K O N I N K L I J K N E D E R L A N D S
M E T E O R O L O G I S C H I N S T I T U U T

D e B i l t

WETENSCHAPPELIJK RAPPORT
W.R. 75-1

P.J. Rijkoort

Een algemeen lineair statistisch model voor de bepaling van relaties tussen een aantal grootheden.

De Bilt, 1975

Publikatienummer: W.R. 75-1 (Stat. Bur.)

U.D.C.: 519.2

EEN ALGEMEEN LINEAIR STATISTISCH MODEL VOOR DE BEPALING VAN RELATIES TUSSEN EEN AANTAL GROOTHEDEN

Samenvatting.

Wetenschappelijk onderzoek is voor een groot deel het zoeken naar relaties tussen één bepaald verschijnsel en een aantal andere verschijnselen.

Mathematisch geformuleerd, hoe ziet de functie eruit in: $y_j = f(x_{ij}; \alpha_r)(i = 1,\ldots,k;\ r = 1,\ldots,t;\ j = 1,\ldots,N)$ waarin y_i resp. x_{ij} kwantificaties van de verschijnselen voorstellen en α_r parameterwaarden zijn. Het probleem is in zijn algemeenheid slechts oplosbaar als men ten aanzien van de vorm van de functie een bepaalde onderstelling invoert, waarna het verder gaat om de schatting van de waarden van de parameters α_r en het geven van een statistische uitspraak betreffende de realiteit van de veronderstelde relatie.

In die gevallen, waarin r = i en t = k kan f vaak als een lineaire functie in de α_i worden beschouwd:
$y_j = \sum_{i=1}^{k} \alpha_i x_{ij}$.

In principe zijn in dit model zeer veel relaties samengevat. Ten eerste omdat de x_{ij} onderling afhankelijk mogen zijn, bijvoorbeeld $x_{3j} = x_{1j}x_{2j}^2$ enz. Ten tweede omdat de x_{ij} zowel continue variabelen kunnen voorstellen als ook z.g. indicatorvariabelen. Dit laatste wil zeggen, dat slechts de waarde 1 of 0 (voor iedere j is één $x_{ij} = 1$ en de overigen zijn 0) worden aangenomen, wat er op neer komt, dat de y_j in klassen i worden ingedeeld. In geval van continue variabelen heet de statistische behandeling regressie-analyse; in geval van indicatorvariabelen variantie-analyse. De gemengde vorm heet covariantie-analyse.

Voor de oplossing van het hier aangeduide probleem zijn mathematische-statistische methoden ontwikkeld die o.a. te vinden zijn in de boeken van Scheffé en van Dempster (zie literatuur).

In dit rapport wordt de algemene theorie volgens bovengenoemde boeken zo kort mogelijk behandeld in hoofdstuk 1, terwijl daarna in 2 numerieke toepassingen worden gegeven. De hoofdstukken 3 en 4 behandelen een bijzonder geval van regressie-analyse namelijk een methode om "met een computer" tot de optimale keuze van r variabelen x_i uit de totaal k variabelen te komen ter verklaring van de variatie van y. Deze methode is te vinden in een artikel van Efroymson.

De bedoeling van dit rapport is de behandelde materie toegankelijk te maken voor diegenen, die niet de tijd hebben de uitvoerige literatuur door te werken.

Summary.

Scientific research is mainly conserned with attempts to find relations between one phenomenon and different other phenomena.

Mathematically formulated, what can be said about: $y_j = f(x_{ij}; \alpha_r)(i = 1,...k; r = 1,...,t; j = 1,...,N)$ in which y_j resp. x_{ij} represent the phenomena and the α_r are parameters to be chosen. In general a solution is only possible if some form of the function is postulated. The problem is reduced then to finding estimates of the α_r and to giving a statistical judgement with regard to the reality of the relation so formed.

If $r = i$ and $t = k$ it is often possible to use a lineair relation (lineair with regard to the α_i):

$$y_j = \sum_{i=1}^{k} \alpha_i x_{ij}.$$

In principle many sorts of relations are contained in this model. Firstly because the x_{ij} need not to be independent of each other, for instance $x_{3j} = x_{1j}x_{2j}^2$ ect. may be possible and secondly because the x_{ij} can be continuous variables as well as so called, indicator variables. The latter means that for each j only $x_{ij} = 1$ and all other $x_{ij} = 0$; so that the y_j are arranged in different classes. The statistical treatment is called regression analysis in the case of continuous variables and analysis of variance in the case of indicator variables. Analysis of covariance is a mixed form.

The mathematical statistical solution of the problem can be found in different textbooks for instance in those written by Scheffé and by Dempster (see literature).

In the present report the general theory according to Scheffé and Dempster is treated shortly in chapter 1, while chapter 2 gives some numerical examples. The chapter 3 and 4 contain a treatment of a special regression problem viz. the solution (which a computor) of the problem of finding an optimum choice of only r from the k variables x_j to explain the variation of y. This method has been developted by Efroymson, using results of Orden (see literature).

This report is meant to give an introduction to the method to be used by investigators who do not have time to study the books of Scheffé, Dempster and other themselves.

Een algemeen lineair statistisch model voor de bepaling van relaties tussen een aantal grootheden.

Dr. P.J. Rijkoort

Algemene inleiding.

Een groot gedeelte van het onderzoek dat in wetenschappelijke instellingen wordt uitgevoerd, bestaat in het zoeken naar het verband tussen één bepaalde grootheid (y) en één of meer grootheden (x_i).

De statistische behandeling van dergelijke problemen, voor zover die in dit verslag zal worden uiteengezet, zal zich beperken tot die gevallen waarbij het verband de volgende lineaire vorm heeft:

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots\ldots + \beta_p x_p \qquad (0.1)$$

De lineairiteit van het model is beperkt tot de parameters β_i en niet wezenlijk tot de x_i, daardoor kunnen met dit model meer relaties beschreven worden dan men in eerste instantie zou denken. Naast die gevallen waarin inderdaad een rechtstreekse lineaire relatie bestaat, kan men voor gevallen waarbij dit niet zo is, door transformaties toch tot bovenstaande vorm komen; voor een exponentieel verband bijvoorbeeld kan dit met een logarithmische transformatie. Bovendien kan men een gecompliceerde vorm van x_i's opvatten als een nieuwe x_i; het kwadratische verband $y = ax^2 + bx + c$ kan bijvoorbeeld ook in de vorm (0.1) gezien worden als men x^2 door x_1 en x zelf door x_2 voorstelt; verder resulteert het opvatten van x_3 als de constante 1 in het opnemen van een constante in de regressievergelijking.

De behandeling van de in het voorgaande ter sprake komende problemen, waarbij de x_i's evenals y continue variabele grootheden zijn, wordt enkelvoudige regressie genoemd als er één x is, of meervoudige (multiple) regressie als er meerdere x_i's zijn.

Het is ook mogelijk een ander soort problemen in de vorm (0.1) te brengen n.l. die waarbij y een grootheid is die onder verschillende omstandigheden wordt waargenomen en waarbij het er dan om gaat vast te stellen of y zich verschillend gedraagt onder deze verschillende omstandigheden. Als de verschillende omstandigheden met klassen A, B, C enz. worden aangeduid dan kan gesteld worden :

$x_{(A)} = 1$; $x_{(B)} = 0$ enz. als $y \in A$ (y behoort tot A)

$x_{(A)} = 0$; $x_{(B)} = 1$; $x_{(C)} = 0$ enz. als $y \in B$

enz.

De grootheid x wordt nu een "indicator" variabele genoemd. Onze aanpak van dit soort problemen zal blijken tot de bekende "variantie-analyse" te voeren.

In geval er zowel continue als indicator variabelen in een probleem voorkomen, wordt de statistische behandeling in de literatuur met covariantie-analyse aangeduidt.

De oplossing van de gestelde problemen kan men in diverse statistische handboeken vinden. Een vrij algemene behandeling, later speciaal op de variantie-analyse gericht, is te vinden in H. Scheffé: "Analysis of variance" Wiley - New York 1959 - 1967, terwijl een algemene behandeling met matrices en verder toepassing op de regressie-analyse uitvoerig wordt uiteengezet in A.P. Dempster: "Elements of continuous multivariate analysis" - Addison - Wesley - 1969.

Hoewel beide boeken niet bepaald bijzonder moeilijk zijn, is het toch voor onderzoekers die niet speciaal statistisch zijn onderlegd, bezwaarlijk eerst deze boeken te moeten doorwerken, alvorens de resultaten te kunnen toepassen. Dit is vooral zo, omdat de methode van behandeling in deze boeken, zoals vaak in de mathematische literatuur, zo is, dat eerst allerlei stellingen en hulpstellingen worden ontwikkeld en bewezen, waarvan de betekenis pas veel later duidelijk wordt.

Men kan natuurlijk wel volstaan met korte recepten van statistische methoden die in artikelen in de statistische literatuur wel te vinden zijn, maar dan kan de onderzoeker niet beoordelen of er correct gewerkt wordt en bovendien is het bij de interpretatie van de resultaten veel beter als men inzicht heeft in de achtergrond van de gebruikte methode, in het hoe en waarom.

Om deze redenen zal in dit verslag getracht worden een zoveel mogelijk recht op het doel afgaande uiteenzetting van de algemene theorie te geven. Ter toelichting zullen één geval van variantie-analyse en één geval van regressie-analyse worden behandeld. Bovendien zal een artikel waarin multiple regressie als een soort recept wordt behandeld nader worden beschouwd.

Uiteraard moet voor het volgen van de behandeling van de theorie in dit rapport wel enige statistische basiskennis aanwezig zijn en tevens bekendheid met het werken met vectoren in een n-dimensionale ruimte en met matrices.

In eerste instantie zal het boek van Scheffé worden gevolgd en later Dempster. De notatie zal uniform worden gemaakt.

Dit rapport is bedoeld als basisrapport, waarin alleen de algemene zaken ter sprake komen. Later zal getracht worden in vervolgrapporten op bijzondere gevallen in te gaan.

1. Theoretische behandeling van een algemeen "multivariabele" model

1.1. Inleiding.

In de algemene inleiding is de lineaire vorm (0.1) gegeven voor het onderstelde verband tussen een variabele y, die afhankelijk variabele zal worden genoemd en variabelen x_i, die onafhankelijke variabelen heten. (N.B. de uitdrukking onafhankelijk met betrekking tot de x_i's is niet helemaal correct, er kan onderlinge afhankelijkheid tussen de x_i voorkomen; onafhankelijk betekent hier alleen dat het geheel van de x_i's primair varieert en dat de y als gevolg daarvan varieert).

Waarnemingen van y zullen meestal niet exact aan (0.1) voldoen maar met een zekere fout ε behept zijn. Daarom zal als algemeen model, waarvoor de notatie Ω zal worden gebruikt, worden ingevoerd:

$$\Omega: \; y_k = \sum_{i=1}^{p} \beta_i x_{ik} + \varepsilon_k \qquad (1.1.1)$$

We onderstellen dat er n waarnemingen zijn van p + 1 grootheden y en x_i. Verder nemen we aan dat ε_k onafhankelijk is van ε_m voor iedere k en m en dat ε_k voor iedere k een normale verdeling met verwachtingswaarde nul ($E(\varepsilon_k) = 0$) en standaarddeviatie σ volgt. Dit wordt als: ε_k is $N(0,\sigma)$, genoteerd. De statistische verwachtingswaarde van y_k is derhalve:

$$E(y_k) = \sum_{i=1}^{p} \beta_i x_{ik} \qquad (1.1.2)$$

1.2. De beste schattingen van de parameters.

Het eerste probleem is nu: Wat zijn voor gegeven waarden van y_k en x_{ik} voor $i = 1,\ldots,p$; $k = 1,\ldots,n$ de beste schattingen

voor de parameters β_i? Er zijn voor de oplossing verschillende methoden beschikbaar, die echter alle aequivalent zijn. De bekendste is de kleinste kwadratenmethode. Een andere methode berust op het principe van de z.g. waarschijnlijkheidsfunctie voor een gegeven steekproef; maximaliseren van deze functie geeft z.g. "Maximum Likelihood" schatters. In dit rapport wordt er de voorkeur aangegeven de afleiding van de formule voor de schattingen van de β_i te baseren op een geometrische voorstelling. De waarden y_k $(k = 1,\ldots,n)$ zullen daartoe als componenten van een vector $\vec{y}$ in een n-dim.ruimte R_n worden opgevat. Evenzo zijn er p-vectoren $\vec{x}_i (i = 1,\ldots,p)$ in R_n die onafhankelijk variabelen voorstellen. Deze p-vectoren $\vec{x}_i$ vormen een deelruimte van R_n die we met V_p zullen aangeven en die p of minder dimensies kan hebben. Het is namelijk mogelijk dat er tussen enkele der $\vec{x}_i$'s lineaire afhankelijkheid bestaat en dan zijn er dus minder dan p onafhankelijke vectoren en heeft de deelruimte V minder dan p dimensies. In principe is de algemene theorie zo te ontwikkelen, dat met deze mogelijkheid wordt rekening gehouden. In dit rapport wordt echter gemakshalve ondersteld, dat V de dimensie p heeft.

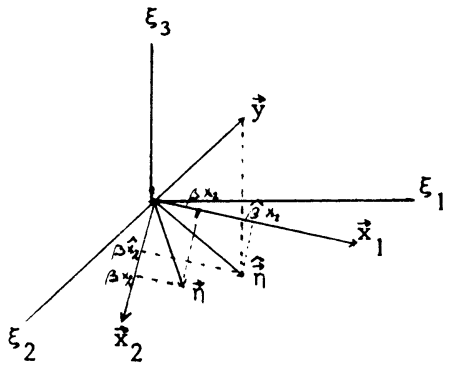
Zoals in (1.2) is gesteld vormen de sommen $\sum_{i=1}^{p} \beta_i x_{ik}$ de verwachtingswaarden van y_k; deze kunnen ook als componenten van een vector in R_n beschouwd worden die met $\vec{\eta}(=E(\vec{y}))$ zal worden aangegeven. Aangezien de componenten van deze vector $\vec{\eta}$ geheel zijn opgebouwd uit lineaire combinaties van de componenten van de vectoren $\vec{x}_i$ zal de vector $\vec{\eta}$ dus ook in V_p liggen. De vector $\vec{y}$ stelt een bepaalde steekproef $(y_1,\ldots,y_n)$ voor, een andere steekproef van n-waarden $(y_1,\ldots,y_n)$ zal een andere vector $\vec{y}$ leveren met dezelfde verwachtingswaarde $\vec{\eta}$.

De vector $\vec{y}$ is een stochastische vector. De verschillende vectoren $\vec{y}$ kunnen in principe overal in R_n liggen, maar zullen in het algemeen geconcentreerd zijn rond $\vec{\eta}$.

Aangezien er maar één steekproef $(y_1,\ldots,y_n)$ en dus één vector $\vec{y}$ beschikbaar is kan $\vec{\eta}$ niet precies bepaald worden. Het enige wat gedaan kan worden is een vector $\vec{\eta}_0$ bepalen die zo dicht mogelijk bij $\vec{y}$ ligt maar wel tot V_p behoort. $\vec{y}$ zelf ligt vrijwel zeker niet in V_p, hoewel dit niet absoluut onmogelijk is; hiervoor zou moeten gelden: $\varepsilon_k = 0$ voor alle k, en de kans hierop is wel te verwaarlozen; en bovendien is er dan statistisch gezien geen probleem meer.

Nu is de vector in V_p die zo dicht mogelijk bij $\vec{y}$ ligt de projectie van $\vec{y}$ op V_p, deze zal als schatting van $\vec{\eta}$ worden genomen en met $\hat{\vec{\eta}}$ worden aangeduid.

fig. 1.



Ter illustratie fig. 1 waarbij om een tekening mogelijk te maken i tot 2 is beperkt en n tot 3. Het assenkruis van R_n is zo gekozen, dat ξ_1 en ξ_2 het vlak van $\vec{x}_1$ en $\vec{x}_2$ aangeven. Het assenkruis (de basis in n-dim.geometrie), waarop de componenten van $\vec{y}$ en $\vec{x}_i$ zijn aangegeven, kan een ander zijn. De vector $\vec{\eta}$ is in werkelijkheid onbekend.

Er geldt nu $\vec{\eta} = \sum_{i=1}^{p} \beta_i \vec{x}_i$ en $\hat{\vec{\eta}} = \sum_{i=1}^{p} \hat{\beta}_i \vec{x}_i$, waarin de $\hat{\beta}_i$ de gezochte schattingen van de regressiecoëfficiënten zijn. Verder is $\vec{y} - \hat{\vec{\eta}} \perp V_p$ en dus $\perp \vec{x}_i$ voor $i = 1,\ldots,p$. De verschillen $\vec{y} - \vec{\eta}$ en $\vec{y} - \hat{\vec{\eta}}$ stellen vectoren voor waarvan de componenten resp. ε_k en $\hat{\varepsilon}_k$ kunnen worden genoemd. Het kwadraat van de lengten van deze vectoren is dus resp. $\sum_{k=1}^{n} \varepsilon_k$ en $\sum_{k=1}^{n} \hat{\varepsilon}_k^2$.

Wegens $(\vec{y} - \hat{\vec{\eta}}) \perp V_p$ is $\sum_{k=1}^{n} \hat{\varepsilon}_k^2$ de kleinste waarde die $\sum_{k=1}^{p} \varepsilon_k^2$ kan aannemen als $\vec{\eta}$ over V_p varieert. Nu is

$\sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (y_k - \sum_{i=1}^{p} \beta_i x_{ik})^2$; als dit $\mathcal{S}_{(\vec{y},\vec{\beta})}$ genoemd wordt dan zijn de $\hat{\beta}_i$ die β_i waarden die $\mathcal{S}_{(\vec{y},\vec{\beta})}$ minimaal maken, met andere woorden, de geometrische oplossing van het probleem is identiek aan de kleinste kwadratenmethode. $\mathcal{S}_{(\vec{y},\hat{\vec{\beta}})}$ zal als $\mathcal{S}_\Omega$ worden genoteerd.

Het minimaliseren van de kwadraatsom geeft de zogenaamde normaalvergelijkinger. Dus:

$\frac{\partial \mathcal{S}(\vec{y},\vec{\beta})}{\partial \beta_i} = 0$ geeft

$$\sum_{k=1}^{n} y_k x_{jk} = \sum_{k=1}^{n} (x_{jk} \sum_{i=1}^{p} \hat{\beta}_i x_{ik}) = \sum_{i=1}^{p} (\hat{\beta}_i \sum_{k=1}^{n} x_{jk} x_{ik}) \qquad (1.2.1)$$

voor $j = 1,\ldots,p$. Hieruit kunnen de $\hat{\beta}_i$ worden berekend.

Men kan de geometrische oplossing ook als volgt formuleren: De vector $\vec{y} \in R_n$ die de waarnemingen voorstelt wordt ontbonden in een vector $\vec{y}_{//}$ die $// V_p$ is en die de beste aanpassing van de waarnemingen aan het model weergeeft $(\vec{y}_{//} \equiv \hat{\vec{\eta}})$ en een vector $\vec{y}_\perp$ die $\perp V_p$ is en waarvan de lengte een maat is voor de afwijking tussen waarnemingen en model $\vec{y}_\perp \equiv \vec{y} - \hat{\vec{\eta}}$.

1.3. Toetsing van hypothesen t.a.v. de parameters.

In het voorgaande zijn de beste schattingen $\hat{\beta}_i$ van de β_i gevonden in de onderstelling dat het algemene model volledig geldt. De β_i zullen in het algemeen onderling ongelijk zijn. Het is echter mogelijk dat van de β_i een aantal nul is of dat er enige aan elkaar gelijk zijn.

De vraag is dan of deze onderstellingen ten aanzien van de $\hat{\beta}_i$ getoetst kunnen, met andere woorden kunnen de gevonden $\hat{\beta}_i$, de onnauwkeurigheid van deze schattingen in aanmerking genomen, in overeenstemming zijn met bepaald veronderstellingen ten aanzien van de β_i. Dit kan wat algemener gesteld worden, namelijk dat er q onderling onafhankelijke lineaire functies Ψ_j van de β_i's zijn. Als nulhypothese H_o wordt dan ingevoerd:

$$H_o: \Psi_j(\beta_1,\ldots,\beta_p) = 0 \qquad j = 1,\ldots,q \qquad (1.3.1)$$

De onderstelling $\beta_3 = \beta_4$ kan bijvoorbeeld ingevoerd worden met $\Psi_1 = \beta_3 - \beta_4$ enz.

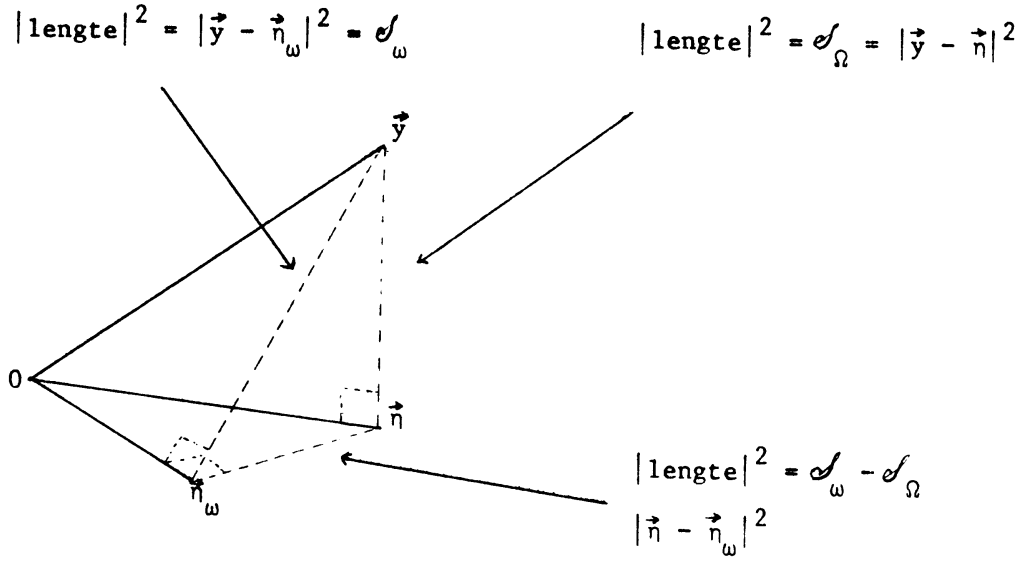
Onder deze nulhypothese zijn er aan de p β_i's q beperkingen opgelegd en dat wil zeggen dat de schattingen $\hat{\beta}_i$ die aan H_o voldoen, gevonden worden door de deelruimte V_p te vervangen door een deelruimte V_{p-q} waarin q beperkingen tot uiting komen. Immers als bijvoorbeeld $\beta_1 = 0$ moet gelden dan moet x_1 buiten beschouwing blijven, V_p wordt vervangen door een V_{p-1} opgespannen door $x_2,\ldots,x_p$; als $\beta_2 = \beta_3$ moet worden getoetst, dan moeten x_1 en x_2 vervangen worden door hun som enz.

De beste schattingen van β_i vindt men nu door $\vec{y}$ te projecteren op V_{p-q}. Deze projectie zal door $\hat{\vec{\eta}}_\omega$ worden voorgesteld, waarbij $\omega = \Omega \frown H_o$, dat wil zeggen het algemene model Ω in combinatie met de nulhypothese H_o. Wat de "ware" vector $\vec{\eta}$ betreft, kan dus nu gesteld worden:

Onder Ω geldt: $\eta \in V_p$
Onder ω geldt: $\eta \in V_{p-q}$.

Het kwadraat van de lengte van de verschilvector $\vec{y} - \hat{\vec{\eta}}_\omega$ stelt de kwadraatsom van de "afwijkingen" $\hat{\varepsilon}$ voor onder ω. Deze zal door $\mathcal{S}_\omega$ worden aangeduid. In fig. 2 zijn de betrokken vectoren voorgesteld,

fig. 2



waarbij dus bedacht moet worden dat $\vec{y} \in R_n$, $\hat{\vec{\eta}} \in V_p \in R_n$ en $\hat{\vec{\eta}}_\omega \in V_{p-q} \in V_p \in R_n$. Er geldt dus $\vec{y} - \hat{\vec{\eta}} \perp \hat{\vec{\eta}}$ en ook $\perp \hat{\vec{\eta}}_\omega$ omdat $\vec{y} - \hat{\vec{\eta}} \perp V_p$ en $\hat{\vec{\eta}}_\omega$ ook in V. Derhalve $(\vec{y} - \hat{\vec{\eta}}) \perp (\hat{\vec{\eta}} - \hat{\vec{\eta}}_\omega)$ waaruit volgt: $|\vec{y} - \vec{\eta}_\omega|^2 = |\vec{y} - \hat{\vec{\eta}}|^2 + |\vec{\eta} - \vec{\eta}_\omega|^2$ of $|\hat{\vec{\eta}} - \hat{\vec{\eta}}_\omega|^2 = \mathcal{S}_\omega - \mathcal{S}_\Omega$. Er geldt bovendien nog $\hat{\vec{\eta}}_\omega \perp (\vec{y} - \hat{\vec{\eta}}_\omega)$ en $\perp (\vec{y} - \hat{\vec{\eta}})$ dus $\perp (\hat{\vec{\eta}} - \hat{\vec{\eta}}_\omega)$, met andere woorden $\hat{\vec{\eta}}_\omega$ is ook de projectie van $\hat{\vec{\eta}}$ op V_{p-q}.

Om nu een uitspraak omtrent het al of niet gelden van de nulhypothese H_o te doen, suggereert fig. 2 dat H_o geaccepteerd kan worden als $\hat{\vec{\eta}}_\omega$ dicht genoeg bij $\hat{\vec{\eta}}$ ligt, met andere woorden, naarmate $\hat{\vec{\eta}}_\omega$ verder van $\hat{\vec{\eta}}$ verwijderd ligt is het onwaarschijnlijker dat H_o geldt.

Hierbij moet natuurlijk ook de positie van $\vec{y}$ ten opzichte van $\hat{\vec{\eta}}$ beschouwd worden, immers naarmate $\vec{y}$ dichter bij $\hat{\vec{\eta}}$ ligt zal ook $\hat{\vec{\eta}}_\omega$ dichter bij $\hat{\vec{\eta}}$ moeten liggen om met gelijkblijvende waarschijnlijkheid een uitspraak te kunnen doen, dit is althans plausibel.

Derhalve zal het quotiënt $\dfrac{\mathcal{S}_\omega - \mathcal{S}_\Omega}{\mathcal{S}_\Omega}$ (1.3.2)

bepalend moeten zijn voor het al of niet verwerpen van H_o. Hiertoe is nodig de kans te kunnen aangeven dat dit quotiënt bepaalde numerieke waarden bereikt; de frekwentieverdeling van $\dfrac{\mathcal{S}_\omega - \mathcal{S}_\Omega}{\mathcal{S}_\Omega}$ onder de nulhypothese moet worden bepaald. Alvorens hiertoe over te gaan zal eerst nog het probleem en de oplossing, met behulp van matrices worden geformuleerd.

1.4. Matrixvoorstelling van het model.

Als voor de grootheden die in het probleem een rol spelen de volgende notatie wordt ingevoerd

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad ; \quad \vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad ; \quad \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \qquad (1.4.1)$$

en het geheel van de p x n grootheden x_{ik} als een matrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ x_{21} & \cdots & x_{2n} \\ \vdots & & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix} \qquad (1.4.2)$$

wordt voorgesteld, is met behulp van de bekende rekenregels voor vectoren en matrices voor (1.1.1) te schrijven:

$$\vec{y} = X'\vec{\beta} + \vec{\varepsilon} \qquad (1.4.3)$$

en $\vec{\eta} = E(\vec{y}) = X'\vec{\beta}$ (1.4.4)

waarin X' de gespiegelde matrix $\begin{pmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & & \vdots \\ x_{1n} & \cdots & x_{pn} \end{pmatrix}$ is.

Er moet nu om $\hat{\beta}$ te vinden gelden: $(\vec{y} - \hat{\vec{\eta}}) \perp \hat{\vec{\eta}}$ of met de notatie voor het inwendig vectorproduct: $(\vec{y} - \hat{\vec{\eta}}, \hat{\vec{\eta}}) = 0$, dus: $(\vec{y} - X'\hat{\vec{\beta}}, X'\hat{\vec{\beta}}) = (\vec{y}, X'\hat{\vec{\beta}}) - (X'\hat{\vec{\beta}}, X'\hat{\vec{\beta}}) = 0$, waaruit: $\vec{y}'X'\hat{\vec{\beta}} = \hat{\vec{\beta}}XX'\hat{\vec{\beta}} \equiv \hat{\vec{\beta}}'S\hat{\vec{\beta}}$ als:

$XX' \equiv S$ (1.4.5)

wordt gesteld. Er onstaat dus:

$\vec{y}X' = \hat{\vec{\beta}}S$ of $S\hat{\vec{\beta}} = X\vec{y}$ (1.4.6)

Nu is S een p x p symetrische matrix waarvan de elementen bestaan uit de inwendige productsommen $S_{ij} = \sum_{k=1}^{n} x_{ik}x_{kj}$. Onder de voorwaarde dat S nonsingulier is, dat wil zeggen, dat de determinantwaarde $S \neq 0$ is en er dus een inverse S^{-1} bestaat, kan (1.4.6) geschreven worden als: $\hat{\vec{\beta}}' = \vec{y}'X'S^{-1}$ of $\hat{\vec{\beta}} = S^{-1}X\vec{y}$ (1.4.7)

en $\hat{\vec{\eta}} = X'\hat{\vec{\beta}} = X'S^{-1}X\vec{y}$ (1.4.8)

In 1.2 was reeds ingevoerd $\mathcal{S}(\vec{y},\hat{\vec{\beta}}) = \mathcal{S}_\Omega = ||\vec{y}-\hat{\vec{\eta}}||^2 = ||\vec{\varepsilon}||^2 = \sum_{k=1}^{n} \varepsilon_k^2$ dus:

$\mathcal{S}_\Omega = ||\vec{y}-X'\hat{\vec{\beta}}||^2 = ||\vec{y}-X'S^{-1}X\vec{y}||^2 = \vec{y}'\vec{y}-\vec{y}'X'S^{-1}X\vec{y}$ (1.4.9)

met $D = X'S^{-1}X$ is ook te schrijven $\mathcal{S}_\Omega = \vec{y}'(I-D)\vec{y}$ (1.4.10)

$\frac{\partial \mathcal{S}(\vec{y},\hat{\vec{\beta}})}{\partial(\beta)} = X(\vec{y}-X'\hat{\vec{\beta}}) = 0$ geeft $X\vec{y} = XX'\hat{\vec{\beta}} = S\hat{\vec{\beta}}$ of $\hat{\vec{\beta}} = S^{-1}X\vec{y}$.

Uitwerking geeft de normaalvergelijkingen zoals deze reeds in (1.2.1) gevonden waren.

Onder Ω was ondersteld, dat variantie y_k $= E(y_k - Ey_k)^2 = \sigma^2$. Als voor de covariantie matrix van ε de notatie M_ε wordt gebruikt, dan is deze te schrijven als:

$$M_\varepsilon = \sigma^2 I \qquad (1.4.11)$$

Nu geldt in de matrixtheorie nog de volgende stelling:

$$\text{Als } \vec{x} = M\vec{y} \text{ dan } M_{\vec{x}} = MM_{\vec{y}}M' \qquad (1.4.12)$$

$$\text{Hiermede: } M_{\vec{\beta}} = S^{-1}XM_{\vec{y}}X'S^{-1} = \sigma^2 S^{-1}XIX'S^{-1} = \sigma^2 S^{-1} \qquad (1.4.12)$$

Uit deze formule is de standaarddeviatie van de schattingen $\hat{\beta}_i$ af te leiden.

In 1.2 waren de q lineaire functies $\Psi_j(\vec{\beta}_i)$ ingevoerd. Deze zijn te schrijven:

$$\Psi_j = \sum_{i=1}^{p} c_{ji}\beta_i$$

in de matrix notatie $\vec{\Psi} = C\vec{\beta}$ (1.4.13)

met:

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & & \vdots \\ c_{j1} & \cdots & c_{jp} \end{pmatrix}$$

Voor de uit de waarnemingen verkregen schattingen van de parameters geldt: $\hat{\Psi} = C\hat{\vec{\beta}} = CS^{-1}X\vec{y} = A\vec{y}$ (1.4.14)

met $A = CS^{-1}X$.

Nu invoeren $B = AA' = CS^{-1}XX'S^{-1} = CS^{-1}C'$ (1.4.15)

dan wordt de covariantie matrix van $\hat{\vec{\Psi}}$: $M_{\hat{\Psi}} = \sigma^2 B$ (1.4.16)

Onder de voorwaarde dat S nonsingulier is, laat $\mathcal{S}_\Omega$ zich schrijven volgens (1.4.9) of (1.4.10).

Een analoge uitdrukking voor $\mathcal{S}_\omega$ is als volgt te verkrijgen: Nodig is die waarden van $\vec{\beta}$ te vinden waarvoor $\mathcal{S}(\vec{y},\vec{\beta})$ een minimum heeft onder de nevenvoorwaarden dat $\vec{\Psi} = 0$. Dit is op te lossen met behulp van Langrange multiplicatoren.

Hiertoe wordt ingevoerd: $\breve{\mathcal{L}}(\vec{y},\vec{\beta}) = \mathcal{L}(\vec{y},\vec{\beta}) + \vec{\lambda}'C\beta$ (1.4.17)
waarin $\vec{\lambda}'$ een 1 x q vector is, die q constanten (multiplicatoren) q_j als componenten heeft. Als (1.4.17) naar β wordt gedifferentieerd ontstaat: $\frac{\delta\breve{\mathcal{L}}(\vec{y},\vec{\beta})}{\delta\vec{\beta}} = -2X(\vec{y}-X'\vec{\beta}) + C'\vec{\lambda}$,
$\frac{\delta\mathcal{L}}{\delta\vec{\beta}} = 0 \rightarrow 2S\breve{\vec{\beta}} = 2X\vec{y} - C'\vec{\lambda}$ of $\breve{\vec{\beta}} = S^{-1}X\vec{y} - \frac{1}{2}S^{-1}C'\vec{\lambda}$ waarin $\breve{\vec{\beta}}$ de notatie is voor de schattingen van $\vec{\beta}$ als ω geldt. Nu moet aan $\vec{\psi} = 0$ voldaan zijn. Dus $C\breve{\vec{\beta}} = CS^{-1}X\vec{y} - \frac{1}{2}CS^{-1}C'\vec{\lambda} = 0$. De matrix $CS^{-1}C'$ is in (1.4.15) reeds als B ingevoerd. Bewezen kan worden, dat B nonsingulier is (zie Scheffé, blz. 30), zodat geschreven kan worden: $\vec{\lambda} = 2B^{-1}CS^{-1}X\vec{y}$ derhalve
$\vec{\beta} = S^{-1}X\vec{y} - S^{-1}C'B^{-1}CS^{-1}X\vec{y}$ waarna:

$\mathcal{L}_\omega = \mathcal{L}(\vec{y},\breve{\vec{\beta}}) = \vec{y}'\vec{y} - 2\vec{y}'X'(S^{-1}X\vec{y} - S^{-1}C'B^{-1}CS^{-1}X\vec{y}) +$
$(\vec{y}'X'S^{-1} - \vec{y}'X'S^{-1}C'B^{-1}CS^{-1})\, S\, (S^{-1}X\vec{y} - S^{-1}C'B^{-1}CS^{-1}X\vec{y}) =$
$\vec{y}'\vec{y} - \vec{y}'X'\{2S^{-1} - 2S^{-1}C'B^{-1}CS^{-1} - S^{-1} + S^{-1}C'B^{-1}CS^{-1} +$
$S^{-1}C'B^{-1}CS^{-1} - S^{-1}C'B^{-1}CS^{-1}C'B^{-1}CS^{-1}\}\, X\vec{y}$. Tenslotte:

$\mathcal{L}_\omega = \vec{y}\,\vec{y} - \vec{y}'X'(S^{-1} - S^{-1}C'B^{-1}CS^{-1})Xy$ (1.4.18)

In 1.3 was gevonden dat voor de beslissing tot al of niet verwerpen van H_o een toetsingsgrootheid nodig is, die niet $\mathcal{L}_\omega$ maar het verschil $\mathcal{L}_\omega - \mathcal{L}_\Omega$ bevat. Voor het verschil is een eenvoudige uitdrukking te vinden.

Uit (1.4.10) en (1.4.18) volgt: $\mathcal{L}_\omega - \mathcal{L}_\Omega = \vec{y}X'S^{-1}CB^{-1}CS^{-1}X\vec{y}$.
Nu is volgens (1.4.14) $\hat{\vec{\psi}} = CS^{-1}X\vec{y}$ zodat tenslotte gevonden wordt:

$\mathcal{L}_\omega - \mathcal{L}_\Omega = \hat{\vec{\psi}}'B^{-1}\hat{\vec{\psi}}$ (1.4.19)

1.5. De frequentieverdeling van de toetsingsgrootheid.

De vectoren $\vec{y},\vec{\eta}$ en $\vec{\eta}_\omega$ zijn gegeven in coördinaten ten aanzien van een of andere basis in R_n die met $\{v_i\}$ zal worden aangegeven.

De verschilvectoren $\vec{y}-\vec{\eta}$ en $\vec{\eta}-\vec{\eta}_\omega$, die loodrecht op elkaar staan zoals in 1.3 is aangetoond, zijn beide in principe dus uitdrukkingen in alle v_i's. Het is nu handig een andere basis te kiezen en wel een orthonormale basis $\alpha_1 \ldots \alpha_n$ zodanig, dat $1^o)$ p-q basisvectoren $\vec{\alpha}_{q+1} \ldots \vec{\alpha}_p$ de ruimte V_{p-q} opspannen, $2^o)$ q andere basisvectoren $\vec{\alpha}_1 \ldots \vec{\alpha}_q$ met die van V_{p-q} samen, V_p opspannen en $3^o)$ de resterende n-p vectoren $\vec{\alpha}_{p+1} \ldots \vec{\alpha}_n$ er voor zorgen, dat R_n gevormd wordt.

Als nu $z_1,\ldots,z_n$ de coördinaten van $\vec{y}$ ten opzichte van de basis $\{\vec{\alpha}_i\}$ zijn, dan is de coördinaat z_i gelijk aan de projectie van $\vec{y}$ op $\vec{\alpha}_i$ en daar $\vec{\alpha}_i$ een eenheidsvector is, is deze projectie gelijk aan het inproduct $(\vec{\alpha}_i,\vec{y})$. De vector $\vec{z}$, die identiek is aan de vector $\vec{y}$, kan dus in de orthonormale basis $\{\vec{\alpha}_j\}$ uitgedrukt, geschreven worden als $P\vec{y}$, waarin P een n x n matrix is, waarvan de elementen de componenten zijn van de $\vec{\alpha}_i$ ten opzichte van de oude basis $\{v_i\}$.

$$P = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix} \qquad (1.5.1)$$

De rijen van P vormen dus de vectoren $\vec{\alpha}_i$. Volgens de regels van de matrixtheorie is:

$$P\vec{y} \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} \alpha_{1j} y_j \\ \vdots \\ \sum_{j=1}^{n} \alpha_{nj} y_j \end{pmatrix} = \begin{pmatrix} (\alpha_1, y) \\ \vdots \\ (\alpha_n, y) \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$$

De z_i zijn dus lineaire combinaties van de y_j's. Nu zijn de y_j's stochastische variabelen en dus zijn de z_i's ook stochastisch. De verwachtingswaarde van z_i zal door ζ_i worden voorgesteld: $Ez_i = \zeta_i$.

In vectornotatie verder: $\vec{\zeta} = E(\vec{z}) = E(P\vec{y}) = PE(\vec{y}) = P\vec{\eta}$ (1.5.2)
Als combinatie van normaal verdeelde grootheden y_j is z_j zelf ook normaal verdeeld en wel met gemiddelde ζ_i en standaarddeviatie σ. Dit laatste volgt zonder meer uit de eigenschap, dat geometrische grootheden invariant zijn t.o.v. de overgang op een andere basis. Het volgt echter ook uit de matrixnotatie met (1.4.9) namelijk: $M_{\vec{z}} = PM_{\vec{y}}P' = M_{\vec{y}}PP' = M_{\vec{y}} = \sigma^2 I$ ($PP' = I$, omdat de $\vec{\alpha}_i$ onderling loodrechte eenheidsvectoren zijn). De covariant is matrix van $\vec{z}$ is dus $\sigma^2 I$.

Nu stellen dus $\vec{\eta} = E(\vec{y})$ en $\vec{\zeta} = E(\vec{z})$ één en dezelfde vector voor, maar genoteerd respectievelijk t.o.v. de basis $\{v_i\}$ en $\{\alpha_i\}$; daar $\vec{\eta} \in V_p$ ook $\vec{z} \in V_p$ en dat wil zeggen $\zeta_{p+1} = \ldots = \zeta_n = 0$.

Als ω geldt dan is $E(\vec{y}) = \vec{\eta}_\omega$ en derhalve $\vec{\zeta}_\omega = E_\omega(\vec{z}) = P\vec{\eta}_\omega$. Daar $\vec{\eta}_\omega \in V_{p-q}$ moet ook gelden $\vec{\zeta}_\omega \in V_{p-q}$ en dus $\zeta_1 = \ldots = \zeta_q = 0$. Voor de grootheden $\mathscr{S}$ wordt nu respectievelijk gevonden:

$$\mathscr{S}_\Omega = //\vec{y}-\hat{\vec{\eta}}//^2 = //\vec{z}-\hat{\vec{\zeta}}//^2 = //\sum_{i=p+1}^{n} z_i\alpha_i//^2 = \sum_{i=p+1}^{n} z_i^2 \qquad (1.5.3)$$

$$\mathscr{S}_\omega = //\vec{y}-\hat{\vec{\eta}}_\omega//^2 = //\vec{z}-\hat{\vec{\zeta}}_\omega//^2 = //\sum_{1}^{q} z_i\alpha_i + \sum_{p+1}^{n} z_i\alpha_i//^2 = \sum_{1}^{q} z_i^2 + \sum_{p+1}^{n} z_i^2 \qquad (1.5.4)$$

en dus: $$\mathscr{S}_\omega - \mathscr{S}_\Omega = //\hat{\vec{\eta}}-\hat{\vec{\eta}}_\omega//^2 = //\hat{\vec{\zeta}}-\hat{\vec{\zeta}}_\omega//^2 = \sum_{i=1}^{q} z_i^2 . \qquad (1.5.5)$$

Volgens de mathematische statistiek volgt de kwadraatsom van ν onderling onafhankelijk normaal verdeelde grootheden x_i met $\sigma = 1$ een zogenaamde χ^2 (Chi kwadraat) verdeling met ν vrijheidsgraden. Is $E(x_i) = \xi_i \neq 0$ dan is er sprake van een niet centrale χ^2 met centraliteitsparameter

$\delta = (\sum_1^\nu \xi_i^2)^{\frac{1}{2}}$, is $\xi_i = 0$ dan wordt de verdeling centraal genoemd, respectievelijk aangeduid met $\chi^2_{\nu,\delta}$ of χ^2_ν. Er geldt: $E\chi^2_{\nu,\delta} = \nu + \delta^2$. Uit (1.5.5) volgt:

$$E(\mathscr{S}_\omega - \mathscr{S}_\Omega) = E\sum_1^q z_i^2 = \sum_1^q Ez_i^2 = \sum_1^q (E(z_i-\zeta_i)^2 + \zeta_i^2) =$$

$$q\sigma^2 + \sum_1^q \zeta_i^2 = q\sigma^2 + (\sigma\delta)^2 \text{ met } \delta = (\sum_1^q \zeta_i^2/\sigma^2)^{\frac{1}{2}} \qquad (1.5.6)$$

Dus $E(\frac{\mathscr{S}_\omega - \mathscr{S}_\Omega}{\sigma^2}) = q + \delta^2$. Verder $E\mathscr{S}_\Omega = E\sum_{p+1}^n z_i^2 = (n-p)\sigma^2$ wegens $\zeta_{p+1} = \ldots = \zeta_n = 0$.

Derhalve $E\frac{\mathscr{S}_\Omega}{\sigma^2} = (n-p)$. (1.5.7)

$\frac{\mathscr{S}_\omega - \mathscr{S}_\Omega}{\sigma^2}$ volgt dus een $\chi^2_{q,\delta}$ verdeling onder Ω en een χ^2_q verdeling onder ω. $\frac{\mathscr{S}_\Omega}{\sigma^2}$ volgt een χ^2_{n-p} verdeling; immers als Ω geldt, dan ligt $\vec{n}$ in V_p en geldt $\zeta_1,\ldots,\zeta_q \neq 0$; geldt ω dan ligt $\vec{n}$ in V_{p-q} en geldt $\zeta_1 = \ldots = \zeta_q = 0$.

Het quotiënt van twee χ^2 verdelingen is een F verdeling, die ook weer centraal of niet centraal kan zijn. Derhalve is

$$\frac{(n-p)(\mathscr{S}_\omega - \mathscr{S}_\Omega)}{q\mathscr{S}_\Omega} \qquad (1.5.8)$$

verdeeld volgens $F_{q,n-p,\sigma}$ onder Ω en volgens $F_{q,n-p}$ onder ω.

Volgens (1.5.6) en (1.5.2) is $\sigma^2\delta^2 = \sum_1^q \zeta_i^2 = \sum_{i=1}^q \left(\sum_{j=1}^n p_{ij}\eta_j \right)^2$

Verder is $\mathcal{S}_\omega - \mathcal{S}_\Omega = \sum_1^q z_1^2 = \sum_{i=1}^q \left(\sum_{j=1}^n p_{ij}y_j \right)^2$ met andere woorden de non-centraliteitsparameter σ is te vinden door in de uitdrukking $\mathcal{S}_\omega - \mathcal{S}_\Omega$ de y's te vervangen door $Ey_i = \eta_j$

1.6. Een aecquivalente formulering van de F-toets.

In 't voorgaande is aangetoond, dat $\frac{\mathcal{S}_\omega - \mathcal{S}_\Omega}{\sigma^2}$ een $\chi^2_{q,\sigma}$ verdeling volgt, terwijl ook geldt dat $\mathcal{S}_\omega - \mathcal{S}_\Omega = \hat{\vec{\psi}}' B^{-1} \hat{\vec{\psi}}$. Verder was σ te vinden door in de uitdrukking voor $\mathcal{S}_\omega - \mathcal{S}_\Omega$ als functie van de y_j, deze y_j door $\eta_j = Ey_j$ te vervangen. Dit is dus acquivalent met $E(\mathcal{S}_\omega - \mathcal{S}_\Omega) = E\hat{\vec{\psi}}' B^{-1} \hat{\vec{\psi}} = \vec{\psi} B^{-1} \vec{\psi} = \sigma^2\delta^2$. Derhalve volgt $\frac{(\hat{\vec{\psi}} - \vec{\psi})' B^{-1} (\hat{\vec{\psi}} - \vec{\psi})}{\sigma^2}$ een χ^2_q terwijl $\frac{\mathcal{S}_\Omega}{\sigma^2}$ een χ^2_{n-p} verdeling volgt. Als $S^2 = \frac{\mathcal{S}_\Omega}{n-p}$ wordt gesteld, dan volgt $\frac{(\hat{\vec{\psi}} - \vec{\psi})' B^{-1} (\hat{\vec{\psi}} - \vec{\psi})}{qS^2}$ een $F_{q,n-p}$ verdeling. Dit is ook zo op te vatten, dat $(\hat{\psi}-\psi)' B^{-1} (\hat{\psi}-\psi) \leq qS^2 F_{q,n-p}$ een vertrouwensgebied vormt, een ellipsoide voor $\vec{\psi}$ in de q-dimensionale ψ-ruimte met centrum $\hat{\psi}_1, .., \hat{\psi}_q$. In het geval q = 1 gaat dit over in een vertrouwensinterval: $b^{-1}(\hat{\psi}-\psi)^2 \leq S^2 F_{1,n-p}$ (1.6.1)

Nu is een F-verdeling met $\nu_1 = 1$ acquivalent met een t-verdeling dat wil zeggen, als $\underline{t}$ een t-verdeling met ν vrijheidsgraden volgt, dan volgt $\underline{t}^2$ een $F_{1,\nu}$ verdeling; derhalve is het vertrouwensinterval voor $\vec{\psi}$ ook te schrijven als:

$$\hat{\vec{\psi}} - t_{n-p} S\hat{\vec{\psi}} \leq \vec{\psi} \leq \hat{\vec{\psi}} + t_{n-p} S\hat{\vec{\psi}} \qquad (1.6.2)$$

waarbij bS^2 door $S^2_{\hat{\vec{\psi}}}$ is vervangen omdat bS^2 gelijk is aan de variantie van $\hat{\vec{\psi}}$.

2. Voorbeelden.

2.1. Enkelvoudige variantie analyse.

Als een grootheid y onder bijvoorbeeld k verschillende omstandigheden wordt waargenomen en men wil weten of deze verschillende omstandigheden invloed hebben op de waarde van y dan kan men dit onderzoeken met behulp van enkelvoudige variantie analyse. (Het is dan zo, dat een eventuele invloed van de omstandigheden zich uit in een niveauverandering van y). De verschillende omstandigheden betekenen voor de waarnemingen en voor de theoretische grootheden van het betrokken probleem een indeling in klassen, die met een index i zullen worden aangegeven. In de termen van 1 is dit probleem als volgt te formuleren:

Ω: $y_{t,r} = \sum_{i=1}^{k} \beta_i x_{i,t} + \varepsilon_{t,r}$ waarin $t = 1,\ldots,k \quad r = 1,\ldots,n_k$

$\sum_{t=1}^{k} n_t = n$ en $\delta_{i,t}$ = (Kronecker-symbool ; dat wil zeggen $\delta_{i,t} = 1$ als $t = i$; $\delta_{i,t} = 0$ als $t \neq i$) $x_{i,t}$ is de in (0.3) genoemde indicator variabele. Korter kan ook $y_{t,r} = \beta_t + \varepsilon_{t,r}$ (2.1.1) geschreven worden. Voor de $\varepsilon_{t,r}$ geldt: onderling onafhankelijk normaal verdeeld met verwachtingswaarde nul en standaarddeviatie σ. Derhalve $Ey_{t,r} = \beta_t$.

Er zijn dus k steekproeven van de grootte n_i (i=1,...,k). De eerste vraag die men zich kan stellen is of er inderdaad een reëel verschil tussen de diverse β_i bestaat. In statistische zin ligt het dan voor de hand te toetsen of de β's mogelijk alle aan elkaar gelijk zijn. Dat wil zeggen, onderzocht wordt de nulhypothese: $H_o : \beta_1 = \beta_2 = \ldots = \beta_k$ (2.1.2) Nu is dit een vrij eenvoudig probleem, dat zonder meer rechtstreeks met de methode der kleinste kwadraten is op te lossen; daarom echter juist geschikt als illustratie.

Naast de rechtstreekse oplossing zal ook de formele oplossing worden gegeven.

2.1.2 Kleinste kwadratische oplossing.

De kwadraatsom $S(\vec{y},\vec{\beta})$ is onder Ω: $S(\vec{y},\vec{\beta}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \beta_i)^2$

Deze is minimaal als $\frac{\delta S(\vec{y},\vec{\beta})}{\delta \beta_\nu} = -2\sum_{j=1}^{n_\nu}(y_{\nu j}-\beta_\nu) = -2\sum_{j=1}^{n_\nu} y_{\nu j} + 2n_\nu\beta_\nu = 0$,

waaruit: $\hat{\beta}_\nu = \sum_{j=1}^{n_\nu} y_{\nu j}/n_\nu = \bar{y}_\nu$. Met andere woorden, de beste schattingen van de β_ν is het gemiddelde van de y waarden onder de met ν aangegeven omstandigheden. Als ook H_o geldt (notatie van 1.2: $\omega = H_o \cap \Omega$) dan zijn alle β_i gelijk, bijvoorbeeld $= \beta$ en dan is $S(\vec{y},\vec{\beta}) = \sum\sum_{ij}(y_{ij} - \beta)^2$ en dus

$$\frac{\delta S(\vec{y},\vec{\beta})}{\delta\beta} = -2\sum_i\sum_j(y_{ij} - \beta) = -2\sum\sum_{ij} y_{ij} + 2\beta\sum_j n_j = 0$$

waaruit: $\hat{\beta} = \bar{\bar{y}}(=\frac{1}{n_{ij}}\sum\sum y_{ij})$.

In de notatie van 1.2 is nu:

$\vec{y} = \{y_{11},\ldots y_{1n_1}, y_{2.1}\ldots y_{2n_2}\ldots y_{k,1}\ldots y_{k,n_k}\}$. Verder is wegens $\eta_{ij} = Ey_{ij} = \beta_i$: $\hat{\vec{\eta}} = \{\hat{\beta}_1\ldots\hat{\beta}_1, \hat{\beta}_2\ldots\hat{\beta}_2, \ldots\ldots, \hat{\beta}_k\ldots\hat{\beta}_k\}$ en dus $\hat{\vec{\eta}}_\omega = \{\hat{\beta}, \ldots \hat{\beta}\}$ waarmee:

$$\mathcal{S}_\Omega = ||\vec{y} - \vec{\eta}||^2 = \sum\sum_{ij}(y_{ij} - \hat{\beta}_i)^2 = \sum\sum_{ij}(y_{ij} - \bar{y}_i)^2 = \sum\sum_{ij} y_{ij}^2 - \sum_i n_i\bar{y}_i^2$$

en $\mathcal{S}_\omega = ||\vec{y} - \vec{\eta}_\omega||^2 = \sum\sum_{ij}(y_{ij} - \hat{\beta})^2 = \sum\sum_{ij}(y_{ij} - \bar{\bar{y}})^2 = \sum\sum_{ij} y_{ij}^2 - n\bar{\bar{y}}^2$

Dus $\mathcal{S}_\omega - \mathcal{S}_\Omega = \sum_i n_i\bar{y}_i^2 - n\bar{\bar{y}}^2$ wat ook te vinden is als:

$$||\hat{\vec{\eta}} - \vec{\eta}_\omega||^2 = \sum\sum_{ij}(\bar{y}_i - \bar{\bar{y}})^2 = \sum_i n_i(\bar{y}_i - \bar{\bar{y}})^2.$$

$S_\omega - S_\Omega = \sum_i n_i(\bar{y}_i - \bar{\bar{y}})^2$ is een gewogen maat voor de spreiding van de steekproefgemiddelden van de k steekproeven rond het algemeen gemiddelde. Dit wordt ook wel als $S_{t(ussen)}$ genoteerd en heet dan spreidingskwadraatsom tussen de steekproeven. $\mathcal{S}_\Omega = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ is een gecombineerde maat voor de spreiding binnen de steekproeven, genoteerd als $S_{b(innen)}$.

Volgens 1.5 volgt $\mathcal{S}_\Omega$ een χ^2 verdeling met n-p vrijheidsgraden waarin p het aantal te schatten β's is, dus k in bovenstaand geval. S_{binnen} heeft dus n-k vrijheidsgraden. $\mathcal{S}_\omega - \mathcal{S}_\Omega$ heeft 9 vrijheidsgraden, waarin 9 het aantal functies ψ_j is dat H_o bepaalt. Met invoering van : $\psi_1 = \beta_1 - \beta_k$, $\psi_2 = \beta_2 - \beta_k, \ldots, \psi_{k-1} = \beta_{k-1} - \beta_k$ is $H_o : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ te schrijven als $\psi_i = 0$ voor $i = 1,\ldots,k-1$. Derhalve is het aantal vrijheidsgraden dat voor $\mathcal{S}_\omega - \mathcal{S}_\Omega$ in rekening moet worden genomen k-1. $\mathcal{S}_\omega$ heeft tenslotte n-1 vrijheidsgraden.

Volgens (1.5.7) geldt cok: $E\, \mathcal{S}_\Omega/n-p = \sigma^2$ (dus $\mathcal{S}_\Omega/n-p$ is een "zuivere" schatter van σ^2), $E\, \frac{\mathcal{S}_\omega - \mathcal{S}_\Omega}{q} = \sigma^2 + \frac{1}{q}\sigma^2\delta^2 = \sigma^2 + \frac{1}{q}\sum_1^q \zeta_i^2$

In bovenstaand geval wordt dit volgens de slotregel in (1.5):

$$E\, \frac{\mathcal{S}_\omega - \mathcal{S}_\Omega}{q} = \sigma^2 + \frac{1}{q}\sum_{i=1}^{k-1} n_i(\beta_i - \beta)^2.$$

De resultaten zijn in de volgende tabel samen te vatten:

$S_t \equiv S_{tussen}$

Bron	kwadraat-sommen	aantal vrijheids-graden	variantie	E.(variantie)
tussen de groepen	$S_t = \sum_i n_i(\bar{y}_i-\bar{\bar{y}})^2$	k-1	$S_t/k-1$	$\sigma^2 + \frac{\sum n_i(\beta_i-\beta)^2}{k-1}$
binnen de groepen	$S_b = \sum_i \sum (y_{ij}-\bar{y})^2$	n-k	$S_b/n-k$	σ^2
totaal	$S\ "tot" = \sum_{ij} \sum (y_{ij}-y)^2$	n-1	-	-

De uitspraak of H_o verworpen moet worden volgt uit de bepaling van de overschrijdingskans van

$$\frac{n-k}{k-1}\frac{S_{tussen}}{S_{binnen}} = \frac{n-k}{k-1}\frac{\sum_{i=1}^{k} n_i(y_{i.}-\bar{y})^2}{\sum_{ij}\sum (y_{ij}-y_{i.})^2}$$ op grond van

de $F_{k-1,n-k}$ verdeling.

2.1.3 Formele matrix oplossing.

Uitgangspunt:

Ω : $y_{ij} = \beta_i + \varepsilon_{ij}$ $(i=1,\ldots,k\ ;\ j=1,\ldots,n_i\ ;\ \Sigma n_i=n; \varepsilon : N(0,\sigma^2))$

H_o : $\beta_1 = \beta_2 = \ldots = \beta_k$

$\psi_j = \beta_k - \beta_j$ $(j=1,\ldots,k-1)$

dus ook:

H_o : $\psi_j = 0$ $(j=1,\ldots,k-1)$

$\vec{y}$: $\{y_{11}\ldots y_{1n_1} y_{21}\ldots y_{2n_2}\ldots y_{kn_k}\}$

De in (0.1) ingevoerde grootheden $\vec{x}_i$ zijn in dit geval

$\vec{x}_1 = \{1,\ldots 1, 0,\ldots,0\}$ enz.

n_1

$\vec{x}_2 = \{\underbrace{0,\ldots 0}_{n_1}, \underbrace{1,\ldots,1}_{n_2}, 0,\ldots,0\}$ enz.

dus:

$$X = \begin{bmatrix} 1\ldots\ldots 1\ 0\ldots 0 & & & 0 \\ 0\ldots\ldots 0\ 1\ldots 1\ 0\ 0 & & & 0 \\ \vdots & & & \vdots \\ 0 & & 0\ 0\ 1\ldots 1 & 0 \end{bmatrix}$$

waarmee $S = XX' = \begin{bmatrix} n_1 & & 0 \\ & n_2 & \\ & & \ddots \\ 0 & & n_k \end{bmatrix}$; $S^{-1} = \begin{bmatrix} \frac{1}{n_1} & & 0 \\ & \frac{1}{n_2} & \\ & & \ddots \\ 0 & & \frac{1}{n_k} \end{bmatrix}$

volgens (1.4.7) is:

$$\hat{\vec{\beta}} = S^{-1}X\vec{y} = \begin{bmatrix} \frac{1}{n_1} & & 0 \\ & \frac{1}{n_2} & \\ & & \ddots \\ 0 & & \frac{1}{n_k} \end{bmatrix} \begin{bmatrix} 1\ldots 1 & & 0 \\ & \ddots & \\ 0 & & 1\ldots 1 \end{bmatrix} \cdot \begin{bmatrix} y_{11} \\ \vdots \\ y_{kn_k} \end{bmatrix} =$$

$$\begin{bmatrix} \frac{1}{n_1} & & 0 \\ & \frac{1}{n_2} & \\ & & \ddots \\ 0 & & \frac{1}{n} \end{bmatrix} \cdot \begin{bmatrix} \Sigma y_{1j} \\ \Sigma y_{2j} \\ \vdots \\ \Sigma y_{kj} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ y_2 \\ \vdots \\ \bar{y}_k \end{bmatrix} \quad \text{waaruit } \hat{\beta}_j = \bar{y}_j$$

verder kan men σ bepalen als $||\vec{y} - \hat{\vec{n}}||^2$ of als $\vec{y}'(I-D)\vec{y}$ volgens (1.4.10).

Nu is $\hat{\vec{\eta}}' = (\bar{y}_1,\ldots,\bar{y}_1,\ \bar{y}_2,\ldots,\bar{y}_2,\ldots\bar{y}_k,\ldots,\bar{y}_k)$

Dus $||\vec{y} - \hat{\vec{\eta}}||^2 = \sum\sum_{ij}(y_{ij} - \bar{y}_i)^2 = \sum\sum_{ij} y_{ij}^2 - \sum_i n_i \bar{y}_i^2$.

ook

$$D = X'S^{-1}X = \begin{pmatrix} 1 & & 0 \\ \vdots & \ddots & \\ 1 & & \\ & \ddots & \vdots \\ 0 & & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \frac{1}{n_2} & \\ & & \ddots \\ 0 & & \frac{1}{n_k} \end{pmatrix} \cdot \begin{pmatrix} 11\ldots1 & & 0 \\ & \ddots & \\ 0 & & \ddots \\ & & 11\ldots1 \end{pmatrix} =$$

$$\begin{pmatrix} \begin{pmatrix} \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ \vdots & & \vdots \\ \frac{1}{n_1} & \cdots & \frac{1}{n_1} \end{pmatrix} & & \\ & \begin{pmatrix} \frac{1}{n_2} & \cdots & \frac{1}{n_2} \\ \vdots & & \vdots \\ \frac{1}{n_2} & \cdots & \frac{1}{n_2} \end{pmatrix} & \\ 0 & & \begin{pmatrix} \frac{1}{n_k} & \cdots & \frac{1}{n_k} \\ \vdots & & \vdots \\ \frac{1}{n_k} & \cdots & \frac{1}{n_k} \end{pmatrix} \end{pmatrix}$$

Dus:

$$I-D = \begin{pmatrix} \begin{pmatrix} (1-\frac{1}{n_1}) & -\frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ -\frac{1}{n_1} & & & \vdots \\ \vdots & & \ddots & \\ -\frac{1}{n_1} & \cdots & \cdots & (1-\frac{1}{n_1}) \end{pmatrix} & & \\ & \begin{pmatrix} (1-\frac{1}{n_2}) & -\frac{1}{n_2} & \cdots \\ -\frac{1}{n_2} & \ddots & \\ \vdots & & \end{pmatrix} & \\ 0 & & \ddots \end{pmatrix}$$

$(I - D)y$ wordt nu $\begin{pmatrix} y_{11} - \bar{y}_1 \\ y_{12} - \bar{y}_1 \\ \vdots \\ y_{21} - \bar{y}_2 \\ \vdots \end{pmatrix}$

zodat $\mathcal{S}_\Omega = \sum_{ij}\sum (y_{ij} - \bar{y}_i)y_{ij} = \sum_{ij}\sum y_{ij}^2 - \sum n_i \bar{y}_i^2$.

wegens H_0: $\psi_j = \beta_k - \beta_j$ wordt de matrix C:

$$C = \begin{pmatrix} -1 & 0 & \cdots & 1 \\ 0 & -1 & & \vdots \\ \vdots & & \ddots & \vdots \\ & & -1 & 1 \end{pmatrix}$$

Dan is $B = CS^{-1}C' =$

$$\begin{pmatrix} -1 & 0 & \cdots & 0 & 1 \\ 0 & -1 & & & \vdots \\ \vdots & & \ddots & & \vdots \\ & & & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{n_1} & & & 0 \\ & \frac{1}{n_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{n_k} \end{pmatrix} \begin{pmatrix} -1 & & 0 \\ & \ddots & \\ 0 & & -1 \\ 1 & & 1 \end{pmatrix}$$

Hieruit:

$$B^{-1} = \begin{pmatrix} n_1 - \frac{n_1^2}{n} & -\frac{n_1 n_2}{n} & \cdots & -\frac{n_1 n_{k-1}}{n} \\ -\frac{n_2 n_1}{n} & n_2 - \frac{n_2^2}{n} & & \vdots \\ \vdots & & \ddots & \vdots \\ & & & \end{pmatrix}$$

Verder is:

$$\hat{\psi} = \begin{pmatrix} y_k - \bar{y}_1 \\ y_k - \bar{y}_2 \\ \vdots \\ y_k - \bar{y}_{k-1} \end{pmatrix}$$

Derhalve $\mathcal{S}_\omega - \mathcal{S}_\Omega = \hat{\psi}' B^{-1} \hat{\psi} =$

$$(\bar{y}_k - \bar{y}_1, \dots) \begin{pmatrix} n_1 - \frac{n_1^2}{n} & -\frac{n_1 n_2}{n} & \cdots \\ -\frac{n_1 n_2}{n} & n_2 - \frac{n_1^2}{n} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \cdot \begin{pmatrix} \bar{y}_k - \bar{y}_1 \\ \vdots \end{pmatrix} =$$

$$(\bar{y}_k - \bar{y}_1, \dots) \begin{pmatrix} n_1(\bar{y}_k - \bar{y}_1) - \frac{n_1}{n} \sum_{j=1}^{k} n_j (\bar{y}_k - \bar{y}_j) \\ n_2(\bar{y}_k - \bar{y}_2) - \frac{n_2}{n} \sum_{j=1}^{k} n_j (\bar{y}_k - \bar{y}_j) \\ \vdots \end{pmatrix} =$$

$$(\bar{y}_k - \bar{y}_1, \dots) \cdot \begin{pmatrix} n_1(\bar{y} - \bar{y}_1) \\ n_2(\bar{y} - \bar{y}_2) \\ \vdots \end{pmatrix} = \sum_{i=1}^{k} n_i (y - y_i)(y_k - y_i) =$$

$$\sum_{i=1}^{k} n_i (\bar{y} - \bar{y}_i)(\bar{y} - \bar{y}_i + \bar{y}_k - \bar{y}) = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 = \Sigma n_i \bar{y}_i^2 - n\bar{y}^2.$$

Berekening van ζ_ω volgens (1.4.18) levert resp.:

$$S^{-1} C' = \begin{pmatrix} -\frac{1}{n_1} & & 0 \\ & -\frac{1}{n_2} & \\ 0 & & \ddots \\ & & \frac{1}{n_{k-1}} \\ \frac{1}{n_k} & \cdots & \frac{1}{n_k} \end{pmatrix}$$

$$S^{-1}C'B^{-1} = \begin{pmatrix} \frac{n_1}{n} - 1 & \frac{n_2}{n} & \cdots & \frac{n_{k-1}}{n} \\ \frac{n_1}{n} & \frac{n_2}{n} - 1 & & \frac{n_{k-1}}{n} \\ \vdots & & \ddots & \vdots \\ \frac{n_1}{n} & & & \frac{n_{k-1}}{n} - 1 \\ \frac{n_1}{n} & \cdots & & \frac{n_{k-1}}{n} \end{pmatrix}$$

$$S^{-1}C'B^{-1}CS^{-1} = \begin{pmatrix} \frac{1}{n_1} - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{1}{n_2} - \frac{1}{n} & & & \vdots \\ \vdots & & \ddots & & \\ & & & \frac{1}{n_{k-1}} - \frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \cdots & & -\frac{1}{n} & \frac{1}{n_k} - \frac{1}{n} \end{pmatrix}$$

dus

$$S^{-1} - S^{-1}C'B^{-1}CS^{-1} = \begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

waarmee

$$\vec{y}'X'(S^{-1} - S^{-1}C'B^{-1}CS^{-1})X\vec{y} = (n_1\bar{y}_1 \quad n_2\bar{y}_2 \ldots n_k\bar{y}_k)\cdot$$

$$\begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \cdot \begin{pmatrix} n_1\bar{y}_1 \\ n_2\bar{y}_2 \\ \vdots \\ n_k\bar{y}_k \end{pmatrix} = \frac{1}{n}(\Sigma n_i y_i)^2 = n\bar{y}^2. \quad \text{Dus } \mathcal{S}_\omega = \Sigma y_i^2 - n\bar{y}^2.$$

Zodat inderdaad $\mathcal{S}_\omega - \mathcal{S}_\Omega = \Sigma n_i y_i^2 - n\bar{y}^2$.
Hiermede zijn dus dezelfde resultaten verkregen als met de rechtstreekse methode.

2.1.4 Numeriek voorbeeld.

Een adequaat voorbeeld van enkelvoudige variantie-analyse is voor meteorologische waarnemingen niet makkelijk te vinden. Meestal zijn de verschijnselen zo complex, dat meerdere effecten door elkaar een rol spelen, of is er sprake van tijdreeksen, waarvoor andere methoden ten onderzoek beter geschikt zijn. Toch is een voorbeeld uit een tijdreeks gekozen namelijk de vorstexcentriciteiten die door F. Ynsen zijn samengesteld ter karakterisering van de winters en gepubliceerd in W.R.74-2. Deze gegevens hebben het voordeel, dat ze te beschouwen zijn als onafhankelijk verdeelde normale grootheden.

Gebruikt zullen worden de vorstexcentriciteiten van 210 jaar namelijk van 1760 tot en met 1969, waarvan 7 steekproeven van 30 worden samengesteld. De vraag is nu of de gemiddelde vorst-excentriciteiten van de 7 opeenvolgende 30-jaar-perioden gelijk zijn. De nulhypothese luidt in de notatie van 2.2.1 dus: $\beta_1 = \beta_2 = \ldots = \beta_7 = \beta$. De schattingen van deze β's blijken te zijn:

$\hat{\beta}_1 = 0,0733$

$\hat{\beta}_2 = -0,0823$

$\hat{\beta}_3 = -0,1226$

$\hat{\beta}_4 = 0,1256$

$\hat{\beta}_5 = 0,1710$

$\hat{\beta}_6 = -0,2066$

$\hat{\beta}_7 = 0,2690$

Het algemeen gemiddelde is $\bar{\beta} = 0,0324$

De kwadraatsom tussen de groepen $S_{tussen} = \mathcal{S}_\omega - \mathcal{S}_\Omega = 5,398$

De kwadraatsom binnen de groepen $S_{binnen} = \mathcal{S}_\omega = 225,704$

De toetsingsgrootheid: $\frac{n-k}{k-1} \cdot \frac{S_{tussen}}{S_{binnen}} = \frac{203}{6} \cdot \frac{5,398}{225,704} = 0,829$

De overschrijdingskans is > 20%, zodat er geen enkele reden is de nulhypothese te verwerpen. Er is dus geen verschil in niveau van de winterkarakterisering in de zeven 30-jaar-tijdvakken te constateren. Natuurlijk in overeenstemming met de constatering in W.R.74-2 dat de reeks vorst-excentriciteiten als een volkomen toevalsreeks is te beschouwen.

Opmerking: Een uitkomst zoals hier, dat de nulhypothese niet verworpen behoeft te worden, geeft verder geen problemen. Een wel verwerpen van de nulhypothese geeft direct aanleiding tot de vraag : Welke waarden voor de β_i moeten dan als reëel worden aanvaard? Het is natuurlijk helemaal niet nodig de berekende schattingen $\hat{\beta}_i$ zonder meer als de beste schattingen te gebruiken, er kan in beperktere mate onderlinge gelijkheid bestaan. Het is de bedoeling hierop in een vervolgrapport nader in te gaan.

2.2. Regressie-analyse.

2.2.1. Toetsing van een kwadratische regressie.

Als in geval van een relatie tussen y en x vermoed wordt, dat deze relatie kwadratisch is, namelijk $y = a + bx + cx^2$ dan kan men toetsen of dit inderdaad het geval is of dat men toch met een lineair verband kan volstaan. Volgens de algemene theorie kan gesteld worden:

$\underline{\Omega}$: $y_i = \alpha + \beta x_{1i} + \gamma x_{2i} + \varepsilon_i$ (2.2.1)

met $x_1 = x$ en $x_2 = x^2$ verder weer $E\varepsilon_i = 0$, $E\varepsilon_i\varepsilon_j = \sigma^2\delta_{ij}$ waarin $\delta_{ij} = 1$ als $i = j$, $\delta_{ij} = 0$ als $i \neq j$.

$\underline{\omega}$: $y_i = \alpha + \beta x_i + \varepsilon_i$ $(i = 1,\ldots,n)$ (2.2.2)

Nu heeft de matrix X de volgende vorm:

$$X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_2^2 \end{pmatrix}$$

$$S = XX' = \begin{pmatrix} n & \Sigma x & \Sigma x^2 \\ \Sigma x & \Sigma x^2 & \Sigma x^3 \\ \Sigma x^2 & \Sigma x^3 & \Sigma x^4 \end{pmatrix}$$

Als de determinant van S door D_s wordt voorgesteld, dan geldt:

$$D_s = n\Sigma x^2\Sigma x^4 - n(\Sigma x^3)^2 + 2\Sigma x\Sigma x^2\Sigma x^3 - (\Sigma x)^2\Sigma x^4 - (\Sigma x^2)^3$$

met $t_i = \frac{\Sigma x^i}{\sqrt{D_s}}$ is S^{-1} te schrijven als:

$$S^{-1} = \begin{pmatrix} t_2t_4 - t_3^2 & t_2t_3 - t_1t_4 & t_1t_3 - t_2^2 \\ t_2t_3 - t_1t_4 & t_0t_4 - t_2^2 & t_1t_2 - t_0t_3 \\ t_1t_3 - t_2 & t_1t_2 - t_0t_3 & t_3t_2 - t_1^2 \end{pmatrix} = \begin{pmatrix} T_{24} & T_{23} & T_{13} \\ T_{23} & T_{04} & T_{12} \\ T_{13} & T_{12} & T_{02} \end{pmatrix}$$

als $T_{24} = t_2t_4 - t_3^2$ enz. Hiermede worden de kolommen van $S^{-1}X$:

$$T_{24} + T_{23}x_i + T_{13}x_i^2$$

$$T_{23} + T_{04}x_i + T_{12}x_i^2$$

$$T_{13} + T_{12}x_i + T_{02}x_i^2$$

en derhalve:
$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = S^{-1}X\vec{y} = \begin{pmatrix} T_{24}\Sigma y + T_{23}\Sigma yx + T_{13}\Sigma yx^2 \\ T_{23}\Sigma y + T_{04}\Sigma yx + T_{12}\Sigma yx^2 \\ T_{13}\Sigma y + T_{12}\Sigma yx + T_{02}\Sigma yx^2 \end{pmatrix} \qquad (2.2.4)$$

De elementen van de matrix $D = X'S^{-1}X$ worden nu:

$$D_{ij} = T_{24} + T_{23}x_i + T_{13}x_i^2 + T_{23}x_j + T_{04}x_ix_j + T_{12}x_i^2x_j + T_{13}x_j^2 + T_{12}x_ix_j^2 + T_{02}x_i^2x_j^2$$

en $\mathcal{S}_\Omega = ||\vec{y} - \hat{\vec{n}}||^2 = \vec{y}'\vec{y} - \vec{y}'D\vec{y} = \Sigma y^2 - \Sigma\Sigma\, y_iD_{ij}y_j =$

$$\Sigma y_{ij}^2 - T_{24}(\Sigma y)^2 - 2T_{23}\Sigma y\,\Sigma xy - 2T_{13}\Sigma y\Sigma x^2y - T_{04}\Sigma x^2y - 2T_{12}\Sigma x^2y\,\Sigma xy - T_{02}(\Sigma x^2y)^2 \qquad (2.2.5)$$

Rechtstreeks berekening van $\mathcal{S}_\omega$ gaat als volgt: De nulhypothese onder ω luidt $\psi = \gamma$. Dus $C = (001)$, waarmee $B = C'S^{-1}C = T_{02}$ derhalve $B^{-1} = T_{02}^{-1}$.

$S^{-1}C'$ wordt nu $\begin{pmatrix} T_{13} \\ T_{12} \\ T_{02} \end{pmatrix}$

en $S^{-1}C'B^{-1}CS^{-1} = \begin{pmatrix} T_{13}^2/T_{02} & T_{12}T_{13}/T_{02} & T_{13} \\ T_{12}T_{13}/T_{02} & T_{12}^2/T_{02} & T_{12} \\ T_{13} & T_{12} & T_{02} \end{pmatrix}$

zodat $\mathcal{S}_\omega = \Sigma y_{ij}^2 - \vec{y}'X' \begin{pmatrix} T_{24}-T_{13}^2/T_{02} & T_{23}-T_{12}T_{13}/T_{02} & 0 \\ T_{23}-T_{12}T_{13}/T_{02} & T_{04}-T_{12}^2/T_{02} & 0 \\ 0 & 0 & 0 \end{pmatrix} Xy$

$$= \Sigma y_{ij}^2 - (T_{24} - \frac{T_{13}^2}{T_{02}})(\Sigma y)^2 - 2(T_{23} - \frac{T_{12}T_{13}}{T_{02}})\Sigma y \Sigma xy - (T_{04} - \frac{T_{12}^2}{T_{02}})(\Sigma xy)^2 \qquad (2.2.7)$$

en dus: $\mathcal{S}_\omega - \mathcal{S}_\Omega = \frac{T_{13}^2}{T_{02}}(\Sigma y)^2 + 2\frac{T_{12}T_{13}}{T_{02}}\Sigma y \Sigma xy + \frac{T_{12}^2}{T_{02}}\Sigma x^2 y + 2T_{13}\Sigma y(\Sigma xy)^2 +$

$2T_{12}\Sigma xy\Sigma x^2 y + T_{02}(\Sigma x^2 y)^2 = \frac{1}{T_{02}}(T_{13}\Sigma y + T_{12}\Sigma xy + T_{02}\Sigma x^2 y)^2$

Dit resultaat is sneller te vinden met (1.4.19): $\mathcal{S}_\omega - \mathcal{S}_\Omega = \hat{\psi}'B^{-1}\hat{\psi}$

waarin $\hat{\psi} = \hat{r} = T_{13}\Sigma y + T_{12}\Sigma yx + T_{02}\Sigma yx^2$ volgens (2.2.4) en

$B^{-1} = T_{02}^{-1}$ volgens (2.2.6)

Er valt op te merken, dat $\mathscr{S}_\omega$ ook te berekenen is op analoge wijze als $\mathscr{S}_\Omega$, omdat in dit geval de X-matrix die geldt als ω in plaats van Ω wordt genomen direct is op te schrijven, namelijk:

$$\tilde{X} = \begin{pmatrix} 1 & 1 & & 1 \\ x_1 & x_2 & & x_n \end{pmatrix}$$

$$\text{en dus } \tilde{S} = \tilde{X}\tilde{X}' = \begin{pmatrix} n & \Sigma x \\ \Sigma x & \Sigma x^2 \end{pmatrix}$$

Als de determinant van $\tilde{S}$ met $\tilde{D}_s$ wordt aangegeven, dan:
$\tilde{D}_s = n\Sigma x^2 - (\Sigma x)^2$ waarna met $r_i = \frac{\Sigma x^i}{\tilde{D}_s}$ is te schrijven:

$$S^{-1} = \begin{pmatrix} r_2 & -r_1 \\ -r_1 & r_0 \end{pmatrix} \text{ en } \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = S^{-1}Xy = \begin{pmatrix} r_2\Sigma_y - r_1\Sigma xy \\ -r_1\Sigma_y - r_0\Sigma xy \end{pmatrix} .$$

Opmerking:

$$\text{Dus wordt } \hat{\beta} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - \Sigma x^2} = \frac{\Sigma xy - nxv}{\Sigma x^2 - nx^2} = \frac{\Sigma(x - x)(y - y)}{\Sigma(x - x)^2}$$

waarmee de bekende formule voor de lineaire regressie-coëfficiënt van y op x is gevonden. Verder is

$$\hat{\alpha} = \frac{\Sigma x^2\Sigma y - \Sigma x\Sigma xy}{n\Sigma x^2 - (\Sigma x)^2} = \frac{v\Sigma x^2 - x\Sigma xy}{\Sigma x^2 - nx^2} = y - \hat{\beta}x$$

de bekende formule voor de constante term in de lineaire regressie.

De elementen van de matrix $D = XS^{-1}X$ worden nu
$\tilde{D}_{ij} = r_2 - r_1x_i - r_1x_i - r_1x_i + r_0x_ix_j$ en dus:
$\mathscr{S}_\omega = \Sigma v^2 - \Sigma_i\Sigma_j v_i\tilde{D}_{ij}y_i = \Sigma v^2 - r_2\Sigma v^2 + 2r_1\Sigma y\Sigma xy - r_0\Sigma x^2y$ (2.2.8)

Dat (2.2.7) identiek is aan (2.2.6) is tenslotte na enig rekenwerk aan te tonen.

De toetsing van H_o kan nu weer op de bekende manier worden uitgevoerd door $(n - l)\,\frac{\mathcal{S}_\omega - \mathcal{S}_\Omega}{\mathcal{S}_\Omega}$ te vergelijken met het gekozen significantie-niveau volgens de $F_{l.n-l}$ verdeling.

Zoals in 1.6 is aangegeven kan ook gebruik worden gemaakt van een vertrouwensinterval: $\hat{\psi} - t_{n-p}S_{\hat{\psi}} < \psi < \hat{\psi} + t_{n-p}S_{\hat{\psi}}$

wat nu wordt: $\hat{\gamma} - t_{n-p}S_{\hat{\gamma}} < \gamma < \hat{\gamma} + t_{n-p}S_{\hat{\gamma}}$ (2.2.8)

waarin $S_{\hat{\gamma}}^2 = \frac{\mathcal{S}_\Omega}{n - p} \cdot T_{02}$ volgens (1.4.15); immers $\frac{\mathcal{S}_\Omega}{n - p}$ is een schatting van σ^2 en T_{02} is B.

2.2.2 Numeriek voorbeeld.

Voor een numeriek voorbeeld zullen gegevens uit een artikel van W. Gerstmann "Regressionsbeziehungen für Turbulenz parameter" worden gebruikt. In Abb. 2 van dit artikel geeft Gerstmann meetresultaten van de turbulente warmtestroom H voor onstabiele situaties in vergelijking met een soort Richardsongetal $\frac{U_4 - U_{0,25}}{t_{0,25} - t_4} = r$,

waarin U_i de windsnelheid is op i meter hoogte en t_i de temperatuur op dezelfde hoogte. Er wordt lineaire regressie-analyse toegepast en getoetst of regressiecoëfficiënt en constantsterm significant van nul afwijken; wat echter niet getoetst wordt is of er misschien een niet lineaire relatie zou kunnen zijn. Dit zal als voorbeeld volgens de in 2.2.1 geschetste methode worden uitgevoerd.

Aangezien de basisgegevens waaruit Gerstmann Abb. 2 samenstelde, niet in het artikel voorkomen zijn de waarden van de grootheden H en r in centimeters uit de figuur afgelezen en y en x genoemd, waarbij $x \sim 2{,}5\,\frac{U_4 - U_{0,25}}{r_{0,25} - t_4}$ en $y \sim 25H$ is. In bijgaande figuur 1 zijn x en y tegen elkaar uitgezet.

Het regressiemodel Ω: $y = \alpha + \beta x + \gamma x^2 + \varepsilon$ wordt nu aangepast. De parameters α, β en γ worden volgens (2.2.3) berekend, met als resultaat:

$\hat{\alpha} = 0{,}206 \qquad \hat{\beta} = 1{,}608 \qquad \hat{\gamma} = -0{,}093$

terwijl volgens (1.4.12) de standaard deviaties zijn:

$S_{\hat{\alpha}} = 0{,}507 \qquad S_{\hat{\beta}} = 0{,}347 \qquad S_{\hat{\gamma}} = 0{,}042$.

Aangezien het de bedoeling is te toetsen of het lineaire model ω: $y = \alpha + \beta x + \varepsilon$ voldoende is, wordt de nulhypothese H_o: $\psi = \gamma = 0$ ingevoerd.

Voor de F-toets zijn de numerieke resultaten nu:

$\mathcal{S}_\Omega = 101{,}109$ berekend volgens (2.2.4)

$\mathcal{S}_\omega - \mathcal{S}_\Omega = 12{,}230$ berekend volgens (2.2.7)

zodat $\dfrac{n-p}{q} \cdot \dfrac{\mathcal{S}_\omega - \mathcal{S}_\Omega}{\mathcal{S}_\Omega} = \dfrac{43{,}3}{1} \cdot \dfrac{12{,}230}{101{,}109} = 4{,}84.$

De overschrijdingskans van deze waarde is volgens de $F_{1,40}$ verdeling ongeveer 4%. Bij een 5% significantiedrempel moet de nulhypothese $\gamma = 0$ dus verworpen worden, met andere woorden de onderstelling dat men met een lineaire relatie kan volstaan moeten worden verworpen ten gunste van een kwadratische relatie. Men verkrijgt hetzelfde resultaat als gebruik gemaakt wordt van het vertrouwensinterval volgens (1.6.2). De 5% t_{40} waarde is namelijk 2,02 zodat $\hat{\psi} \pm t_{n-p} S_{\hat{\psi}}$ wordt $-0{,}093 \pm 2{,}02 \times 0{,}042$. Dus $-0{,}178 < \psi = \gamma < -0{,}008$. De waarde $\gamma = 0$ ligt buiten het vertrouwensinterval; de nulhypothese moet verworpen worden.

$\mathcal{S}_\omega$ is ook rechtstreeks te berekenen als $y = \tilde{\alpha} + \tilde{\beta} x$ wordt aangepast. Er wordt dan gevonden $\tilde{\alpha} = 0{,}973$ en $\tilde{\beta} = 0{,}878$ met $S_{\tilde{\alpha}} = 0{,}386$ en $S_{\tilde{\alpha}} = 0{,}105$ terwijl dan $\mathcal{S}_\omega$ berekend volgens $\mathcal{S}_\omega = ||\vec{y} - \hat{\vec{\eta}}||^2$ gelijk aan 133,335 is.

In het voorgaande werd berekend $\mathcal{S}_{\Omega} = 101,109$ en $\mathcal{S}_{\omega} - \mathcal{S}_{\Omega} = 12,230$, zodat $\mathcal{S}_{\omega} = 101,109 + 12,230 = 113,339$; dit klopt dus zeer fraai met de rechtstreekse berekening.

In figuur 1 is zowel het lineaire verband $y = 0,973 + 0,878x$, als het kwadratische $y = 0,206 + 1,608x - 0,093x^2$ getekend. Het lineaire verband moet natuurlijk hetzelfde zijn als Gerstmann in Abb. 2 geeft. Dit klopt op kleine verschillen na, die het gevolg zijn van afleesonnauwkeurigheden.

Tenslotte kan men zich nog afvragen of het nodig is een constante term in de regressievergelijking mee te nemen. Het vertrouwensinterval (2.2.8) voor α is $0,206 - 2,02 \times 0,507 \leq \alpha < 0,206 + 2,02 \times 0,507$ of $-0,892 < \alpha < 1,230$ zodat er geen reden is $\alpha \neq 0$ te stellen. Wordt nu $y = \beta^* x + \gamma^* x^2$ aangepast, dan blijkt $\beta^* = 1,723$ en $\gamma^* = -0,105$. De kromme $y = 1,723x - 0,105x^2$ is eveneens in figuur 1 getekend.

3. Een bewerkingstechniek voor multivariate analyse.

3.1. Inleiding.

In hoofdstuk 1 is de oplossing gezocht voor de beste schatting van een lineaire relatie tussen een grootheid y en p grootheden x_i. De geometrische voorstelling gaf aan, dat deze oplossing te vinden is door in een n-dim. ruimte een vector y te projecteren op een p-dim. deelruimte V_p gevormd door p-vectoren $x_i (i = 1,\ldots,p)$. In feite is er daarbij, zoals reeds in 1.2. is gesteld, sprake van de ontbinding van een vector $\vec{y}$ in een component evenwijdig aan V_p ($\vec{y}_{//}$) en een component loodrecht op V_p ($\vec{y}_{\perp}$). De eerste geeft de "aanpassing" aan het lineaire model en de tweede is een maat voor de afwijking tussen waarneming en model.

Nu is het meestal zo dat de p-grootheden x_i niet allen even belangrijk zijn, dat wil zeggen, dat de variatie van y grotendeels bepaald wordt door een beperkt aantal x_i's terwijl de anderen nauwelijks een rol spelen. Voor het bepalen van de x_i's die tezamen een optimaal resultaat geven ter "verklaring" van de variatie van y zijn diverse berekeningsmethoden ontwikkeld, onder anderen door Efroymson. Hierop zal in 4 nader worden ingegaan.

Voor de berekeningen van dergelijke problemen zijn in het algemeen zeer grote aantallen bewerkingen nodig, zodat een rekenmachine een noodzakelijk hulpmiddel is.

Het gaat daarbij dan om de oplossing van een stelsel lineaire vergelijkingen, waarvoor een vector, matrixvorm in (1.4.7) was gevonden, namelijk: $\hat{\vec{\beta}} = S^{-1}X\vec{y}$ (met voorwaarde, dat de matrix S nonsingulier is).

Mathematisch gezien is hiermede het probleem opgelost. Numeriek kunnen er echter moeilijkheden ontstaan, als de matrix S slechts geconditioneerd is, dat wil zeggen, dat kleine variaties in de componenten van de vector y tot zeer grote variaties in de β leiden; n.l. hetzij ten gevolge van de afrondingsfouten die bij de vele computerbewerkingen cumulatief een groot effect kunnen hebben, hetzij omdat ergens nagenoeg even grote getallen van elkaar worden afgetrokken, waardoor de relatieve fout zeer groot wordt. Er zijn methoden ontwikkeld waarmee deze moeilijkheden kunnen worden vermeden. Hierbij wordt de vectorvoorstelling van het probleem door ontbinding van de vectoren in een onderling orthogonaal stelsel getransformeerd.

Voor het regressieprobleem, dat in dit rapport aan de orde, is ook orthogonalisatie toegepast maar niet zozeer om het conditieprobleem te ontlopen, maar om over een rekentechniek te beschikken waarmee de optimale keuze van het beperkte aantal x_i's kan worden bereikt.

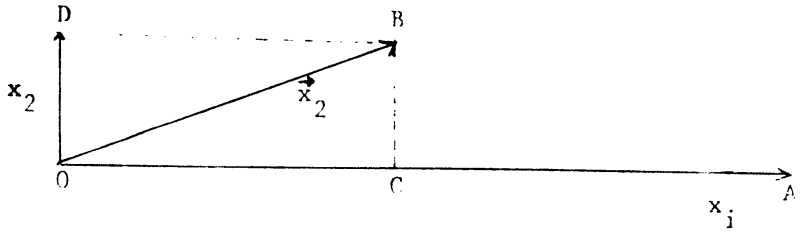
Het conditieprobleem is bij het relatief kleine aantal matrixelementen in dit geval doorgaans nauwelijks aanwezig (zie Orden p. 41).

In het volgende wordt nu de techniek, die de in het voorgaande genoemde orthogonalisatie realiseert, uiteengezet. Daarbij zullen ook de standaarddeviaties van de geschatte parameters van de lineaire relaties worden verkregen. Een uitvoerige beschrijving van de methode is te vinden in: A.P. Dempster: "Elements of continous multivariate analysis".

3.2. Orthogonalisatie.

Uitgangspunt is een p-dim. ruimte V_p opgespannen door p vectoren $\vec{x}_i (i = 1,\ldots,p)$. Deze p vectoren dienen te worden vervangen door p vectoren $\vec{x}_i^*$ die alle onderling loodrecht op elkaar staan en dezelfde ruimte V_p bepalen. De $\vec{x}_i$ vormen een willekeurig basis voor V_p; de $\vec{x}_i^*$ een orthogonale basis van V_p. Men kan de vectoren $\vec{x}_i^*$ achtereenvolgens als volgt vinden: $\vec{x}_1$ wordt $\vec{x}_1^*$ genoemd, daarna wordt x_2 ontbonden in een vector evenwijdig $\vec{x}_1 = \vec{x}_1^*$ en één loodrecht op $\vec{x}_1 = \vec{x}_1^*$. Deze laatste wordt $\vec{x}_2^*$ genoemd. $\vec{x}_3^*$ wordt bepaald als de component van x_3 die loodrecht is $\vec{x}_2^*$ zowel als op $\vec{x}_1^*$. Op overeenkomstige wijze worden de overigen $\vec{x}_i^*$ verkregen.

Een uitdrukking voor dit proces is te vinden met behulp van het inproduct van twee vectoren, zie bijgaande figuur:



Als OC de projectie van $\vec{x}_2$ op $\vec{x}_1 = \vec{x}_1^*$ is, dan is de component van $\vec{x}_2$ die $\perp \vec{x}_1^*$ is, te vinden door $\vec{x}_2$ met OC te verminderen; resultaat: $OD = \vec{x}_2^*$. OC als vector is te schrijven als $\frac{OC}{OA} \cdot \vec{x}_1^*$. Nu is het inproduct $(\vec{x}_2 \vec{x}_1^*) =$ OC x OA terwijl $(\vec{x}_1^* \vec{x}_1^*) = OA^2$. Dus is $\frac{OC}{OA} = \frac{(\vec{x}_2 x_1^*)}{(\vec{x}_1^* \vec{x}_1^*)}$. Zo ontstaat $\vec{x}_2^* = \vec{x}_2 - \frac{(\vec{x}_2 \vec{x}_1^*)}{(\vec{x}_1^* \vec{x}_1^*)} X_1$.

In het algemeen: $$\vec{x}_r^* = \vec{x}_r - \sum_{j=1}^{r-1} \frac{(\vec{x}_r \vec{x}_j^*)}{(\vec{x}_j^* \vec{x}_j^*)} \vec{x}_j^* \qquad (3.2.1)$$

Of ook: $$\vec{x}_r = \vec{x}_r^* + \sum_{j=1}^{r-1} \frac{(x_r x_j^*)}{(x_j^* x_j^*)} x_j \qquad (3.2.2)$$

Voeren we de matrix M in, volgens:

$$M = \begin{bmatrix} 1 & & & \\ \frac{(\vec{x}_2 \vec{x}_1^*)}{(\vec{x}_1^* \vec{x}_1^*)} & 1 & & \\ \frac{(\vec{x}_3 \vec{x}_1^*)}{(\vec{x}_1^* \vec{x}_1^*)} & \frac{(\vec{x}_3 \vec{x}_2^*)}{(\vec{x}_2^* \vec{x}_2^*)} & 1 & \\ \vdots & & & \ddots \\ \frac{(\vec{x}_p \vec{x}_1^*)}{(\vec{x}_1^* \vec{x}_1^*)} & \cdots & \cdots & 1 \end{bmatrix}$$

dan is te schrijven: $\vec{x} = M\vec{x}^*$ (3.2.4)

en $\vec{x}^* = N\vec{x}$ met $N = M^{-1}$ (3.2.5)

Een matrix Q waarvan de elementen de inproducten $(\vec{x}_i \vec{x}_j)$ zijn, is dan te schrijven als:

$$Q = \{\vec{x}\vec{x}'\} = M\{x\ x'\} = MTM' \tag{3.2.6}$$

waarin T een diagonaalmatrix met elementen $(\vec{x}_j^* \vec{x}_j^*)$ in de diagonaal is; de elementen $(\vec{x}_i^* \vec{x}_j^*)$ zijn immers nul wegens de orthogonaliteit van de $\vec{x}_i^*$.

De matrix Q is het uitgangsgegeven voor een operatie waarbij stap voor stap de orthogonalisatie wordt doorgevoerd. Als met s het aantal vectoren $\vec{x}_i$ wordt aangegeven dat op zeker moment georthogonaliseerd is, dan zal een operator worden gedefinieerd die uitgaande van de beginsituatie de orthogonalisatie tot een willekeurige s tot stand brengt. Daaruit is de operator af te leiden die het proces van s naar s + 1 voert. Het ligt voor de hand voor vectoren en matrices een splitsing in te voeren op de volgende wijze:

Als v en K respectievelijk een willekeurige (p x 1) vector en (p x p) matrix is:

$$v = \begin{pmatrix} \vec{v}_1 \\ \vec{v}_2 \end{pmatrix} = \left(\begin{pmatrix} v_1 \\ \vdots \\ v_s \end{pmatrix} \begin{pmatrix} v_{s+1} \\ \vdots \\ v_p \end{pmatrix} \right) \text{en} \quad K = \begin{pmatrix} K_{11}^{\ s \times s} & K_{12}^{\ s \times (p-s)} \\ K_{21}^{\ (p-s) \times s} & K_{22}^{\ (n-s) \times (p-s)} \end{pmatrix} \tag{3.2.7}$$

Als deze splitsing op (3.2.4) wordt toegepast ontstaat:

$$\begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \end{pmatrix} = \begin{pmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} \vec{x}_1^* \\ \vec{x}_2^* \end{pmatrix} = \begin{pmatrix} M_{11}x_1 \\ M_{21}x_1 + M_{22}x_2 \end{pmatrix}$$

waarbij: $\vec{x}_1 = \begin{pmatrix} x_1 \\ \vdots \\ x_s \end{pmatrix}$ en $\vec{x}_2 = \begin{pmatrix} x_{s+1} \\ \vdots \\ x_p \end{pmatrix}$.

Dus: $\vec{x}_1 = M_{11}\vec{x}_1^*$ (3.2.8)

$\vec{x}_2 = M_{21}\vec{x}_1^* + M_{22}\vec{x}_2^*$ (3.2.9)

Nu is iedere vector van $\vec{x}_2^*$ loodrecht op iedere vector, die in $\vec{x}_1^*$ is begrepen, dus loodrecht op de ruimte door deze vectoren $\vec{x}_1^*$ gevormd en dit is dezelfde ruimte die door de $\vec{x}_1$ vectoren wordt opgespannen. (3.2.9) betekent dus een splitsing van de $\vec{x}_2$ vectoren in componenten $M_{21}\vec{x}_1^*$ die evenwijdig zijn aan de ruimte van de $\vec{x}_1$ vectoren en componenten $M_{22}\vec{x}_2^*$ die loodrecht op die ruimte staan. Hiervoor wordt de notatie:

$\vec{x}_{2.1} = M_{22}\vec{x}_2^*$ (3.2.10)

ingevoerd.

Verder geldt: MN = I. Dus:

$$\begin{pmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{pmatrix} \cdot \begin{pmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{pmatrix} = \begin{pmatrix} N_{11}N_{11} & M_{11}N_{12} \\ M_{21}N_{11} + M_{22}N_{21} & M_{21}N_{12} + M_{22}N_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

Dus $M_{11}N_{11} = I$, derhalve $N_{11} = M_{11}^{-1}$ (3.2.11)

$M_{11}N_{12} = 0$, wegens $M_{11} \neq 0$ dus $N_{12} = 0$ (3.2.12)

$M_{21}N_{12} + M_{22}N_{22} = M_{22}N_{22} = I$ waaruit $N_{22} = -M_{22}^{-1}$ (3.2.13)

$M_{21}N_{11} + M_{22}N_{21} = 0$ of $M_{21}N_{11} = -M_{22}N_{21}$

Voor de laatste formule zal ingevoerd worden de notatie H_{21}:

$H_{21} = M_{21}N_{11} = -M_{22}N_{21}$ (3.2.14)

Nu zijn N_{11} en M_{22} beide driehoeksmatrices met diagonaal elementen die alle 1 zijn, derhalve is uit (3.2.11) af te lezen:

(i) De eerste rij van H_{21} is gelijk aan het tegengestelde van de eerste rij van N_{21}.

(ii) De laatste kolom van H_{21} is gelijk aan de laatste kolom van M_{21}.

Uit (3.2.9) en (3.2.10) volgt: $\vec{x}_2 = M_{21}\vec{x}_1^* + \vec{x}_{2.1}$

Uit (3.2.8) en (3.2.11) : $\vec{x}_1^* = M_{11}^{-1}\vec{x}_1 = N_{11}\vec{x}_1$

Dus: $\vec{x}_{2.1} = \vec{x}_2 - M_{21}\vec{x}_1^* = \vec{x}_2 - M_{21}N_{11}\vec{x}_1 = \vec{x}_2 - H_{21}\vec{x}_1$ (3.2.15)

3.3. De "Sweep" Operator.

Nu kan de operator ingevoerd worden die de overgang van begin tot de toestand s moet bewerkstellen. Hiervoor zal de notatie SWP, die "sweep operator" [1] betekent, worden gebruikt:

$$SWP(0, 1,\ldots,s)\ Q = SWP(0, 1,\ldots,s) \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} = \begin{pmatrix} -Q_{11}^{-1} & Q_{11}^{-1}Q_{12} \\ Q_{21}Q_{11}^{-1} & Q_{22}-Q_{21}Q_{11}^{-1}Q_{12} \end{pmatrix}$$

met $SWP(0)\ Q = Q$ en $SWP(0, 1,\ldots,p)\ Q = -Q^{-1}$ (3.3.1)

Als het splitsingsprocedé op (3.2.6) wordt toegepast, ontstaat:

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} = \begin{pmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{pmatrix} \cdot \begin{pmatrix} T_{11} & 0 \\ 0 & T_{22} \end{pmatrix} \cdot \begin{pmatrix} M'_{11} & M'_{21} \\ 0 & M'_{22} \end{pmatrix} =$$

$$\begin{pmatrix} M_{11}T_{11}M'_{11} & M_{11}T_{11}M'_{21} \\ M_{21}T_{11}M'_{11} & M_{21}T_{11}M'_{21} + M_{22}M'_{22} \end{pmatrix}$$

Hierop SWP(0, 1,...,s) toepassen levert:

[1] Deze term is volgens Dempster afkomstig van A.E. Beaton (1964)

$$\text{SWP}(0, 1,\ldots,s)\ Q = \begin{pmatrix} -M_{11}'^{-1}T_{11}^{-1}M_{11}^{-1} & M_{11}^{-1}M_{21}' \\ M_{21}M_{11}^{-1} & M_{22}T_{22}M_{22}' \end{pmatrix}$$

Nu is $M_{11}'T_{11}M_{11} = M_{11}'(\vec{x}_1^{*\prime}\vec{x}_1^{*})M_{11} = (\vec{x}_1'\vec{x}_1) = Q_{11}$.
Verder is volgens (3.2.11) en (3.2.13): $M_{21}M_{11}^{-1} = H_{21}$ en
$M_{11}^{-1}M_{21}' = H_{21}'$ en tenslotte met (3.2.10):
$M_{22}T_{22}M_{22}' = M_{22}(\vec{x}_2^{*\prime}\vec{x}_2^{*})M_{22}' = (\vec{x}_{21} \cdot \vec{x}_{21}')$ hiervoor invoeren
$M_{22.1}$, dus: $M_{22.1} = (\vec{x}_{2.1} \cdot \vec{x}_{2.1}') = M_{22}T_{22}M_{22}'$ (3.3.2)
hieruit zien we dat het element op de eerste rij en in de eerste kolom van $M_{22.1}$ gelijk is aan het eerste diagonaal element van T_{22}. Nu kan geschreven worden:

$$\text{SWP}(0, 1,\ldots,s)\ M = \begin{pmatrix} -M_{11}^{-1} & H_{21}' \\ H_{21} & M_{22.1} \end{pmatrix} \qquad (3.3.3)$$

Uit het voorgaande blijkt dus dat bij toepassing van SWP(0, 1,...,s)M, voor achtereenvolgens s = 1,...,p, uit de submatrices H_{21} de transformatie matrices M en N zijn af te leiden met de regels (i) en (ii), terwijl uit de submatrix $M_{22.1}$ de matrix T volgt.

Als de formule (3.2.15) als $\vec{x}_2 = \vec{x}_{2.1} + H_{21}\vec{x}_1$ geschreven wordt, dan is in te zien dat de matrix H_{21} in zijn rijen de coëfficiënten levert van de projecties van de vectoren uit $\vec{x}_2$ op de ruimte van de geschatte regressiecoëfficiënten tussen één der x_j (j= s +1,...,p) en alle x_k (k=1,...,s).

Verder leveren de diagonaalelementen van $M_{22.1}$ de kwadraten van de componenten van de $\vec{x}_2$ vectoren die loodrecht op $\vec{x}_1$ staan; dit is in de regressie-analyse de rest variantie nadat de regressie tussen de $\vec{x}_2$ en de $\vec{x}_1$ grootheden is tot stand gebracht. Tenslotte levert M_{11}^{-1} informatie over de standaarddeviatie van de regressiecoëfficiënten (zie (1.4.12). M is gelijk aan de matrix S. Dit zal nog nader worden bekeken.

Eerst moet nu nog worden aangegeven hoe de opeenvolgende operaties SWP(0, 1,...,s) voor s = 1,...,p uit elkaar worden afgeleid. Hiertoe de reciproke operator RSW invoeren volgens:

$$\mathrm{RSW}(0,1,\ldots,s)\begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} = \begin{pmatrix} -M_{11}^{-1} & -M_{11}^{-1}M_{12} \\ -M_{21}M_{11}^{-1} & M_{21}-M_{21}M_{11}^{-1}M_{12} \end{pmatrix} \qquad (3.3.4)$$

Gemakkelijk is in te zien dat inderdaad:

$\mathrm{RSW}(0,1,\ldots,s)\ \mathrm{SWP}(0,1,\ldots,s)\ M = M$

evenals:

$\mathrm{SWP}(0,1,\ldots,s)\ \mathrm{RSW}(0,1,\ldots,s)\ M = M$

Er wordt nu een SWP(s) ingevoerd waarvoor geldt:

$\mathrm{SWP}(s)\ \mathrm{SWP}(0,1,\ldots,s-1) = \mathrm{SWP}(0,1,\ldots,s)$

en dus:

$\mathrm{SWP}(s)\ M = \mathrm{SWP}(0,1,\ldots,s)\ \mathrm{RSW}(0,1,\ldots,s-1)M$

M eerst als volgt schrijven:

$$M = \begin{pmatrix} \tilde{M}_{11} & m' & \tilde{M}_{12} \\ m & c & r \\ \tilde{M}_{21} & r' & \tilde{M}_{22} \end{pmatrix}$$

Waarin $\tilde{M}_{11}$ een (s-1) x (s-1) matrix,
m een 1 x (s-1) matrix,
r een 1 x (p-s) matrix en
c één enkel element is.

Dan is:

$$\mathrm{RSW}(0,1,\ldots,s-1)M = \begin{pmatrix} -\tilde{M}_{11}^{-1} & -\tilde{M}_{11}^{-1}m' & -\tilde{M}_{11}^{-1}\tilde{M}_{12} \\ -m\tilde{M}_{11}^{-1} & c-m\tilde{M}_{11}m' & r-m\tilde{M}_{11}^{-1}\tilde{M}_{12} \\ -\tilde{M}_{21}\tilde{M}_{11}^{-1} & r'-\tilde{M}_{21}\tilde{M}_{11}m' & M_{22}-\tilde{M}_{21}\tilde{M}_{11}^{-1}\tilde{M}_{12} \end{pmatrix}$$

Nu hierop SWP(0, 1,...,s) toepassen; hiervoor is eerst de reciproke nodig van de matrix:

$$\begin{pmatrix} -M_{11}^{-1} & -M_{11}^{-1}m' \\ -mM_{11}^{-1} & c-mM_{11}^{-1}m' \end{pmatrix}$$

Stel dat deze te schrijven is als $\begin{pmatrix} A & B' \\ B & D \end{pmatrix}$ dan moet gelden:

$$\begin{pmatrix} A & B' \\ B & D \end{pmatrix} \cdot \begin{pmatrix} -\tilde{M}_{11}^{-1} & -\tilde{M}_{11}^{-1}m' \\ -m\tilde{M}_{11}^{-1} & c-m\tilde{M}_{11}^{-1}m' \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

of:

$$\begin{pmatrix} -A\tilde{M}_{11}^{-1}-B'm\tilde{M}_{11}^{-1} & -A\tilde{M}_{11}^{-1}m' + B'(c-m\tilde{M}_{11}^{-1}m') \\ -B\tilde{M}_{11}^{-1}-Dm\tilde{M}_{11}^{-1} & -B\tilde{M}_{11}^{-1}m' + D(c-m\tilde{M}_{11}^{-1}m') \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

A en B oplossen uit gelijkstelling van de elementen van de eerste rijen geeft $B' = -mc^{-1}$ en $A = -\tilde{M}_{11} + m'c^{-1}m$.

B en D uit elementen van de tweede rijen geeft $D = c^{-1}$ en $B = -c^{-1}m$.

Derhalve:
$$\begin{pmatrix} -\tilde{M}_{11}^{-1} & -\tilde{M}_{11}^{-1}m' \\ -m\tilde{M}_{11}^{-1} & c-m\tilde{M}_{11}^{-1}m' \end{pmatrix}^{-1} = \begin{pmatrix} -\tilde{M}_{11}^{-1} + m'c^{-1}m & -m'c^{-1} \\ -c^{-1}m & c^{-1} \end{pmatrix}$$

SWP(0, 1,...,s) RSW(0, 1,...,s-1)M wordt nu:

$$\text{SWP(s)M} = \begin{pmatrix} \tilde{M}_{11}^{-1}-m'c^{-1}m & m'c^{-1} & \tilde{M}_{12}-m'c^{-1}r \\ c^{-1}m & -c^{-1} & c^{-1}r \\ \tilde{M}_{21}-r'c^{-1} & r'c^{-1} & Q_{22}-r'c^{-1}r \end{pmatrix}$$

Met het bovenstaande is de operator SWP(s) afgeleid die de opeenvolgende operaties SWP(0, 1,...,s) voor s = 1,.... uit elkaar kan afleiden. Deze is ook als volgt te definieren, als m_{ij} de elementen van M en $\tilde{m}_{ij}$ die van SWP(s)M resp. RSW(s)M zijn:

$$
\text{SWP}(s) = \left\{ \begin{array}{lll} \tilde{m}_{ss} = & -1/m_{ss} & \\ \tilde{m}_{is} = & m_{is}/m_{ss} & \\ \tilde{m}_{sj} = & m_{sj}/m_{ss} & i \text{ en } j \neq s \\ \tilde{m}_{ij} = & m_{ij} - m_{is}m_{sj}/m_{ss} & \end{array} \right.
$$

en

$$
\text{RSW}(s) \left\{ \begin{array}{lll} \tilde{m}_{ss} = & -1/m_{ss} & \\ \tilde{m}_{is} = & -m_{is}/m_{ss} & \\ \tilde{m}_{sj} = & -m_{sj}/m_{ss} & \\ \tilde{m}_{ij} = & m_{ij} - m_{is}m_{jj}/m_{ss} & \end{array} \right. \qquad (3.3.5)
$$

N.B. Door Dempster wordt een geheel andere volgorde bij de afleiding van de methode aangehouden. Eerst wordt de overgang van s naar s + 1 in het proces moeizaam omschreven, waarbij nagegaan wordt, wat er met de elementen van M_{11}^{-1}, H_{21} en $M_{22.1}$ gebeurt bij de orthogonalisatie. Daarna wordt de operator (3.3.5) ingevoerd en wordt aangetoond, dat deze aan de eisen voldoet. (3.3.3) komt ergens uit de lucht vallen, hiervan blijkt pas later de bedoeling, terwijl tenslotte (3.3.1) als eigenschap blijkt te gelden.

Alternatieve Sweepoperatie definitie.

In afwijking van de definitie, zoals die door Dempster is ingevoerd, kan men ook

$$
\text{SWP}^{*}(0, 1, \ldots, s) \begin{pmatrix} M_{21} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} = \begin{pmatrix} M_{11}^{-1} & M_{11}^{-1}M_{12} \\ -M_{21}^{-1}M_{11}^{-1} & M_{22} - M_{21}M_{11}^{-1} \quad M_{12} \end{pmatrix}
$$

definieren. En

$$
\text{SWP}^{*}(s) : \left\{ \begin{array}{lll} \tilde{m}_{ss} = & 1/m_{ss} & \\ \tilde{m}_{is} = & -m_{is}/m_{ss} & \\ \tilde{m}_{sj} = & m_{sj}/m_{ss} & i; j \neq s \\ \tilde{m}_{ij} = & m_{ij}m_{is}m_{sj}/m_{ss} & \end{array} \right. \qquad (3.3.6)
$$

Dit heeft het voordeel dat nu geen afzonderlijke inverse operator behoeft te worden ingevoerd. $SWP^*(s)\ SWP^*(s)\ M = M$ zoals gemakkelijk is in te zien. Wel moet bedacht worden, dat de rijen van H_{21} nu de regressiecoëfficiënt met tegengesteld teken leveren.

4. Toepassing van de Sweepoperator techniek bij "optimalisering" van multiple lineaire regressie.

4.1 Inleiding.

Van de techniek die in 3 is ontwikkeld, wordt gebruik gemaakt voor de oplossing van het probleem van de multiple regressie, waarbij het er om gaat voor de regressie van y op $x_1,\ldots,x_p$ te bepalen, welke van de p-grootheden x_1 men in de regressie moet betrekken of met andere woorden, welke x_i's geven een significante bijdrage ter verklaring van de variatie van y en welke x_i's kunnen buiten beschouwing blijven.

De optimale oplossing is natuurlijk die welke mogelijke combinaties van 1 tot en met p variabelen uit de p x_i's in beschouwing neemt en voor iedere combinatie toetst of de nulhypothese, dat de regressiecoëfficiënten voor de niet in de combinatie voorkomende x_i's gelijk aan nul zijn, moet worden verworpen. Van de combinaties waarvoor deze nulhypothese niet verworpen behoeft te worden kiest men dan die welke de kleinste restvariantie geeft. Nu is dat aantal combinaties 2^p-1 en dit wordt bijvoorbeeld bij p = 10 al meer dan 1000. De hoeveelheid rekenwerk wordt dan zeer groot. Een methode die minder rekenwerk eist is bijvoorbeeld door Efroymson beschreven in het artikel "Multiple Regression Analysis".

In de eerste plaats zal hier een korte algemene beschrijving van de methode worden gegeven, daarna zullen in 4.4 enkele details en onduidelijkheden in het artikel van Efroymson nader worden besproken.

4.2. Algemene beschrijving van de "optimaliserings" methode.

De methode bestaat in het successievelijk toevoegen aan een verwijdering uit de regressie van x_i's en wel als volgt:

In ieder stadium wordt nagegaan hoe groot de restvariantie-reductie is die de niet in de regressie meespelende x_i's ieder afzonderlijk bij toevoegen zouden veroorzaken. Van de maximale waarde van deze restvariantie-reductie wordt getoetst of de bijdrage significant is. Is dit het geval dan wordt de betrokken x_i in de regressie opgenomen. Vervolgens wordt nagegaan of de reeds meespelende x_i's moeten blijven meedoen. Het is namelijk zo, dat door toevoeging van een nieuwe x_i de bijdrage van de reeds aanwezig x_i's verandert, in het algemeen kleiner wordt. Het is dus mogelijk dat een bepaald x_i nu niet langer een significante bijdrage levert. Van de meedoende x_i's wordt dus nagegaan welke de kleinste restvariantie reductie geeft en vervolgens wordt getoetst of deze reductie significant is; zo ja, dan wordt de betrokken x_i uit de regressie verwijderd. Daarna wordt opnieuw onderzocht of nog een nieuwe x_i moet worden toegevoegd. Dit proces wordt zolang voortgezet totdat van de niet meespelende x_i's er geen meer is die een significante bijdrage tot restvariantie reductie kan leveren.

De vraag die nu naar voren komt is natuurlijk: "Vindt men op bovenstaande wijze ook de optimale oplossing?" Helaas is op deze vraag geen antwoord te geven. Het is mogelijk dat bepaalde gunstige combinaties in bovenstaande methode niet aan bod komen. Te verwachten is wel dat deze heen en weer gaande methode een beter resultaat geeft dan 1° de simpele "voorwaarts"methode die alleen maar stopsgewijs nagaat welke x_i's moeten worden toegevoegd, dan wel 2° de "achterwaarts"methode die, uitgaande van volledige regressie op alle p x_i's nagaat welke x_i verwijderd kunnen worden.

4.3. Toetsing tijdens het "optimaliserings"procedé.

Uitgaande van Ω: $y_k = \sum_{i=1}^{p} b_i x_{ik} + \varepsilon_k$; $\varepsilon_k \Rightarrow N(0,\sigma^2)$; $E\varepsilon_k\varepsilon_l = 0$;

$k = 1,\ldots,n$

wordt de matrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{p1} & \cdots & x_{pn} \\ y_1 & \cdots & y_n \end{pmatrix}$$

gevormd, waaruit:

$$M = XX' = \begin{pmatrix} \sum_{j=1}^{n} x_{1j}^2 & \sum_{j=1}^{n} x_{1j}x_{2j} & \cdots & \sum_{j=1}^{n} x_{1j}\Sigma x_{pj} & \sum_{j=1}^{n} x_{1j}y_j \\ \vdots & & & \vdots & \vdots \\ \sum_{j=1}^{n} x_{pj}x_{1j} & \cdots & & \sum_{j=1}^{n} y_j x_{pj}^2 & \sum_{j=1}^{n} x_{pj}y_j \\ \sum_{j=1}^{n} y_j x_{1j} & \cdots & & \sum_{j=1}^{n} y_j x_{pj} & \Sigma y_j^2 \end{pmatrix}$$

Stel nu dat in een zeker stadium van de bewerking de grootheden $x_{r_1},\ldots,x_{r_t}$ in de regressie zijn betrokken, dat wil zeggen, dat $SWP^{*}(0, r_1,\ldots,r_t)$ is uitgevoerd. M is veranderd in een matrix voor te stellen is door: $\tilde{M}^t = SWP\ (0, r_1,\ldots,r_t)M$

$$\begin{pmatrix} (\tilde{M}^t)^{-1} & \tilde{H}_{12}^t \\ -\tilde{H}_{21}^t & \tilde{M}_{22.1}^t \end{pmatrix}$$

waarvoor (3.3.6) is gebruikt.

De elementen van M zullen door m_{ij} en die van $\tilde{M}^{t}$ door $\tilde{m}^{t}_{ij}$ worden voorgesteld, derhalve: als i en j beide tot $r_1,\ldots,r_l$ behoren dan $\tilde{m}^{t}_{ij} \in M_{11}^{-1}$; als i wel en j niet tot $r_1,\ldots,r_l$ behoren dan $\tilde{m}^{t}_{ij} \in H_{12}$, enz.

Er wordt nu voor alle s de waarde berekend van $V_s = \frac{\tilde{m}^{t}_{p+1,s}\,\tilde{m}^{t}_{s,p+1}}{\tilde{m}_{s}\,\tilde{m}_{s}}$. Dit is volgens (3.3.6) de verandering die de term $\tilde{m}^{t}_{p+1,p+1}$ ondergaat als SWP*(s) op $\tilde{M}^{t}$ wordt toegepast en dit is dus volgens de regressie-analyse de verandering van de restvariantie. Behoort s niet tot $r_1,\ldots,r_t$ dan is x_s niet in de regressie opgenomen en zijn $\tilde{m}^{t}_{p+1,s}$ en $\tilde{m}^{t}_{s,p+1}$ elementen van $\tilde{M}^{t}_{22.1}$ en dus aan elkaar gelijk, derhalve is $V_s > 0$; behoort s wel tot $r_1,\ldots,r_t$ dan geldt $\tilde{m}^{t}_{p+1,s} \in -\tilde{H}^{t}_{21}$ en $\tilde{m}^{t}_{s,p+1} \in \tilde{H}^{t}_{12}$ dus $\tilde{m}^{t}_{p+1,s} = -\tilde{m}^{t}_{s,p+1}$ derhalve $V_s < 0$. Toepassing van SWP(s) betekent in het geval s niet tot $r_1,\ldots,r_t$ behoort toevoeging van x_s aan de regressie en volgens (3.3.6) dus verkleining van $m_{p+1,p+1}$, dus verkleining van de restvariantie; terwijl toepassing op s als deze wel tot $r_1,\ldots,r_t$ behoort betekent dat x_s uit de regressie wordt verwijderd, waarbij de restvariantie dus toeneemt.

Is de situatie $r_1,\ldots,r_t$ onstaan na toevoeging van een nieuwe x_i, dan wordt voor alle negatieve V_i, het minimum $|V_m|$ van $|V_i|$ bepaald. Stel $|V_m|$ behoort bij x_m. Getoetst wordt dan of de variantiereductie nog significant van nul afwijkt (in feite wordt getoetst of de bijbehorende regressiecoëfficiënt van nul afwijkt).

Is dit zo, dan volgt verwijdering van x_m. Voor de toetsing geldt dat $\tilde{m}^{t}_{p+1,p+1}$ overeenkomt met $\mathscr{S}_{\Omega}$ (zie 1.2) en $|V_m|$ met $\mathscr{S}_{\omega} - \mathscr{S}_{\Omega}$, verder is p = t en q = 1, dus met (1.5.8) wordt x_m verwijderd als

$$\frac{(n-t)|V_m|}{\tilde{m}^{t}_{p+1,p+1}} < F_{(1,n-t)\alpha} \qquad (4.3.1)$$

waarbij $F_{(1,n-t)\alpha}$ de $\alpha\%$ waarde is van een $F_{1,n-t}$ verdeling.

Is de $r_1,\ldots,r_t$ ontstaan na verwijdering van een x_i, dan wordt uit de positieve V_i de maximale waarde bepaalt stel V_M vervolgens getoetst of deze restvariantiereductie significant groter dan nul is. Nu is $\mathcal{A}_\Omega$ niet zonder meer de reeds berekende $\tilde{m}^{t}_{p+1,p+1}$ maar de niet berekende $\tilde{m}^{t+M}_{p+1,p+1}$ die echter gelijk is aan $\tilde{m}^{t}_{p+1,p+1} - V_M$ Geldt nu: $\frac{(n-t-1).V_M}{\tilde{m}_{p+1,p+1}-V_M} > F_{(1,n-t-1)\alpha}$ (4.3.2)

dan volgt toevoeging van r_m aan de regressie. Is dit niet het geval dan stop het proces. (4.2.1) komt overeen met 17 in het artikel van Efroymson en (4.2.2) met 19.

Verwijdering van x_m wordt uitgevoerd door toepassing van SWP (r_m) en opname van X_M door SWP (r_M).

Is het proces klaar, dan kunnen de gewenste regressiecoëfficiënten afgelezen worden als $\hat{\beta}_i = -\tilde{m}^{t}_{p+1,r_i}$, waarin de elementen zijn uit de rij p + 1 van de submatrix $-H_{21}$.

De standaarddeviatie $S_{\tilde{y}}$ waarin $\tilde{y} = \tilde{y} - \sum_{i=1}^{t} \beta_{r_i} x_{r_i}$ (4.3.3)

is $\sqrt{\tilde{m}^{t}_{p+1,p+1}/t}$ dit is volgens (1.5.7) een zuivere schatting van σ.

De standaarddeviatie van $\hat{\beta}_i$ is volgens (1.4.12) te schatten als

$$S_{\hat{\beta}_i} = S_{\tilde{y}} \sqrt{\tilde{m}^{t}_{r_i r_i}} \qquad (4.3.4)$$

waarin $\tilde{m}^{t}_{r_i r_i}$ een element is van submatrix M_{11}^{-1}. die zoals reeds in 3.2. opgemerkt is, identiek aan S^{-1} in (1.4.12).

4.4. Enkele bijzonderheden in het artikel "Multiple Regression Analysis" van Efroymson.

1. Bij het onderstellen van lineaire regressie wordt veelal aangenomen, dat ook een constante term aanwezig kan zijn. Nu is dit in feite in $y = \sum_{r=1}^{p} \beta_r x_r$ opgesloten; men kan namelijk zonder bezwaar onderstellen dat bijvoorbeeld voor r=1 alle $x_1 \equiv 1$ zijn, dan zal β_1 de constante term voorstellen, waarvan de schatting $\hat{\beta}_o$ als $-m_{p+1,1}$ is af te lezen, terwijl ook de standaarddeviatie van $\hat{\beta}_o$ bepaald kan worden.

Door Efroymson wordt echter die constante term eerst verwijderd door y en x_i te vervangen door $y - \bar{y}$ en $x_i - \bar{x}_i$. Hierbij is $\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$, enz. De constante term wordt dan aan het eind van het proces als $\hat{\beta}_o = \bar{y} - \sum_{i=1}^{p} \hat{\beta}_{r_i} y_{r_i}$ bepaald. In dit geval krijgt men echter niet de standaarddeviatie van $\hat{\beta}_o$ ter beschikking. Efroymson definieert zijn uitgangsmatrix met de elementen

$$q_{ij} = \sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j).$$

Interessant is het op te merken, dat de matrix S voor de op de gemiddelde gereduceerde grootheden door toepassing van de SWP-operator uit de matrix van de ongereduceerde grootheden is af te leiden.

Stel deze laatste is

$$M = \begin{bmatrix} \Sigma x_{ok}^2 & \Sigma x_{ok}x_{1k} & \cdots & \Sigma x_{ok}y_{1k} \\ \vdots & & & \vdots \\ \Sigma y_k x_{ok} & \cdots & & \Sigma y_k^2 \end{bmatrix} = \begin{bmatrix} n & \Sigma x_1 & \cdots & \Sigma y \\ \Sigma x_1 & \Sigma x_1^2 & & \\ \vdots & & \ddots & \\ \Sigma y & \Sigma x_1 y & & \Sigma y^2 \end{bmatrix}$$

Waarin de constante term als een o^{de} term is toegevoegd in plaats van als x_1. Nu kan de Sweepoperator op het constante element worden toegepast :

$$SWP(0)M = \begin{bmatrix} -\frac{1}{n} & \bar{x}_1 & \cdots & \bar{x}_p & \bar{y} \\ \bar{x}_1 & \Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n} & & & \\ \vdots & & \ddots & & \\ \bar{y} & & & \Sigma y^2 - \frac{(\Sigma y)^2}{n} & \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{1}{n} & \bar{x}_1 & \cdots & \bar{x}_p & \bar{y} \\ \bar{x}_1 & \Sigma(x_1 - \bar{x}_1)^2 & & & \\ \vdots & & \ddots & & \\ \bar{x}_p & & & & \\ \bar{y} & & & \Sigma(y - \bar{y})^2 & \end{bmatrix}$$

Behoudens de eerste rij en kolom is dit de matrix waar Efroymson van uitgaat.

2. Dat Efroymson niet (3.3.5.) als Sweep operator gebruikt, maar de vorm met tekenwisselen in de rijen is reeds behandeld. Daarnaast voert Efroymson nog een uitbreiding van de matrix in, waarvan de betekenis bij de computerberekening niet duidelijk is.

Efroymson noteert namelijk de uitgangsmatrix als $\begin{pmatrix} S & T' \\ T & Z \end{pmatrix}$ waarin S zoals reeds onder 1^{o} werd opgemerkt uit de elementen $\Sigma_k(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$ bestaat, terwijl verder $Z = \sum_k (y_k - \bar{y})^2$ en T uit $\sum_k (x_{ik} - \bar{x}_i)(y_k - \bar{y})$ bestaat. Z is dus één enkel element. T één rij van elementen en T' één kolom van elementen.

De matrix wordt uitgebreid tot $\begin{pmatrix} S & T' & I \\ T & Z & D \\ -I & B & C \end{pmatrix}$ (4.4.1.)

waarin B, C en D aanvankelijk nul zijn. I is een identiteitsmatrix met evenveel elementen als S. Er wordt nu gesteld dat toepassing van het algorithme op de elementen van S opname van elementen in de regressie betekent en toepassing op elementen van C verwijdering van x_i's uit de regressie. Na iedere stap bevat B de regressiecoëfficienten en C de inverse elementen van de S matrix, dat wil zeggen de c_{ii} die nodig zijn om de standaarddeviatie van de regressiecoëfficienten te bepalen. Wat B betreft, is het inderdaad zo, dat deze de regressiecoëfficienten bevat, maar op de plaatsen die betrekking hebben op x_i's die verwijderd zijn door toepassing van de Sweep operator op C elementen komen geen nullen; daar blijft iets staan. Dit is als volgt in te zien. Als bovenstaande matrix gesplitst wordt in :

$$\begin{pmatrix} S_{11} & S_{12} & T_1' & I & 0 \\ S_{21} & S_{22} & T_2' & 0 & I \\ T_1 & T_2 & Z & 0 & 0 \\ -I & 0 & 0 & 0 & 0 \\ 0 & -I & 0 & 0 & 0 \end{pmatrix}$$

en hierop wordt SWP*(0, 1,...,s) toegepast, onstaat:

rijen

$$\begin{matrix} 1 \\ \vdots \\ s \end{matrix}\left\{\begin{matrix} \\ \\ \end{matrix}\right.\begin{pmatrix} S_{11}^{-1} & S_{11}^{-1}S_{12} & S_{11}^{-1}T_1' & S_{11}^{-1} & 0 \end{pmatrix}$$

$$\begin{matrix} s+1 \\ \vdots \\ p \end{matrix}\left\{\begin{matrix} \\ \\ \end{matrix}\right.\begin{pmatrix} -S_{21}S_{11}^{-1} & S_{22}-S_{21}S_{11}^{-1}S_{12} & T_2-S_{21}S_{11}^{-1}T_1' & -S_{22}S_{11}^{-1} & 0 \end{pmatrix}$$

$$p+1 \quad \begin{pmatrix} -T_1S_{11}^{-1} & T_2-T_1S_{11}^{-1}S_{12} & Z-T_1S_{11}^{-1}T_1' & -T_1S_{11}^{-1} & 0 \end{pmatrix}$$

$$\begin{matrix} p+2 \\ \vdots \\ p+1+s \end{matrix}\left\{\begin{matrix} \\ \\ \end{matrix}\right.\begin{pmatrix} S_{11}^{-1} & S_{11}^{-1}S_{12} & S_{11}^{-1}T_1' & S_{11}^{-1} & 0 \end{pmatrix}$$

$$\begin{matrix} p+2+s \\ \vdots \\ 2p+1 \end{matrix}\left\{\begin{matrix} \\ \\ \end{matrix}\right.\begin{pmatrix} 0 & -I & 0 & 0 & 0 \end{pmatrix}$$

De regressiecoëfficiënten zijn te vinden in de eerste groep elementen van rij p+1 namelijk $-T_1S_{11}^{-1}$; maar met negatief teken. Ze zijn ook te vinden zowel in de rijen 1,...,s als in p+r,.....p+1+s, namelijk als $S_{11}^{-1}T_1'$ inderdaad in de B submatrix van (4.4.1). Wordt de Sweepoperator nu op de elementen (p+1, p+2)...(p+1+s, p+1+s) toegepast, dan ontstaat:

$$\begin{pmatrix} 0 & 0 & 0 & I & 0 \\ 0 & S_{21} & T_2' & -S_{21} & I \\ 0 & T_2 & Z & -T_1 & 0 \\ -I & -S_{12} & -T_1' & S_{11} & 0 \\ 0 & -I & 0 & 0 & 0 \end{pmatrix}$$

B is dus niet weer 0 geworden maar $\begin{pmatrix} -T_1' \\ \\ 0 \end{pmatrix}$ terwijl C nu de oorspronkelijke S_{11} bevat.

Men vraagt zich echter wel af waarom Efroymson deze uitbreiding invoert, want in het computerprogramma wordt dit niet toegepast; de matrix blijft beperkt tot $\begin{pmatrix} S & T' \\ T & Z \end{pmatrix}$.

3. Tenslotte nog een laatste verandering ten opzichte van de algemene theorie. Er worden in het computerprogramma niet de oorspronkelijke gedefinieerde elementen $q_{ij} = \Sigma(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$ enz. ingevoerd maar de correlatiecoëfficiënten

$$r_{ij} = \frac{\Sigma(x_{ik} - \bar{x}_i)(x_{jt} - \bar{x}_j)}{\sqrt{\Sigma(x_{ik} - \bar{x}_i)^2 \cdot \Sigma(x_{jk} - \bar{x}_j)^2}} = \frac{q_{ij}}{S_i \cdot S_j}$$

hierin is ook y begrepen door y als x_{p+1} te schrijven en i van 1 t/m p+1 te laten lopen. De matrix $\begin{bmatrix} S & T' \\ T & Z \end{bmatrix}$ bestaat dus nu uit de elementenrij voor i, j = 1,...,p+1, en zal door $\begin{bmatrix} \tilde{S} & \tilde{T}' \\ \tilde{T} & \tilde{Z} \end{bmatrix}$ worden aangegeven. Toepassing van SWP*(0, 1,...,s) op beide heeft tot resultaat dat resp.:

$$\begin{bmatrix} S^{-1} & S^{-1}T' \\ -TS^{-1} & Z-TS^{-1}T' \end{bmatrix} \quad \text{en} \quad \begin{bmatrix} \tilde{S}^{-1} & \tilde{S}^{-1}\tilde{T}' \\ -\tilde{T}\tilde{S}^{-1} & \tilde{Z}-\tilde{T}\tilde{S}^{-1}\tilde{T}' \end{bmatrix}$$

ontstaat. De elementen van S zijn q_{ij} voor i,j = 1,...,p, die van $\tilde{S}$ zijn $\frac{q_{ij}}{s_i s_j}$; van T en $\tilde{T}$ resp. $q_{p+1,j}$ en $\frac{q_{p+1,j}}{s_{p+1} s_j}$ en van Z en $\tilde{Z}$ resp. $q_{p+1,p+1}$ en $\frac{q_{p+1,p+1}}{s_{p+1}^2}$. Als de elementen van S^{-1} met $\tilde{q}_{ij}$ worden aangegeven, dan is gemakkelijk in te zien dat die van $\tilde{S}^{-1}$ gelijk zijn aan $q^*_{ij} = s_i s_i \tilde{q}_{ij}$. Nu zijn de elementen van TS^{-1} de regressiecoëfficiënten voor het oorspronkelijke regressiemodel, dus $\hat{\beta}_i$; de elementen van $\tilde{T}\tilde{S}^{-1}$ die van het op standaarddeviaties genormeerde model, die met b_i zullen worden aangegeven. Dan is:

$\hat{\beta}_i = \sum_{k=1}^{p} q_{p+1,k}\tilde{q}_{k,j}$ en derhalve $b_i = \sum_{k=1}^{p} \frac{q_{p+1,k}}{s_{p+1}s_k} \cdot s_k s_j \tilde{q}_{kj} = \frac{s_j}{s_{p+1}} \cdot \hat{\beta}$

of $\hat{\beta}_i = \frac{s_{p+1}}{s_j} b_i$.

Dit is in het artikel van Efroymson, formule (15), genoteerd als $\hat{\beta}_i = \frac{\sigma_n}{\sigma_i} b_{in}$; s_{p+1} is namelijk de standaarddeviatie van y, wat door Efroymson met σ_n wordt aangegeven.

De standaarddeviatie van de volgens de regressie gereduceerde y (in feite van $\hat{\varepsilon}$) is volgens (4.3.3) te schrijven als $s_{\tilde{y}} = \sqrt{\frac{q^*_{p+1,p+1}}{b}}$ waarin $q^*_{p+1,p+1} = Z - TS^{-1}T'$; als $\tilde{Z} - \tilde{T}\tilde{S}^{-1}\tilde{T}' = \tilde{q}_{p+1,p+1}$ wordt genoemd, geldt $\tilde{q}^*_{p+1,p+1} = \frac{q^*_{p+1,p+1}}{s^2_{p+1}}$ derhalve $s_{\tilde{y}} = s_{p+1}\sqrt{\frac{\tilde{q}^*_{p+1,p+1}}{b}}$.

Dit is formule 14 van Efroymson: $s_y = \sigma_n\sqrt{r_{nn}/\phi}$.

Tenslotte is $s_{\hat{\beta}_i} = s_y\sqrt{\tilde{q}_{ii}}$ volgens (4.3.4) waarin $\tilde{q}_{ii}$ een element van S^{-1} is. Nu geldt voor de elementen q^*_{ii} en $\tilde{S}^{-1}$ dat $q^*_{ii} = S_i^2\tilde{q}_{ii}$ derhalve $S_{\hat{\beta}_i} = \frac{s_y}{s_i}\sqrt{q^*_{ii}}$. Dit is de formule (16) van Efroymson: $\frac{s_y}{\sigma_i}\sqrt{c_{ii}}$.

LITERATUUR

H. Scheffé : The Analysis of Variance
Wiley - New York 1959

A.P. Dempster : Elements of continious multivariable analysis
Addison - Wesly Publ. 1969

M.A. Efroymson: Multiple regression analysis
Nr. 17 in Ralston and Wilf (e.d.):
Mathematical Methods for Digital Computors N.Y. 1962 (1967)

A. Orden : Matrix inversion and related topics by direct methods
Nr. 2 in Ralston and Wilf (e.d.):
Mathematical Methods for Digital Computors N.Y. 1962 (1969)

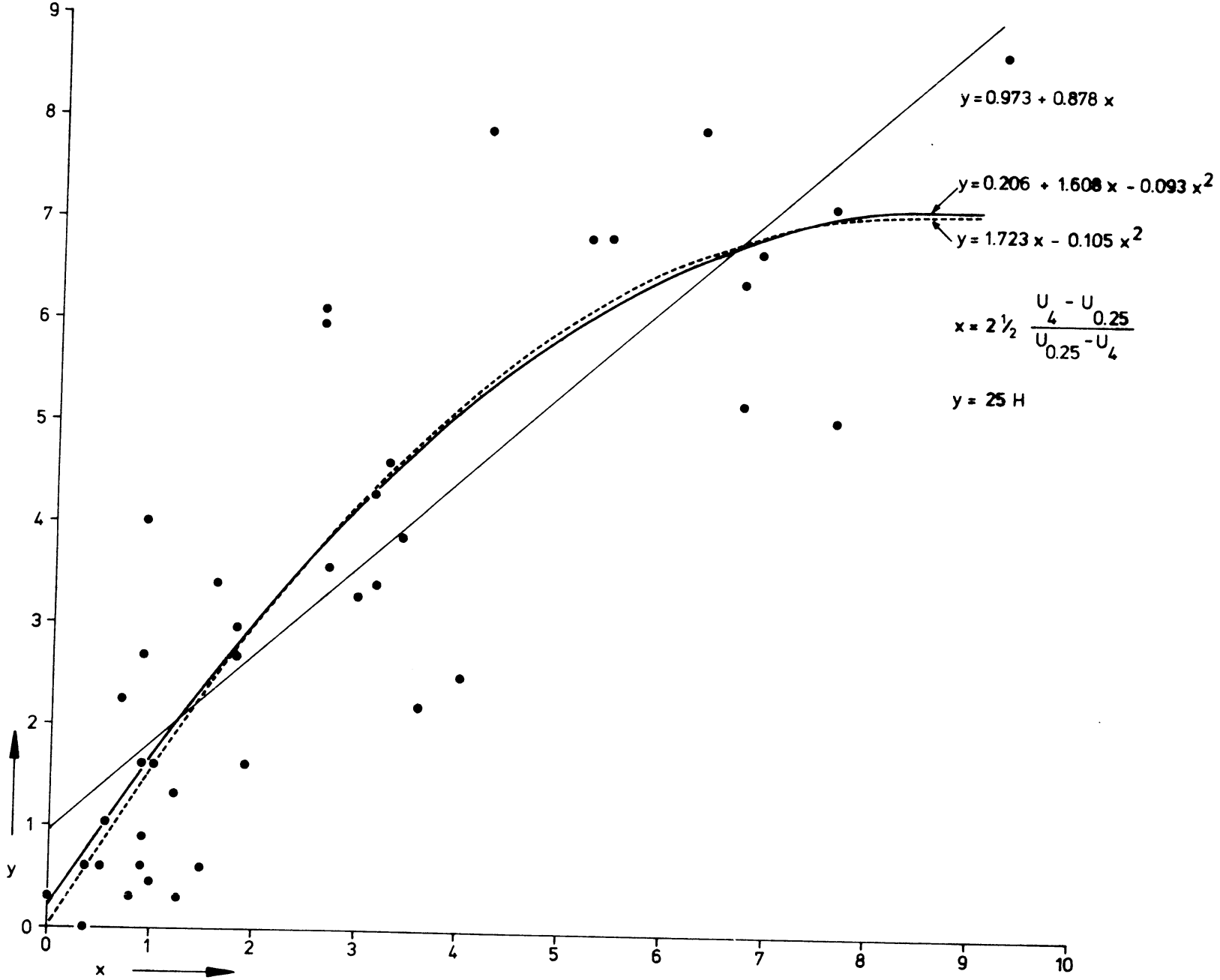


Fig. 1 Relatie tussen warmtestroom en stabiliteitsparameter