[INSTIT]UUT VOOR CULTUURTECHNIEK EN WATERHUISHOUDING
NOTA no. 131 d.d. 23 mei 1962

Lineaire regressie

L.P.Kamil

Inleiding

Bij het empirisch onderzoek naar de wijze waarop een bepaalde grootheid, bijvoorbeeld werkelijke verdamping, opbrengst, enzovoort, af zal hangen van een aantal andere variabele grootheden, wordt uitgegaan van reeksen waarnemingsuitkomsten. Men weet, dat er een zekere samenhang bestaat tussen de gemeten grootheden, dat wil zeggen één der grootheden is uit te drukken als een functie van de andere, en men stelt zich tot doel, aan de hand van een regressiemodel de onderlinge samenhang te bestuderen.

Een voorbeeld hiervan wordt gegeven op pagina 12 en volgende (Lysimeteronderzoek).

In het navolgende worden de te gebruiken formules en hun afleidingen bijeengebracht, terwijl tevens een overzichtelijk rekenschema wordt gegeven.

Symbolen en begrippen

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

Stochastische grootheden worden aangegeven door een onderstreepte letter ($\underline{x}$). Een vector wordt als zodanig in de tekst gedefinieerd.

De stochastische variabelen (vectoren) $\underline{x}$ en $\underline{y}$ zijn isomoor ($\underline{x} \approx \underline{y}$) als ze dezelfde kansfunctie hebben.
In formule:

$$\underline{x} \approx \underline{y} \quad \text{indien} \quad P(\underline{x} \leq c) = P(\underline{y} \leq c) \text{ voor iedere } c$$

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

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

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

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

<u>Formulering</u>

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

Er zijn dus n groepen van naar k geordende waarnemingsuitkomsten:

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

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

Schrijft men de verkregen waarnemingsuitkomsten overzichtelijk in kolommen, dan kan men de eigenschappen als kolomvectoren aanduiden. Men gaat nu uit van het volgende model:

$$\begin{pmatrix} \underline{y}_{10} \\ \underline{y}_{20} \\ \vdots \\ \underline{y}_{i0} \\ \vdots \\ \underline{y}_{n0} \end{pmatrix} \cong \alpha_1 \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{i1} \\ \vdots \\ y_{n1} \end{pmatrix} + \ldots + \alpha_j \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{nj} \end{pmatrix} + \ldots + \alpha_k \begin{pmatrix} y_{1k} \\ y_{2k} \\ \vdots \\ y_{ik} \\ \vdots \\ y_{nk} \end{pmatrix} + \begin{pmatrix} \underline{e}_1 \\ \underline{e}_2 \\ \vdots \\ \underline{e}_i \\ \vdots \\ \underline{e}_n \end{pmatrix} \qquad (1),$$

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

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

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

In vectornotatie wordt (1) en (2) :

$$\underline{y}_o \cong E\underline{y}_o + \sigma \underline{\chi}_n \quad ,$$

waarin $E\underline{y}_o = \sum_{j=1}^{k} \alpha_j \, y_j$

Hiermee is de stochastische vector $\underline{y}_o$ uitgedrukt als een som van een lineaire combinatie van de vectoren $y_1, \ldots, y_k$ en de stochastische vector $\sigma\underline{\chi}_n$.

De parameter α_j $(j = 1,2,\ldots\ldots,k)$ en σ moeten uit de gegevens worden geschat. Aangezien de gegevens stochastisch zijn, zullen ook de schatters stochastisch zijn.

Het probleem leidt tot het vinden van:

1. een schatter $\underline{a}$ van α , waarin met $\underline{a}$ en α respectievelijk worden bedoeld de kolomvectoren $(\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_k)$ en $(\alpha_1, \alpha_2, \ldots, \alpha_k)$.

2. de covariantie matrix van $\underline{a}$, waarmee betrouwbaarheids intervallen voor α bepaald kunnen worden.

3. een schatter van de rest variantie σ^2, welke nodig is voor het oplossen van punt 2

4. een schatter van de meervoudige correlatiecoëfficiënt ρ , welke als maat gezien kan worden van de aanpassing van het model.

Opgemerkt wordt, dat met een "schatter" in het algemeen een functie wordt bedoeld, waarmee een onbekende parameter berekend kan worden, zodanig dat een goede aanpassing aan de waarnemingsuitkomsten wordt verkregen. Een berekende uitkomst van zo'n functie heet een schatting.

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

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

Men beschouwt dus de gestandaardiseerde vectoren:

$$x_j = \frac{y_j - \bar{y}_j}{\sqrt{\Sigma\,(y_j-\bar{y}_j)^2}} \qquad \text{of in vector notatie : } x_j = \frac{y_j - \bar{y}_j}{\sqrt{(y_j-\bar{y}_j)^2}} \qquad (3)$$

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

$$E\,\underline{x}_o = \sum_{j=1}^{k-1} \beta_j \; x_j$$

De niveau vector is nu een o-vector (zie (3)).

In matrixnotatie is nu:

$$E\,\underline{x}_o = X\beta \quad , \qquad (4)$$

waarin X de matrix is van de gestandaardiseerde vectoren $(x_1, x_2, \ldots x_{k-1})$ en β de kolomvector $(\beta_1, \beta_2, \ldots, \beta_{k-1})$.

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

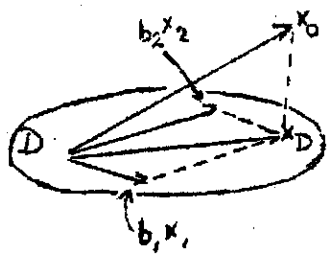
$$\alpha_j = \frac{\sqrt{(y_o - \bar{y}_o)^2}}{\sqrt{(y_j - \bar{y}_j)^2}} \; \beta_j \qquad , \; j = 1, 2, \ldots, k-1$$

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

Een schatter van β

De n waarnemingsuitkomsten van een eigenschap kunnen worden voorgesteld door een vector in de n-dimensionale ruimte. In deze ruimte liggen dus de k vectoren van de eigenschappen ($k < n$). Een schatter $\underline{b}$ van β volgt uit de orthogonale projectie van x_o op de lineaire deelruimte D met basisvectoren ($x_1, x_2, \ldots, x_{k-1}$) (zie tekening).

figuur 1



D is een hypervlak opgespannen door de basisvectoren ($x_1, x_2, \ldots x_{k-1}$). De voorwaarde tot loodrechte projectie is, dat de vector ($x_o - x_D$) loodrecht moet zijn op alle basisvectoren van D.
Bovendien is x_D een lineaire combinatie van deze basisvectoren.

Dus : $\underline{x}_D = X\,\underline{b}$ (5)

Tengevolge van de orthogonale projectie is :

$$(\underline{x}_o - X\,\underline{b}) \perp X$$

Dan is, indien ${}^tX$ de getransponeerde is van X, de voorwaarde voor de loodrechtheid:

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

Hieruit volgen de normaalvergelijkingen door toepassing van de distributieve wet:

$${}^tX\,X\,\underline{b} = {}^tX\,\underline{x}_o \quad ,$$

waaruit na voorvermenigvuldiging met de inverse van ${}^tXX$ volgt:

$$\underline{b} = ({}^tXX)^{-1}\;{}^tX\,\underline{x}_o \qquad (6)$$

Opgemerkt wordt dat de matrix van de normaalvergelijkingen (${}^tXX$) tengevolge van de standaardisatie tevens de correlatiematrix van de "verklarende" variabelen is.

Omgekeerd geldt voor de verwachtingswaarde van $\underline{b}$:

$$E\,\underline{b} = ({}^t XX)^{-1}\ {}^t X\, E\,\underline{x}_o \quad , \qquad (7)$$

en met $E\,\underline{x}_o = X\,\beta$ (zie (4)) :

$$E\,\underline{b} = ({}^t XX)^{-1}\ {}^t XX\,\beta = \beta\ ,$$

waaruit volgt dat $\underline{b}$ een zuivere schatter van β is.

De covariantie-matrix van b

Ter bepaling van de covariantie-matrix van $\underline{b}$ wordt eerst de covariantie-matrix van $\underline{x}_o$ berekend.
Uit de veronderstellingen dat:

$$\underline{e}_1 \cong \underline{e}_2 \cong \ldots\ldots \cong \underline{e}_n \cong \sigma\underline{\chi}$$

en : $\underline{e}_1, \underline{e}_2, \ldots\ldots, \underline{e}_n$ onderling onafhankelijk ,

volgt : $E\,(\underline{x}_{io} - E(\underline{x}_{io}))(\underline{x}_{jo} - E(\underline{x}_{jo})) =$

$$E\,\underline{e}_i\underline{e}_j = \begin{cases} \sigma^2 & \text{voor } i = j \\ 0 & \text{voor } i \neq j \end{cases} \quad ,$$

zodat : $\mathrm{Cov}(\underline{x}_o) = E\,(\underline{x}_o - E(\underline{x}_o))\ {}^t(\underline{x}_o - E(\underline{x}_o)) = \sigma^2 I$,

waarin I de n x n eenheidsmatrix is.

Verder is (zie (7)) $\beta = E(b) = ({}^t XX)^{-1}\ {}^t X\, E\,\underline{x}_o$

Nu is de covariantie-matrix van $\underline{b}$:

$$\begin{aligned}
\mathrm{Cov}(\underline{b}) &= E\,(\underline{b} - \beta)\ {}^t(\underline{b} - \beta) \\
&= E[({}^t XX)^{-1}\ {}^t X(\underline{x}_o - E\underline{x}_o)]\ {}^t[({}^t XX)^{-1}\ {}^t X(\underline{x}_o - E\underline{x}_o)] \\
&= E\,({}^t XX)^{-1}\ {}^t X(\underline{x}_o - E(\underline{x}_o))\ {}^t(\underline{x}_o - E\underline{x}_o)\ X\ ({}^t XX)^{-1} \\
&= ({}^t XX)^{-1}\ {}^t X\ [E(\underline{x}_o - E\underline{x}_o)\ {}^t(\underline{x}_o - E\underline{x}_o)]\ X\ ({}^t XX)^{-1} \\
&= ({}^t XX)^{-1}\ {}^t XX\ ({}^t XX)^{-1}\ \sigma^2 \\
&= ({}^t XX)^{-1}\ \sigma^2
\end{aligned}$$

De covariantie-matrix van $\underline{b}$ volgt dus uit de inverse van de matrix van de normaalvergelijkingen.

De varianties van $\underline{b}$ volgen uit de elementen van de hoofddiagonaal.

Een schatter van σ^2

Voor de bepaling van de covariantie-matrix van $\underline{b}$ is bekendheid met de waarde van de restvariantie σ^2 nodig.
Een schatting van σ^2 wordt gevonden uit de betrekking:

$$^t(\underline{x}_o - X\underline{b})(\underline{x}_o - X\underline{b}) \simeq \sigma^2 \ \underline{\chi}^2_{n-k} \tag{8}$$

De term in het linkerlid is het kwadraat van de lengte van de verschil-vector $x_o - x_D$ uit figuur 1, of dat deel van $\underline{x}_o$ dat niet door een lineaire combinatie van de basis van D verantwoord wordt.

Aangezien $(\underline{x}_o - X\underline{b}) \perp X$ geldt onder toepassing van de distributieve wet op (8) en bedenkende dat:

$$E \ \underline{\chi}^2_{n-k} = n - k \quad :$$

$$E \ (^t\underline{x}_o\underline{x}_o - {}^t\underline{b} \ {}^tX \ \underline{x}_o) = (n - k) \ \sigma^2$$

Hieruit volgt dat:

$$\frac{^t\underline{x}_o\underline{x}_o - {}^t\underline{b} \ {}^tX \ \underline{x}_o}{n - k} = \frac{1 - {}^t\underline{b} \ {}^tX \ \underline{x}_o}{n - k}$$

een zuivere schatter van σ^2 is.

De meervoudige correlatie coëfficiënt

De meervoudige correlatie coëfficiënt komt overeen met de cosinus van de hoek die x_o maakt met zijn component x_D in de deelruimte D. Aangezien de lengte van x_o de eenheidslengte is, is de cosinus van de hoek gelijk aan de lengte van x_D. Voor het kwadraat van de lengte geldt dan onder toepassing van (5) en (6) :

$$: \underline{r}^2 = {}^t(X\,\underline{b})\;X\,\underline{b} = {}^t\underline{b}\;{}^tXX\,\underline{b} = {}^t\underline{b}\;{}^tXX\;({}^tXX)^{-1}\;{}^tX\,\underline{x}_o = {}^t\underline{b}\;{}^tX\,\underline{x}_o$$

Hieruit volgt tevens dat een schatter van σ^2 is :

$$S(\sigma^2) = \frac{1 - \underline{r}^2}{n - k} \tag{9}$$

Overzicht van de te berekenen grootheden

De te berekenen grootheden, die de oplossing van het regressie probleem geven zijn nu samengevat:

$$b = ({}^tXX)^{-1}\;{}^tX\,x_o = A^{-1}c \qquad \text{als} \quad {}^tXX = A \quad \text{en} \quad {}^tX\,x_o = c$$

$$\mathrm{Cov}\;\underline{b} = ({}^tXX)^{-1}\;\sigma^2 = A^{-1}\sigma^2$$

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

$$r^2 = {}^tb\;{}^tX\,x_o = {}^tb\;c$$

Gewichten

Bij de veronderstellingen dat $x_{10}, x_{20}, \dots, x_{no}$ niet gelijke spreidingen bezitten doch respectievelijk de spreidingen $(\sigma_1, \sigma_2, \dots, \sigma_n)$ of $(\sigma/\gamma_1, \sigma/\gamma_2, \dots, \sigma/\gamma_n)$hebben kan men schrijven:

$$\underline{x}_o \cong E\,\underline{x}_o + \sigma\,\Gamma^{-1}\underline{\chi}_n \quad ,$$

waarin Γ^{-1} een diagonale matrix is met diagonaal elementen:

$$1/\gamma_1,\; 1/\gamma_2, \dots,\; 1/\gamma_n$$

Door voorvermenigvuldiging met Γ krijgt men:

$$\Gamma\,\underline{x}_o \cong \Gamma\,E\,\underline{x}_o + \sigma\,\underline{\chi}_n$$

De te berekenen grootheden worden slechts volledigheidshalve vermeld. De bewijzen verlopen volledig parallel aan de voorgaande.

Stelt men $\Gamma^2 = G$ dan volgt:

$$\underline{b} = ({}^tXGX)^{-1} \; {}^tXG \, \underline{x}_o$$

$$\text{Cov } \underline{b} = ({}^tXGX)^{-1} \, \sigma^2$$

$$S(\sigma^2) = \frac{{}^t\underline{x}_o G(\underline{x}_o - X\underline{b})}{n - k}$$

Nabeschouwing

De gebruikte veronderstellingen zijn:

Elke E $\underline{e}_i^2 = \sigma^2$, dat wil zeggen overal langs het regressie vlak moet dezelfde spreiding σ bestaan. Is dit niet het geval dan zijn er twee mogelijkheden, namelijk, de spreiding verloopt functioneel in welk geval transformatie moet worden toegepast door bijvoorbeeld overgang op de logarithme van de variabelen bij afwijkingen die procentueel verlopen, of de spreiding is van punt tot punt verschillend in welk geval door toepassing van gewichten naar gelijke σ moet worden getransformeerd.

Elke $\underline{e}_i \simeq \sigma \underline{\chi}$. Bij het terugbewijzen van de zuiverheid van de schatter van β is van deze veronderstelling geen gebruik gemaakt, zodat het normaal verdeeld zijn geen noodzakelijke voorwaarde is. Wel geldt bij deze veronderstelling, dat dan tevens de meest aannemelijke (maximum likelihood) schattingen worden verkregen.

De eigenschappen $x_1, \ldots\ldots, x_{k-1}$ behoeven niet stochastisch te zijn. Zij kunnen gecontroleerde variabelen zijn, zoals bijvoorbeeld, mestgift, waterstand. Onder deze categorie vallen ook afgeleide variabelen, zoals kwadraten, produkten, reciproken enzovoort.
In het geval van een model met afgeleide variabelen wordt het model ook lineair genoemd, namelijk in:

$$x_1,\ x_2,\ x_1^2,\ x_2^2,\ x_1,\ x_2,\ \ldots\ldots\ \text{enzovoort.}$$

Met het symbool $\underline{\chi}_n$ wordt de standaard-normale bolsymmetrische stochastische variabele bedoeld. In het bovengenoemde geval kan men aan het symbool een ruimere betekenis toekennen. Bij het toepassen van t en F-toets op de schattingen van de parameters is dan echter enige voorzichtigheid geboden.

Rekenschema

De berekening van de gevraagde grootheden geschiedt op overzichtelijke wijze met het schema van Choleski.

Men berekent eerst de matrix ${}^tXX = A$ en de kolom ${}^tX\,x_o = c$. In een schema worden na elkaar geschreven: (A, c, I). Het is mogelijk om A te schrijven als het produkt ${}^tTT$, als T een rechterbovendriehoeksmatrix is, die dus nullen heeft op plaatsen links onder de hoofddiagonaal.

Het berekenen van T kan men voorstellen door het matrix-produkt: $DA = T$, waarin D een linkeronderdriehoeksmatrix is.

Dan is:

$$D(A,c,I) = (DA,\ Dc,\ D) = (T\ ,\ Dc,\ D)$$

Nu is: ${}^tTT = {}^t(DA)(DA) = {}^tA\ {}^tDDA = A$

en : ${}^tA = A$

waaruit volgt dat: ${}^tDD = A^{-1}$

Evenzo geldt: ${}^t(Dc)D = {}^tc\ {}^tDD = {}^tc\ A^{-1} = {}^t(A^{-1}c) = {}^tb$

en : ${}^tbc = {}^tb\ {}^tX\ x_o = r^2$

Resteert nog te vermelden hoe de elementen t_{ij} van T gevonden worden uit de elementen a_{ij} van A.

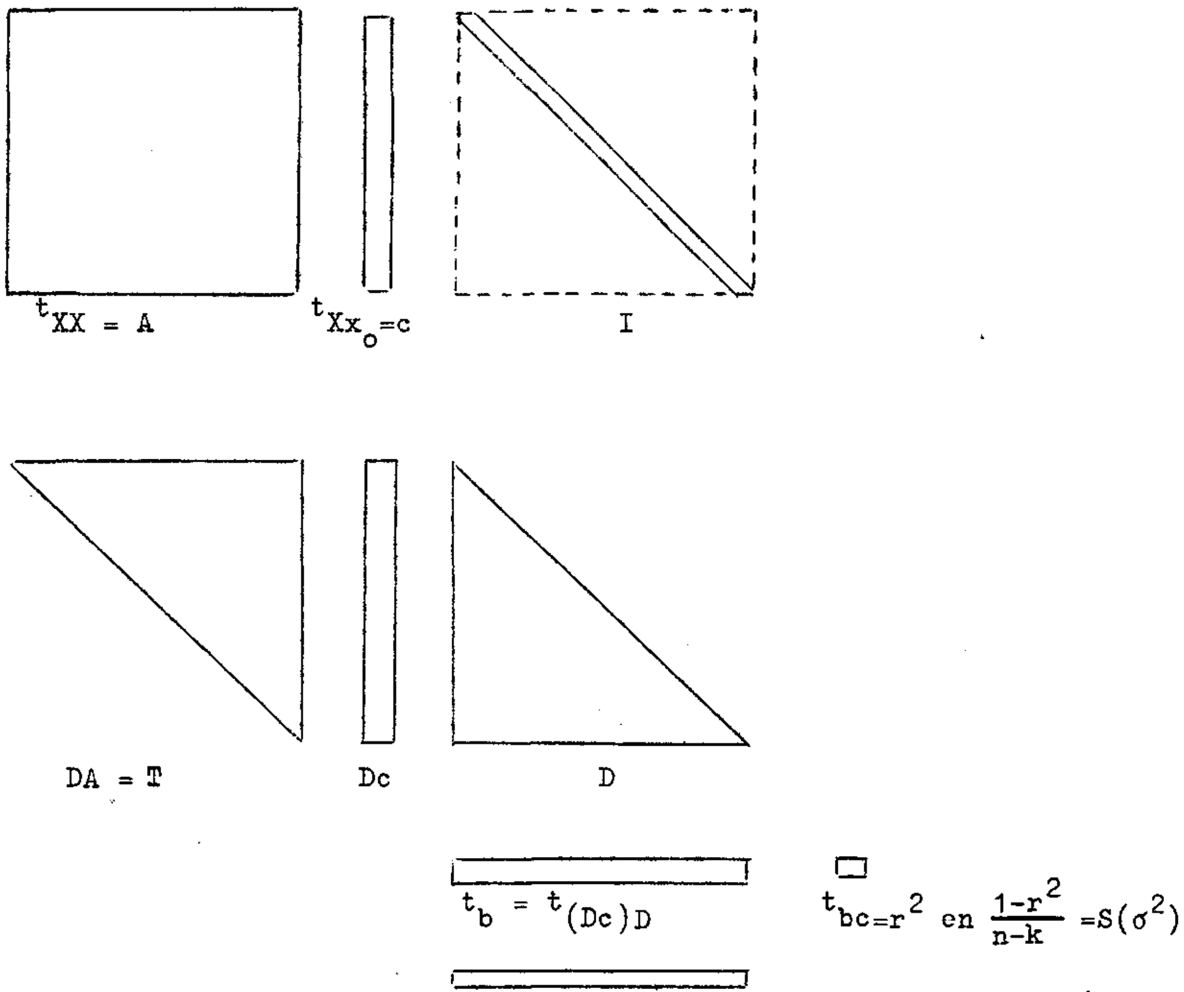
Het voorschrift luidt:

$$t_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} t_{ki}^2}$$

$$\text{en } t_{ij} = \left[a_{ij} - \sum_{k=1}^{i-1} t_{ki}t_{kj}\right] \Big/ t_{ii} \qquad \text{voor } i < j,$$

hetgeen wordt toegepast op de gehele matrix (A, c, I)

De eenvoud van de berekeningen blijkt uit het volgende schematische overzicht:



diagonaal elementen van de matrix A^{-1}
volgen uit het inprodukt van de kolommen
van de matrix D met zichzelf.

Toepassing op Lysimeteronderzoek

Het voorgaande geeft de grondslagen van de bewerkingen die zijn toegepast op lysimeter-gegevens met het doel de werkelijke verdamping (y_o), zo goed mogelijk te verklaren uit:

y_1 = open bak verdamping
y_2 = neerslag
y_3 = infiltratie - afvoer
y_4 = som van de vochtveranderingen in het profiel
y_5 = de graslengte

De n waarnemingsuitkomsten y_{io} ($i = 1, 2, \ldots\ldots, n$) worden beschouwd als uitkomsten van de stochastische variabele $\underline{y}_o$ met verwachtingswaarden:

$$E(y_{io}) = \phi\ (y_{i1}, \ldots\ldots, y_{i5}), \qquad (10)$$

waarvan de functie ϕ echter onbekend is.
Het is mogelijk een groot aantal functies binnen een gegeven interval voldoende te benaderen door een polynomium van voldoend hoge graad.

Een veronderstelling is, dat een polynomium van de tweede graad de functie ϕ voldoende benadert.

Hiertoe worden de variabelen aangevuld met kwadratische- en produkt termen, volgens het schema:

	y_1	y_2	y_3	y_4	y_5
y_1	y_6	y_{11}	y_{12}	y_{13}	y_{14}
y_2		y_7	y_{15}	y_{16}	y_{17}
y_3			y_8	y_{18}	y_{19}
y_4				y_9	y_{20}
y_5					y_{10}

zodat $y_6 = y_1^2$, $y_{11} = y_1 y_2$ enzovoort

De variabele $y_{21} = (1, 1, \ldots\ldots, 1)$ is de verantwoording van een constante.

De onder (10) genoemde verwachtingswaarden worden nu benaderd door:

$$E\,\underline{y}_o = \sum_{j=1}^{21} \alpha_j y_j$$

De meervoudige correlatie-coëfficiënt, die volgens (9) samenhangt met de restvariantie kan worden gebruikt als criterium voor goede benadering. Een hoge correlatie-coëfficiënt geeft bij deze wijze van werken nog niet aan dat het model fysisch voldoende interpreteerbaar is. Het is geboden het model steeds te toetsen aan veronderstellingen en inzichten.

Blijkt de correlatie-coëfficiënt laag te zijn, dan zijn daarvoor verschillende redenen mogelijk:

1. De benadering door een polynomium van de tweede graad is niet voldoende, of de benadering door het gekozen polynomium is niet mogelijk bijvoorbeeld wanneer $\frac{1}{x}$ inplaats van x opgenomen zou moeten worden.
2. In het model zijn niet alle verklarende factoren opgenomen. Opgemerkt wordt, dat door het opnemen van een nieuwe variabele de meervoudige correlatie-coëfficiënt nooit kleiner kan worden, zodat door toeval een schijnbare verbetering kan optreden. Dit is ook het geval bij het opnemen van nieuwe combinaties zoals bijvoorbeeld $y_1^2 y_2$, $y_1 y_2 y_3$, enzovoort.
3. De meetfouten in de gegevens zijn groot.

Tot slot wordt verwezen naar nota no. 103, Instituut voor Cultuurtechniek en Waterhuishouding, waarin de resultaten van bovengenoemd voorbeeld worden vermeld.