ_i en λ_j, dan geldt met toepassing van (5) en (6) :

$$Xz_i \cdot Xz_j = {}^t z_i\, {}^t X X z_j = z_i \cdot A z_j = z_i \cdot \lambda_j z_j = \lambda_j z_i \cdot z_j = 0 \qquad (8)$$

zodat Xz_i en Xz_j onderling loodrecht zijn.

7. <u>Keuze van belangrijke richtingen</u>

Indien een richting Xz_i (nieuwe eigenschap) invloed heeft op $\underline{x}_o$ dan is de projectie van $\underline{x}_o$ op die richting groot.

De projectie van $\underline{x}_o$ op de richting Xz_i kan voorgesteld worden door : $d_i Xz_i$, waarbij d_i gevonden wordt uit de voorwaarde, dat de verschilvector $x_o - d_i Xz_i$ loodrecht staat op die richting en dus overeenkomstig (4) :

$${}^t z_i\, {}^t X (x_o - d_i Xz_i) = z_i \cdot {}^t X x_o - d_i\, {}^t z_i\, {}^t X X z_i$$

wat volgens (5) gelijk is aan :

$$z_i \cdot {}^tXx_o - d_i\lambda_i = 0$$

waaruit volgt dat :

$$d_i = \frac{z_i \cdot {}^tXx_o}{\lambda_i}$$

Het kwadraat van de lengte van de projectie is dan onder toepassing van (5) :

$$(d_iXz_i)^2 = d_i^2 \; {}^tz_i \; {}^tXXz_i = d_i^2\lambda_i$$

7.1. Toets :

Indien $\underline{x}_T$ de component is van $\underline{x}_o$ in de ruimte loodrecht op de deelruimte opgespannen door de vectoren $x_1, x_2, \ldots, x_{k-1}$ (fig. 1), dan is :

$$\underline{x}_T^2 \cong \sigma^2 \underline{\chi}^2_{n-k}$$

Verder geldt, indien B de 1-dimensionale deelruimte is, opgespannen door de vector Xz_i, voor de variantie van de projectie van $\underline{x}_o$ op B, voorgesteld door $\underline{x}_{oB}$:

$$(\underline{x}_{oB} - E(\underline{x}_{oB}))^2 \cong \sigma^2 \underline{\chi}^2_1$$

Bovendien zijn de deelruimten B en T orthogonaal, zodat $\underline{x}_T$ en $\underline{x}_B$ onderling loodrecht zijn.

Dan geldt onder de nulhypothese (H_o : $E(\underline{x}_{oB}) = 0$ of $E(\underline{d}_i) = 0$) :

$$\frac{\underline{x}^2_{oB}}{\underline{x}^2_T/n-k} = \frac{\underline{d}_i^2\lambda_i}{\underline{x}^2_T/n-k} \cong \frac{\sigma^2 \underline{\chi}^2_1}{\sigma^2 \underline{\chi}^2_{n-k}/n-k} \cong \underline{F}^1_{n-k}$$

De nulhypothese wordt verworpen indien :

$$P(\underline{F}^1_{n-k} > \frac{d_i^2\lambda_i}{x^2_T/n-k}) < \alpha \quad ,$$

waarin α een gekozen overschrijdingskans is.

Op grond van deze toets kunnen belangrijke richtingen Xz_i geselecteerd worden.

8. Keuze van de oude eigenschappen

Alvorens over te gaan tot de vraag, welke van de oorspronkelijke vectoren in de basis gekozen moeten worden, is het nodig de componenten van die vectoren langs de gekozen richtingen Xz_i te beschouwen.

Men vindt de componenten van de oorspronkelijke vectoren x_j langs de richtingen Xz_i, algemeen voorgesteld door $a_{ji}Xz_i$, op een overeenkomstige wijze als (4) uit de gezamenlijke voorwaarden, geldend voor elke x_j uit X:

$$ {}^t z_i {}^t X(X - Xz_i {}^t a_i) = {}^t z_i {}^t XX - {}^t z_i {}^t XXz_i {}^t a_i = 0 $$

en door achter vermenigvuldiging met z_i, onder toepassing van (5):

$$ {}^t z_i {}^t XXz_i - {}^t z_i {}^t XXz_i {}^t a_i z_i = \lambda_i - \lambda_i {}^t a_i z_i = 0 $$

waaruit volgt:

$$ {}^t a_i z_i = 1 $$

of

$$ a_i = z_i $$

De component van x_j langs de richting Xz_i wordt dus gegeven door de vector $z_{ji}Xz_i$.

Het kwadraat van de lengte van de component van x_j langs Xz_i is dan:

$$ (z_{ji}Xz_i)^2 = z_{ji}^2 \, {}^t z_i {}^t XXz_i = z_{ji}^2 \lambda_i $$

Het kwadraat van de lengte van de component van x_j in de deelruimte opgespannen door de gekozen richtingen is dan wegens onderlinge loodrechtheid der richtingen (8):

$$ \sum_i z_{ji}^2 \lambda_i, \text{ voor die i waarvan de richtingen gekozen zijn.} $$

Is $\sum_i z_{ji}^2 \lambda_i$ klein, dan ligt de grootste component van x_j in de deelruimte opgespannen door de niet belangrijke richtingen, zodat de vector niet in aanmerking komt voor opname in de nieuwe (beperkte) basis.

Stel x_1 en x_2 zijn twee oude eigenschappen, waarvoor gevonden wordt dat $z_{1i}^2 \lambda_i \approx z_{2i}^2 \lambda_i$, dan verklaren beide vectoren ongeveer evenveel, zodat een keuze uit die twee vectoren gedaan kan worden.

Kiest men langs elke belangrijke richting een oude eigenschap, waarvan de component in die richting groot is, dan verkrijgt men zodoende een aantal combinaties, die aan het gestelde doel kunnen voldoen.

9. Voorbeeld

Gekozen is een voorbeeld op technologisch gebied, omdat op dit voorbeeld reeds twee van de drie voorstellen zijn toegepast [3], terwijl als andere voordelen, de bijna afhankelijkheid van de basisvariabelen en het overzichtelijk aantal variabelen genoemd kunnen worden. Men kan echter voorbeelden bedenken op cultuurtechnisch terrein waarop genoemde werkwijze kan worden toegepast.

Het voorbeeld is uit Hald [2] pag. 647-650 en het betreft de benodigde warmte in calorieën per gram cement voor het harden van klinkers als functie van de samenstelling.

De variabelen zijn:

y_0 = benodigde warmte in calorieën per gram cement

y_1 = gewichtspercentage 3 $CaO.Al_2O_3$

y_2 = " 3 $CaO.SiO_2$

y_3 = " 4 $CaO.Al_2O_3.Fe_2O_3$

y_4 = " 2 $CaO.SiO_2$

Bij de gegevens (zie tabel 1) blijkt, dat $95\% \leqslant \sum_{i=1}^{4} y_i \leqslant 99\%$ zodat een van de vier verklarende variabelen bij benadering **een** lineaire combinatie is van de andere drie.

Tabel 1

De gegevens

y_0	y_1	y_2	y_3	y_4	$\sum_{i=1}^{4} y_i$
78.5	7	26	6	60	99
74.3	1	29	15	52	97
104.3	11	56	8	20	95
87.6	11	31	8	47	97
95.9	7	52	6	33	98
109.2	11	55	9	22	97
102.7	3	71	17	6	97
72.5	1	31	22	44	98
93.1	2	54	18	22	96
115.9	21	47	4	26	98
83.8	1	40	23	34	98
113.3	11	66	9	12	98
109.4	10	68	8	12	98

In tabel 2 wordt de matrix ^tXX alsmede de kolom ^tXx_o gegeven.

Tabel 2

De matrix ^tXX en de kolom ^tXx_o

x_i	1	2	3	4	0
1	1.00000	0.22858	-0.82415	-0.24544	0.73072
2		1.00000	-0.13924	-0.97295	0.81625
3			1.00000	0.02953	-0.53466
4				1.00000	-0.82130

De kwadraten van de projecties van x_o op alle mogelijke deelruimten welke de vectoren x_1 en/of x_2 en/of x_3 en/of x_4 kunnen opspannen worden gegeven in tabel 3, evenals de volgorden welke ontstaan bij selectie en eliminatie.

Tabel 3

De kwadraatprojecties en volgorden bij selectie en eliminatie

vectoren in de basis	kwadraat projectie	selectie	eliminatie	volgorde selectie	volgorde eliminatie
1	0.53395				
2	0.66626		0.66626		4
3	0.28586				
4	0.67453	0.67453		1	
1, 2	0.97867		0.97867		3
1, 3	0.54818				
1, 4	0.97247	0.97247		2	
2, 3	0.84702				
2, 4	0.68006				
3, 4	0.93527				
1, 2, 3	0.98227				
1, 2, 4	0.98233	0.98233	0.98233	3	2
1, 3, 4	0.98126				
2, 3, 4	0.97277				
1, 2, 3, 4	0.98237	0.98237	0.98237	4	1

In tabel 4 worden de eigenwaarden λ_i, eigen vectoren z_i, de coëfficiënten d_i, de kwadraat projecties $d_i^2\lambda_i$ alsmede de waarden van F_8^1 en P gegeven. De 4e eigen vector is niet berekend, aangezien de kwadraat projectie $d_4^2\lambda_4$ te berekenen is volgens:

$$d_4^2\lambda_4 = 0.9824 - \sum_{i=1}^{3} d_i^2\lambda_i$$

Uit de tabel blijkt, dat indien $\alpha = 0{,}05$ wordt gekozen, de richtingen Xz_1 en Xz_3 belangrijk zijn.

Tabel 4

De eigenwaarden, eigen vectoren en toets grootheden

λ	2.23576	1.57608	0.18661	0.00155
x \ z	1	2	3	4
1	0.4758	0.5091	0.6755	-
2	0.5640	-0.4137	-0.3145	-
3	-0.3939	-0.6051	0.6376	-
4	-0.5480	0.4510	-0.1954	-
di	0.6569	-0.0080	0.3029	-
$d_i^2\lambda i$	0.9648	0.0001	0.0171	0.0003
F_8^1	438.6	0.04	7.8	0.15
P	< 5%	> 25%	< 5%	> 25%

In tabel 5 worden de kwadraten van de lengte van de componenten der oorspronkelijke vectoren in de gekozen richtingen gegeven.

Een dergelijke tabel wordt soms wel met de naam aspectentabel voor de basisvectoren aangeduid.

Tabel 5

De kwadraatprojecties der oude vectoren

	Xz_1	Xz_3	Σ
x_1	0.5062	0.0852	0.5914
x_2	0.7111	0.0185	0.7296
x_3	0.3469	0.0759	0.4228
x_4	0.6715	0.0071	0.6786
$\Sigma = \lambda_i$	2.2357	0.1867	2.4224

Op grond van de somkolom in tabel 5 blijkt dat x_3 een grotere component heeft in de restruimte van de 4-dimensionale ruimte loodrecht op en dus onafhankelijk van die, opgespannen door de gekozen richtingen Xz_1 en Xz_3. De vector x_3 wordt dan ook buiten beschouwing gelaten.

Verder blijkt dat x_2 en x_4 grote componenten hebben in de richting Xz_1 en de vector x_1 de enige overblijvende vector is die in aanmerking komt in de richting Xz_3.

Zodoende komt men tot de combinaties x_1 en x_2 of x_1 en x_4, welke als kleiner set kunnen dienen.

10. Samenvatting en conclusies

Men kan door middel van de F-toets aantonen, dat in het gekozen voorbeeld de kwadraten van de projecties van $\underline{x}_o$ op de deelruimten opgespannen door x_i en x_j alle nog vergroot kunnen worden, behalve, indien $i = 1$ en $j = 2$ of $i = 1$ en $j = 4$. De combinaties x_1, x_2 en x_1, x_4 blijken dus te voldoen aan de gestelde vraag.

Volgens de selectie methode wordt slechts de combinatie x_1, x_4 gevonden, terwijl bij eliminatie de combinatie x_1, x_2, overblijft.

Worden beide methoden gebruikt, dan vindt men in dit geval beide combinaties. Men kan echter voorbeelden bedenken, waarbij meer combinaties mogelijk zijn, zodat dan enige combinaties niet worden gevonden.

In het gebruikte voorbeeld voldoet de methode, welke gebruik maakt van richtingen van grote variatie, aangezien beide combinaties als belangrijk worden aangewezen.

Men kan nu op bijvoorbeeld technologische gronden een der combinaties aanwijzen als meest geschikt voor het gestelde doel, namelijk $\underline{x}_o$ zo efficiënt mogelijk bepalen.

Geraadpleegde literatuur

1. L.C.A.Corsten: Selectie van variabelen in een regressieprobleem (stencil, Wageningen zonder jaar)
2. A.Hald : Statistical Theory with Engineering Application(1952)
3. H.C.Hamaker : On Multiple Regression Analysis.Statistica Jrg.16 nr.1