

CORRELATIEREKENING ALS HULPMIDDEL BIJ BEDRIJFSECONOMISCHE ANALYSES

door H. A. A. de Melverda

II

Somformules

Zoals uit artikel I reeds is gebleken, komen wij regelmatig sommeringen tegen. Teneinde de gedachten kort te kunnen uitdrukken geven wij deze sommeringen aan met het symbool Σ (sigma). Zo geven wij de jaarproductie van 1,8 mill. eenheden aan met $\Sigma X = 1800$ (eenheden in duizendtallen) en de jaarkosten evenzo met $\Sigma Y = 3144$. (lees: sigma Y is 3144)

Noemen wij nu het aantal bij elkaar behorende waarnemingen N (in ons voorbeeld is $N = 12$), dan kunnen wij schrijven:

$$\bar{X} = \frac{\Sigma X}{N} \text{ en } \bar{Y} = \frac{\Sigma Y}{N}. \text{ Voorts is gebleken: } \Sigma x = 0, \Sigma y = 0.$$

Aldus worden onze gedachten als het ware stenografisch weergegeven.

In verband met enige bewerkingen van algebraïsche aard leiden wij hieronder enige zeer eenvoudige bewerkingen met sommeringen af:

1. Stel a een constante factor en p een variabele grootheid, die achtereenvolgens de waarden $p_1, p_2, p_3 \dots p_n$ aanneemt, dan is

$$\Sigma ap = a \Sigma p \dots\dots\dots (II \ 1)$$

Immers men kan Σap schrijven als $(ap_1 + ap_2 + ap_3 + \dots ap_n) = a(p_1 + p_2 + p_3 + \dots p_n) = a \Sigma p$.

2. Stel q een andere variabele grootheid, die achtereenvolgens de waarden $q_1, q_2, \dots q_n$ aanneemt, dan is

$$\Sigma (p + q) = \Sigma p + \Sigma q \dots\dots\dots (II \ 2)$$

Immers men kan $\Sigma (p + q)$ schrijven als $(p_1 + q_1 + p_2 + q_2 \dots + p_n + q_n) = (p_1 + p_2 \dots p_n) + (q_1 + q_2 + \dots q_n) = \Sigma p + \Sigma q$.

3. $\Sigma c = Nc \dots\dots\dots (II \ 3)$

Immers men kan Σc schrijven als $c + c + c \dots$ aldus N -maal.

Het volgende voorbeeld geeft een toepassing van deze 3 bewerkingen:

$$\begin{aligned} \Sigma (a + p)(b + q) &= \Sigma (ab + aq + bp + pq) = \\ &= Nab + a \Sigma q + b \Sigma p + \Sigma pq. \end{aligned}$$

Strooïng, variatie en standaarddeviatie

Teneinde de mate van afwijking rondom het gemiddelde in een maat uit te drukken kunnen wij de som dezer afwijkingen bepalen, ongeacht het teken, en deze som delen door het aantal getallenparen (N): (zie staatje op volgende pagina bovenaan).

Hieruit zien wij, dat de gemiddelde afwijking om het gemiddelde bij y groter is dan bij x . Hieruit kunnen wij reeds concluderen, dat de regressielijn steiler moet zijn dan 45° t.o.v. de x -as. Immers dat zouden wij te zien krijgen, als wij de punten $+18,5$; $+21,83$; en $-18,5$; $-21,83$ in Figuur 2 zouden tekenen. Voorts zien wij, dat de waarden $18,5$ en $21,83$ niet ver uiteenliggen, hetgeen zou moeten corresponderen met de veronderstelling, dat de regressielijn met de x -as ten naaste bij een hoek van 45° maakt, hetgeen door figuur 2 aardig wordt bevestigd.

<u>x</u>		<u>y</u>	
pos.	neg.	pos.	neg.
	41		60
	29		20
	21		1
	12		22
	7	5	
	1		20
2			8
10		20	
13		8	
20		34	
28		18	
38		46	
<u>111</u>	<u>111</u>	<u>131</u>	<u>131</u>
222		262	

Strooiing: $222 : 12 = 18,5$; $262 : 12 = 21,83$

Voor de rekentechniek is het gebruik van de strooiing als maatstaf onhandig, want dan moet men van geval tot geval het teken van x en y bepalen. Daarom volgt men een andere methode en maakt gebruik van een eigenschap van quadraten, n.l. dat ze steeds positief zijn. Door nu alle waarden van x en y te quadrateren is men van de teken-kwestie ineens af:

<u>x^2</u>	<u>y^2</u>
1681	3600
841	400
441	1
144	484
49	25
1	400
4	64
100	400
169	64
400	1156
784	324
1444	2116
<u>6058</u>	<u>9034</u>

Om een nog nader duidelijk te maken reden, worden deze uitkomsten nu niet gedeeld door 12, maar door 11, waarna men vindt de *variatie* van x en y , welke wij aanduiden met resp. v_x en v_y .

Wij kunnen nu dus schrijven:

$$v_x = \frac{\sum x^2}{N - 1} = \frac{6058}{11} = 550,73, \text{ en}$$

$$v_y = \frac{\sum y^2}{N - 1} = \frac{9034}{11} = 821,27$$

Trekken wij nu uit deze uitkomsten de wortel (om de quadratering op te heffen), dan krijgen wij de *standaarddeviatie* van x en y , welke wij

aanduiden met resp. s_x en s_y ³⁾. Wij vinden dan dus $s_x = 23,47$ en $s_y = 28,66$.

Wij zien hieruit, dat de standaarddeviatie groter is dan boven gevonden waarden voor de strooiing. Uiteraard is ook hier de standaarddeviatie van y groter dan die van x .

In het algemeen geldt voor de berekening van de standaarddeviatie van een variabele p de formule:

$$s_p = \sqrt{v_p} = \sqrt{\frac{\sum p^2}{N-1}} \dots\dots\dots (II\ 4)$$

Men moet even aan het begrip standaarddeviatie wennen en kan dit het best doen door daaraan de voorstelling te verbinden van gemiddelde afwijking van het gemiddelde.

Vrijheidsgraden

Thans komen wij tot de verklaring, waarom wij de som der quadraten der afwijkingen niet door N , maar door $N-1$ deelden. Men drukt dit uit door te spreken van $(N-1)$ vrijheidsgraden (degrees of freedom).

De bedoeling hiervan kan o.a. duidelijk gemaakt worden aan de hand van de omstandigheid, dat $\sum x = 0$ is. In ons voorbeeld wordt de uitkomst 0 verkregen door een samentelling van 12 deels positieve, deels negatieve getallen. Neemt men nu 11 willekeurige positieve, deels negatieve getallen, dan is de som daarvan in het algemeen ongelijk aan nul. De waarde van deze uitkomst kan ons tot op zekere hoogte onverschillig laten. Immers als wij de waarde van het 12e getal willekeurig kunnen bepalen, dan kunnen wij — ook al zou men de 11 getallen nog zo willekeurig veranderen — steeds de waarde van het 12e getal dusdanig bepalen, dat de som der 12 getallen gelijk aan nul wordt. M.a.w. men kan niet 12 getallen willekeurig kiezen, maar wel 11 getallen; men spreekt dan van 11 vrijheidsgraden.

Dat wij bij deling door N van de som der quadraten door een te groot getal hebben gedeeld moge duidelijk worden uit de volgende overweging. Wij hebben $\bar{X}$ bepaald uit een reeks van N waarnemingen. Dit is een beperkt aantal en eigenlijk zou de uitkomst $\bar{X}$ veel nauwkeuriger zijn, als wij hadden kunnen beschikken over een zeer groot aantal waarnemingen. Wij zouden dan een waarde hebben gevonden, die zowel groter als kleiner zou kunnen zijn dan $\bar{X}$, hetgeen wij kunnen uitdrukken door de juiste waarde $\bar{X}'$ gelijk te stellen aan $\bar{X} \pm$ een klein bedrag ε , dus $\bar{X}' = \bar{X} \pm \varepsilon$. Voor de afwijkingen rondom het gemiddelde zouden wij dan gevonden hebben, b.v. $x_1' = X_1 - (\bar{X} \pm \varepsilon) = x_1 \mp \varepsilon$; $x_2 = x_2 \mp \varepsilon$, enz.

$$\begin{aligned} \text{Quadratering levert dan op:} \quad (x_1')^2 &= x_1^2 \mp 2\varepsilon x_1 + \varepsilon^2 \\ (x_2')^2 &= x_2^2 \mp 2\varepsilon x_2 + \varepsilon^2 \\ &\text{enz.} \end{aligned}$$

$$\text{Totaal: } \sum (x')^2 = \sum x^2 \mp 2\varepsilon \sum x + N\varepsilon^2$$

Nu weten wij, dat $\sum x = 0$, zodat bovenstaande uitkomst neerkomt op:

$$\sum (x')^2 = \sum x^2 + N\varepsilon^2$$

³⁾ Meestal geeft men de standaarddeviatie aan met het kleine sigmasymbool σ_x en σ_y . Wij wijken van deze notatie af, omdat onze notatie o.i. gemakkelijker tot het voorstellingsvermogen spreekt.

Overwegen wij nu, dat wij eigenlijk $\sum (x')^2$ hadden moeten bepalen, in plaats van $\sum x^2$ en dat $N \varepsilon^2$ steeds een positief getal moet zijn, dan is duidelijk, dat de waarde voor de variatie $v_x = \frac{\sum (x')^2}{N} = \frac{\sum x^2}{N} + \varepsilon^2 > \frac{\sum x^2}{N}$ is, hetgeen wij kunnen corrigeren door $\sum x^2$ in plaats van door N door een kleiner getal te delen.

Men kan nu bewijzen, dat men dan door $N - 1$ moet delen.⁴⁾

Dat het getal, waardoor men moet delen 1 minder is dan N , maakte Mather plausibel door het uitzonderlijke geval te nemen, dat $N = 1$ is, d.w.z. dat men slechts beschikt over één enkele waarneming van X . In dat geval is $\bar{X} = X_1$ en dus $x_1 = 0$. Voor de standaarddeviatie zou men dan vinden $\frac{0}{1} = 0$. Deze uitkomst kan echter onmogelijk goed zijn, want uit één enkele waarneming kan men onmogelijk een conclusie trekken omtrent de waarde van de standaarddeviatie, zodat de uitkomst had moeten luiden $\frac{0}{0} =$ onbepaald, d.w.z. men moet hier delen door 1 minder dan het aantal waarnemingen, dus algemeen door $N - 1$.

Gestandaardiseerde afwijkingen van het gemiddelde.

Het levert bepaalde voordelen op, om de afwijkingen van het gemiddelde (x en y) te delen door hun standaarddeviatie (s_x en s_y). Hierdoor komt men vooreerst af van het afhankelijk zijn van de grafiek 2 van de keuze der eenheden. Zou men de eenheden van de productie-hoeveelheid b.v. in honderdtallen hebben gemeten, dan zouden de waarden van x alle 10 maal zo groot zijn en zou de regressielijn in de figuur heel vlak lopen, waardoor een bedriegelijke schijn zou worden gewekt. Aangezien nu echter de standaarddeviatie van x ook 10 maal zo groot wordt, levert deling van elke waarde van x door de standaarddeviatie van x uitkomsten op, die onafhankelijk zijn van de keuze der eenheden.

De aldus berekende getallen worden aangeduid met x_s en y_s en men spreekt dan van *gestandaardiseerde variabelen*.

Men heeft dus $x_s = \frac{x}{s_y}$ en $y_s = \frac{y}{s_x}$ (II 5)

Wanneer men de aldus gestandaardiseerde variabelen in grafiek brengt, dan zal de helling van de regressielijn niet meer beïnvloed worden door de keuze der eenheden. Een ander voordeel van deze procedure zal hieronder nog blijken.

Voor ons getallenvoorbeeld vinden wij dan het staatje, dat op de volgende pagina bovenaan is weergegeven.

Wij zien hieruit, dat $\sum x_s^2$ en $\sum y_s^2$ ten naaste bij gelijk zijn aan 11, d.w.z. aan het aantal vrijheidsgraden. (Dat deze uitkomsten niet precies op 11 uitkomen, wordt veroorzaakt door een afronding in de standaarddeviatie.) Het bewijs hiervoor is zeer eenvoudig:

$$\sum x_s^2 = \sum \left\{ \frac{x}{s_x} \right\}^2 = \sum \frac{x^2}{s_x^2} = \frac{\sum x^2}{s_x^2} = \frac{\sum x^2}{\frac{\sum x^2}{N-1}} = N-1$$

⁴⁾ Dit bewijs nemen wij hier niet op. Het komt neer op het bewijs, dat $N\varepsilon^2 = s_x'^2$ moet zijn. Men kan dan schrijven $\sum (x')^2 = \sum x^2 + \frac{\sum (x')^2}{N}$, waaruit volgt: $s_x'^2 = \frac{\sum (x')^2}{N} = \frac{\sum x^2}{N-1}$.

	x_s	y_s	x_s^2	y_s^2
	— 1,747	— 2,093	3,0520	4,3806
	— 1,235	— 0,698	1,5252	0,4872
	— 0,895	— 0,035	0,8010	0,0012
	— 0,511	— 0,767	0,2611	0,5883
	— 0,298	+ 0,174	0,0888	0,0303
	— 0,043	— 0,698	0,0018	0,4872
	+ 0,085	— 0,279	0,0072	0,0778
	+ 0,426	+ 0,698	0,1815	0,4872
	+ 0,554	+ 0,279	0,3069	0,0778
	+ 0,852	+ 1,186	0,7259	1,4066
	+ 1,193	+ 0,628	1,4232	0,3944
	+ 1,619	+ 1,605	2,6212	2,5760
Totaal	0	0	10,9958	10,9946

Het gevolg van deze eigenschap is, dat de variatie (Σ_i^2 gedeeld door $N - 1$) steeds gelijk aan 1 moet zijn en dat derhalve ook de standaarddeviatie gelijk aan 1 moet zijn:

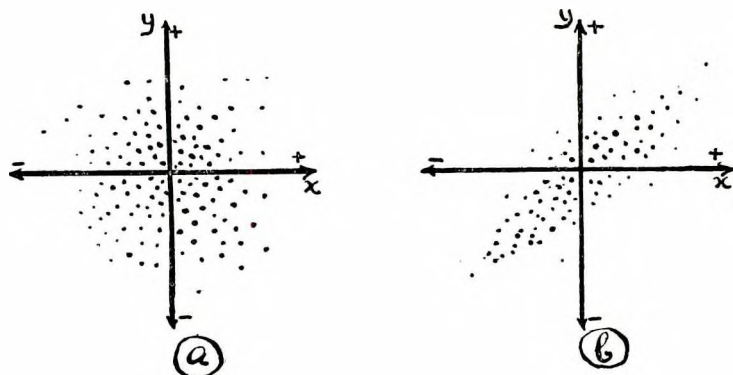
$$s_{x_s} = \sqrt{v_{x_s}} = 1 \dots\dots\dots (II\ 6)$$

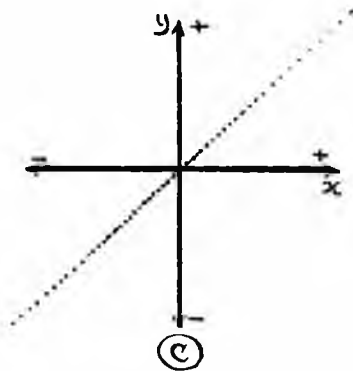
of in woorden: de standaarddeviatie van een gestandaardiseerde variabele is steeds gelijk aan 1.

Van deze eigenschap zullen wij in de hiernavolgende paragraaf gebruik moeten maken. Ter geruststelling delen wij echter mede, dat men in de practijk nooit berekeningen van de gestandaardiseerde variabele behoeft te maken. Bovenstaande afleiding heeft daarom uitsluitend waarde voor de bewijsvoering in de volgende paragraaf.

Covariatie en correlatiecoëfficiënt

Eén van de doeleinden van de correlatierekening is een maatstaf te vinden voor de mate van samenhang. Eenanschouwelijk beeld van deze samenhang kan men verkrijgen, als men de variabelen x en y grafisch voorstelt. De mate van correlatie kan dan onderscheiden worden in vele stadia, waarvan de volgende figuur de drie meest karakteriserende aangeeft:





Figuur 3.

In figuur 3a is er geen verband tussen de variabelen aan te wijzen d.w.z. correlatie is afwezig (men kan ook zeggen, dat de storende invloeden het verband volledig overheersen). In figuur 3b is er een duidelijk verband aan te wijzen, terwijl in figuur 3c de correlatie volmaakt is.

Bezien wij nu eerst de volmaakte correlatie wat nader, dan blijkt deze erop neer te komen, dat bij elke positieve waarde van x ook een positieve waarde van y behoort, zodat wij constateren, dat voor elk punt geldt, dat het product van x en y positief moet zijn.

Bezien wij nu figuur 3b, dan blijkt, dat er hier ook vele punten zijn, waarbij x en y gelijk teken hebben, maar dat er daarnaast andere punten zijn, waarvoor x en y tegengesteld teken hebben, m.a.w. voor vele punten is het product van x en y positief en voor sommige punten negatief. Wij kunnen nu direct aanvoelen, dat $\sum xy$ een maat moet zijn voor de correlatie en dat deze som van producten bij volmaakte correlatie hoger moet zijn (bij een gelijk aantal punten) dan bij onvolmaakte correlatie.

Een bevestiging hiervoor vinden wij in figuur 3a, waarbij het net zo waarschijnlijk is, dat xy positief of negatief is, zodat de waarde van $\sum xy$ de waarde van nul benaderen zal. Daarom kan men ook zeggen, dat correlatie afwezig zal zijn, als $\sum xy = 0$ is.

Voorts kan men aan de waarde van $\sum xy$ zien hoe de punten moeten verlopen. Immers is $\sum xy$ positief, dan moet de puntenvolgorde overwegend van linksonder naar rechtsboven gaan; men spreekt dan van *positieve correlatie*. Is echter $\sum xy$ negatief, dan moet de puntenvolgorde overwegend van linksboven naar rechtsonder gaan; men spreekt dan van *negatieve correlatie*.

Toch is $\sum xy$ nog geen volledig ideale maatstaf voor de correlatie, want de waarde ervan neemt toe als het aantal waarnemingen (resp. aantal vrijheidsgraden) toeneemt. Daarom is de maatstaf al aanmerkelijk beter, als men $\sum xy$ deelt door het aantal vrijheidsgraden. De uitkomst geeft men dan de benaming *covariatie* (ter onderscheiding van de variatie, die betrekking heeft op de producten van x en x , resp. van y en y), welke wij zullen aanduiden met c_{xy} . Wij hebben derhalve

$$c_{xy} = \frac{\sum xy}{N - 1} \dots\dots\dots (II\ 7)$$

De waarde van de covariatie van ons getallenvoorbeeld wordt dus berekend als volgt:

x	y	xy
— 41	— 60	+ 2460
— 29	— 20	+ 580
— 21	— 1	+ 21
— 12	— 22	+ 264
— 7	+ 5	— 35
— 1	— 20	+ 20
+ 2	— 8	— 16
+ 10	+ 20	+ 200
+ 13	+ 8	+ 104
+ 20	+ 34	+ 680
+ 28	+ 18	+ 504
+ 38	+ 46	+ 1748
0	0	+ 6530

De covariatie is nu dus $\frac{6530}{11} = 593,6$. Uit dit getal kan men echter onmogelijk een conclusie trekken omtrent de mate van correlatie. Voor eerst is het nog afhankelijk van de keuze der eenheden en vervolgens zou men het getal willen vergelijken met het getal, dat men voor volmaakte correlatie zou moeten krijgen.

Daarom volgt men de methode, om — evenals werd gedaan bij de variatie — de covariatie te bepalen voor de gestandaardiseerde variabelen (x_s en y_s):

$$c_{x_s y_s} = \frac{\sum x_s y_s}{N - 1}$$

Van de waarde voor $\sum x_s y_s$ kan nu in het geval van volmaakte correlatie iets naders gezegd worden. In dit geval liggen de punten op een rechte lijn door de oorsprong, zodat voor elk willekeurig punt zal gelden: $y_s = kx_s$, waarin k een bepaalde constante is.

Derhalve moet ook gelden:

$y_s^2 = k^2 x_s^2$ en zal voor de som der aldus berekende waarden moeten gelden

$$\sum y_s^2 = k^2 \sum x_s^2$$

Wij hebben echter gezien, dat zowel $\sum y_s^2$ als $\sum x_s^2$ gelijk moeten zijn aan het aantal vrijheidsgraden, zodat $k^2 = 1$ en $k = \pm 1$ moet zijn. Hieruit volgt dat $y_s = \pm x_s$ moet zijn, en wel $y_s = x_s$ bij positieve volmaakte correlatie en $y_s = -x_s$ bij negatieve volmaakte correlatie. m.a.w. de regressielijn deelt bij volmaakte correlatie in het diagram der gestandaardiseerde variabelen de rechte hoek tussen de assen middendoor. Uit de gevonden gelijkheid volgt nu $\sum x_s y_s = \pm \sum x_s^2$. Nu is $\sum x_s^2$ gelijk aan het aantal vrijheidsgraden ($N - 1$), zodat de conclusie is, dat $c_{x_s y_s} = \pm 1$

Letten wij gemakshalve alleen op de positieve correlatie, dan blijkt, dat voor alle statistische verbanden van enkelvoudige lineaire aard de waarde van $c_{x_s y_s}$ moet liggen tussen 0 en 1; n.l. bij 0 in het geval van afwezigheid van correlatie, bij 1 in het geval van volmaakte correlatie en daartussen in alle andere mogelijke gevallen. De covariatie van de gestandaardiseerde variabelen blijkt dus een ideale maatstaf te zijn voor de beoordeling van

de mate der correlatie. Daarom heeft men deze maatstaf de kortere naam van *correlatie-coëfficiënt* gegeven, aangeduid door: r_{xy} of kortweg: r ; derhalve

$$r = \frac{\sum x_s y_s}{N-1} \dots\dots\dots (II\ 8)$$

Het werken met gestandaardiseerde variabelen is — zoals wij gezien hebben — nogal omslachtig; daarom moet bovenstaande formule herleid worden tot de oorspronkelijke variabelen:

$$\begin{aligned} r &= \frac{\sum \frac{x}{s_x} \cdot \frac{y}{s_y}}{N-1} = \frac{\sum xy}{s_x s_y (N-1)} = \\ &= \frac{\sum xy}{\sqrt{\frac{\sum x^2}{N-1}} \cdot \sqrt{\frac{\sum y^2}{N-1}} \cdot (N-1)} \\ r &= \frac{\sum xy}{\sqrt{\sum x^2} \cdot \sqrt{\sum y^2}} \dots\dots\dots (II\ 9) \end{aligned}$$

Voor ons practisch voorbeeld vinden wij dan:

$$r^2 = \frac{6530^2}{6058 \times 9034} = 0.779 \text{ en dus } r = +0.883.$$

Uit deze uitkomst is te zien, dat de correlatie vrij bevredigend is te noemen.

Op overeenkomstige wijze kan men nu voor alle andere enkelvoudige lineaire correlaties de mate van samenhang door de berekening van de correlatiecoëfficiënt bepalen.

(Wordt vervolgd)