CORRELATIEREKENING ALS HULPMIDDEL BIJ BEDRIJFSECONOMISCHE ANALYSES [1])

door H. A. A. de Melverda

III

Berekening van de enkelvoudige lineaire regressielijn.

De berekening van de gezochte regressielijn is nu na het voorgaande een betrekkelijk eenvoudige aangelegenheid. Onderstaande figuur zal de methode, volgens welke wij te werk zullen gaan, verduidelijken:

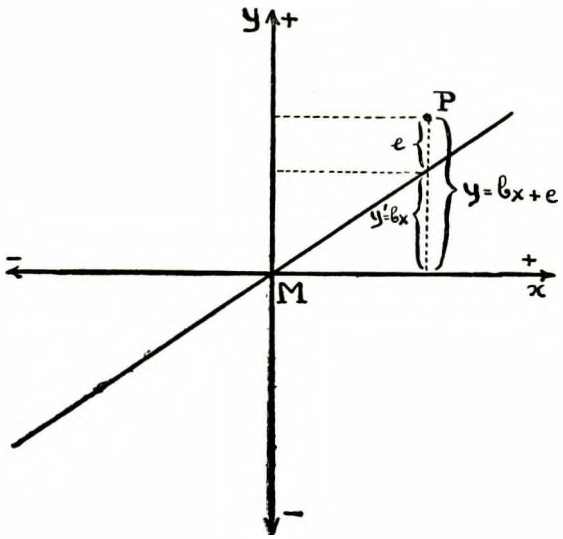


Fig. 4.

Uiteraard zal een willekeurig punt P (behorende bij de waarnemingen x en y) in de regel niet op de gezochte regressielijn liggen. Laten wij een loodlijn uit P neer op de x-as, dan zal de waarde y derhalve groter of kleiner zijn dan de waarde $y' = bx$ (van het punt op de regressielijn bij dezelfde waarde van x). Het verschil $y - y'$ zullen wij aangeven met e (error), zodat wij kunnen schrijven $y = bx + e$, of $e = y - bx$. De te bepalen grootheid b geeft de helling van de regressielijn aan en wordt daarom *regressiecoëfficiënt* genoemd.

Wij zullen nu een veronderstelling moeten maken omtrent de oorzaak van de omstandigheid, dat er een verschil is tussen waargenomen waarde en berekende waarde van y, d.i. de oorzaak waarom er geen volmaakte correlatie is tussen x en y. Wij zullen aannemen, dat de productiehoeveelheid zonder fout is vastgesteld, m.a.w. de oorzaak van afwezigheid van volmaakte correlatie ligt niet bij x maar bij de waarneming van y, d.w.z. de kostprijs wordt beïnvloed door storende invloeden hetgeen wij gemakkelijk kunnen begrijpen als wij denken aan schommelingen in prijspeil, toevallige duurdere of goedkopere inkopen of aan de omstandigheid, dat de kosten behalve van de productiehoeveelheid ook nog afhankelijk kunnen zijn van b.v. de gemiddelde grootte van de bestelling.

[1]) Zie ook de door de Redactie geschreven inleiding bij het 1e gedeelte van deze artikelenreeks in het Mei-nummer 1950.

Hoe het ook zij, het verschil e wordt veroorzaakt door de waarneming y, maar niet door de waarneming x, zodat e en x ten opzichte van elkaar volkomen onafhankelijk moeten zijn of m.a.w. tussen e en x bestaat geen correlatie. Wij hebben in het tweede artikel geleerd, dat wij deze laatste omstandigheid mathematisch kunnen weergeven door

$$\Sigma xe = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{(III 1)}$$

Deze uitdrukking wordt *normaalvergelijking* genoemd en is uitgangspunt van de berekening van de regressielijn. Immers, daar $e = y - bx$ is, kunnen wij (III 1) schrijven als volgt:

$\Sigma xe = \Sigma x(y - bx) = \Sigma (xy - bx^2) = \Sigma xy - b\Sigma x^2 = 0,$

waaruit volgt $b = \dfrac{\Sigma xy}{\Sigma x^2}$ ..... (III 2)

Door deze uitkomst is de gezochte regressielijn volledig bepaald. Voor ons getallenvoorbeeld vinden wij:

$$b = \frac{6.530}{6.058} = 1{,}078$$

Aangezien $\overline{Y} = a + b\overline{X}$ is, kan a berekend worden uit:

$$a = \overline{Y} - b\overline{X} \quad \ldots\ldots\ldots\ldots \text{(III 3)}$$

Voor ons getallenvoorbeeld vinden wij dan $a = 262 - 1{,}078 \times 150 = 100{,}3$.

De gezochte vergelijking moet dus luiden: $Y' = 100{,}3 + 1{,}078\ X$ (zowel Y' als X in duizendtallen). Hieronder volgen de aldus berekende waarden:

Y	Y'	e	e^2
202	217,8	— 15,8	249,64
242	230,7	+ 11,3	127,69
261	239,4	+ 21,6	466,56
240	249,1	— 9,1	82,81
267	254,5	+ 12,5	156,25
242	260,9	— 18,9	357,21
254	264,2	— 10,2	104,04
282	272,8	+ 9,2	84,64
270	276,0	— 6,0	36,00
296	283,5	+ 12,5	156,25
280	292,2	— 12,2	148,84
308	302,9	+ 5,1	26,01
3144	3144,0	0	1995,94

De eerste merkwaardigheid, die wij opmerken, is wel plausibel, n.l. $\Sigma Y = \Sigma Y'$. Het bewijs hiervoor is als volgt:

$$\begin{aligned}\Sigma Y' &= \Sigma (a + b\overline{X}) = Na + b\Sigma X = \\ &= N(\overline{Y} - b\overline{X}) + b\Sigma X = \\ &= N\overline{Y} - b(N\overline{X} - \Sigma X) = \\ &= N\overline{Y}, \text{ waaruit volgt:}\end{aligned}$$

$$\Sigma Y' = \Sigma Y \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{(III 4)}$$

Daar $y = Y - \overline{Y}$ en $y' = Y' - \overline{Y}$ is, is $Y - Y' = y - y' = e$. Uit (III 4) volgt nu rechtstreeks, dat $\Sigma e = 0$ moet zijn. De betekenis van e^2 komt in een latere paragraaf ter sprake.

Berekening met behulp der oorspronkelijke getallen (crude data method).

De berekening van de regressielijn op de wijze, als in het voorgaande werd gedaan kan soms wat te omslachtig zijn. Dit geldt vooral als de gemiddelde waarden $\bar{X}$ en $\bar{Y}$ geen „mooie" getallen zijn, waardoor in de afwijkingen van de gemiddelden decimalen voorkomen. Tenslotte is elke extra-bewerking een extra-mogelijkheid om rekenfouten te maken, zodat het in vele gevallen veel korter moet zijn te werken met de oorspronkelijke getallen dan met de afwijkingen der gemiddelden dier getallen.

Daarbij komt, dat het voor de nauwkeurigheid der berekeningen voldoende is, als de oorspronkelijke getallen bestaan uit *3* à *4* cijfers d.w.z. alle cijfers, die daarachter komen, kan men gevoeglijk weglaten. Het werken met de afwijkingen van de gemiddelden is daarom alleen raadzaam, als daardoor gewerkt kan worden met getallen, die kleiner zijn dan *3* à *4* cijfers.

De hierboven gevolgde methode was echter nodig, om de theorie duidelijk te maken. Het is daarom duidelijk, dat de crude data method neerkomt op een herleiding tot de afwijkingen van de gemiddelden en wel als volgt:

$$\Sigma xy = \Sigma (X - \bar{X})(Y - \bar{Y}) = \Sigma (XY - X\bar{Y} - \bar{X}Y + \bar{X}\bar{Y}) =$$

$$= \Sigma XY - \bar{Y}\Sigma X - \bar{X}\Sigma Y + N\bar{X}\bar{Y} = \Sigma XY - \frac{\Sigma Y}{N}.\Sigma X -$$

$$- \frac{\Sigma X}{N}.\Sigma Y + N\frac{\Sigma Y}{N}.\frac{\Sigma X}{N}.$$

$$\Sigma xy = \Sigma XY - \frac{\Sigma X \Sigma Y}{N} \quad \text{(III 5a)}$$

Op overeenkomstige wijze — of korter: door vervanging van y door x resp. x door y — vindt men ook:

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \quad \text{(III 5b)}$$

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} \quad \text{(III 5c)}$$

Deze formules zijn gemakkelijk te onthouden door hun symmetrische structuur.

Voorts kan het gewenst zijn op de berekeningen een *contrôle* toe te passen, want zowel bij de optellingen als bij de quadrateringen en vermenigvuldigingen kan men fouten maken. De contrôle op de berekeningen verkrijgt men door $X + Y = Q$ te stellen en vervolgens $XQ = X^2 + XY$ en $YQ = XY + Y^2$ te berekenen. Voor de sommeringen vindt men dan:

$$\Sigma XQ = \Sigma X^2 + \Sigma XY \quad \text{(III 6a)}$$

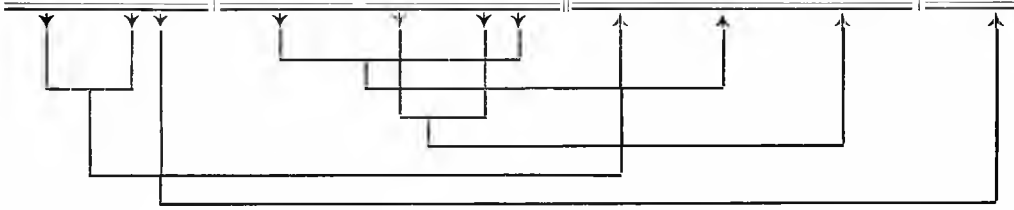
$$\Sigma YQ = \Sigma XY + \Sigma Y^2 \quad \text{(III 6b)}$$

Klopt nu ΣXQ maar ΣYQ niet, dan moet de fout schuilen bij ΣY^2, want ΣXY moet dan goed zijn. Klopt ΣYQ wel maar ΣXQ niet, dan schuilt de fout zeker bij ΣX^2. Kloppen geen van beiden, dan is het zeer waarschijnlijk, dat de fout schuilt bij ΣXQ, want deze term is in beide vergelijkingen gemeenschappelijk, hoewel uiteraard de fouten ook nog kunnen schuilen in de overige termen.

De berekening voor het voorbeeld, dat wij tot dusver volgden, verloopt

nu als volgt (de pijlen onder de sommeringen geven de controles op de juistheid aan):

X	Y	X^2	Y^2	XY	Φ	$X\Phi$	$Y\Phi$	Y'
109	202	11.881	40.804	22.018	311	33.899	62.822	217,8
121	242	14.641	58.564	29.282	363	43.923	87.846	230,7
129	261	16.641	68.121	33.669	390	50.310	101.790	239,4
138	240	19.044	57.600	33.120	378	52.164	90.720	249,1
143	267	20.449	71.289	38.181	410	58.630	109.470	254,5
149	242	22.201	58.564	36.058	391	58.259	94.622	260,9
152	254	23.104	64.516	38.608	406	61.712	103.124	264,2
160	282	25.600	79.524	45.120	442	70.720	124.644	272,8
163	270	26.569	72.900	44.010	433	70.579	116.910	276,0
170	296	28.900	87.616	50.320	466	79.220	137.936	283,5
178	280	31.684	78.400	49.840	458	81.524	128.240	292,2
188	308	35.344	94.864	57.904	496	93.248	152.768	302,9
1800	3144	276.058	832.762	478.130	4944	754.188	1.310.892	3144,0



$$\Sigma x^2 = 276.058 - \frac{1800^2}{12} = 6.058$$

$$\Sigma y^2 = 832.762 - \frac{3144^2}{12} = 9.034$$

$$\Sigma xy = 478.130 - \frac{1800 \times 3144}{12} = 6.530$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{6530}{6058} = 1,078 \qquad r = \sqrt{\frac{(\Sigma xy)^2}{\Sigma x^2 . \Sigma y^2}} = +\,0,883$$

$$a = \overline{Y} - b\overline{X} = \frac{3144 - 1,078 \times 1800}{12} = 100,3.$$

$$\underline{Y' = 100,3 + 1,078\ X}$$

Standaardfout (standard error of estimate).

Evenals bij x en y kan men ook hier de standaarddeviatie berekenen van e. Men spreekt dan van de standaardfout. Het aantal vrijheidsgraden is hier echter $N - 2$; immers is $N = 2$ (d.w.z. er zijn twee punten, waardoorheen altijd een rechte lijn getrokken kan worden), dan kan men geen fout waarnemen; dat kan pas, wanneer $N > 2$ is.

De formule voor de standaardfout is dus:

$$s_e = \sqrt{\frac{\Sigma e^2}{N - 2}} \qquad \text{(III 7)}$$

Boven hadden wij voor $\Sigma e^2 = 1.995,94$ gevonden, zodat

$s_e = \sqrt{199,594} = 14,13$ is.

Grafisch kan men de standaardfout aangeven, door boven en onder de regressielijn een lijn te tekenen op een afstand s_e van de regressielijn, aldus:

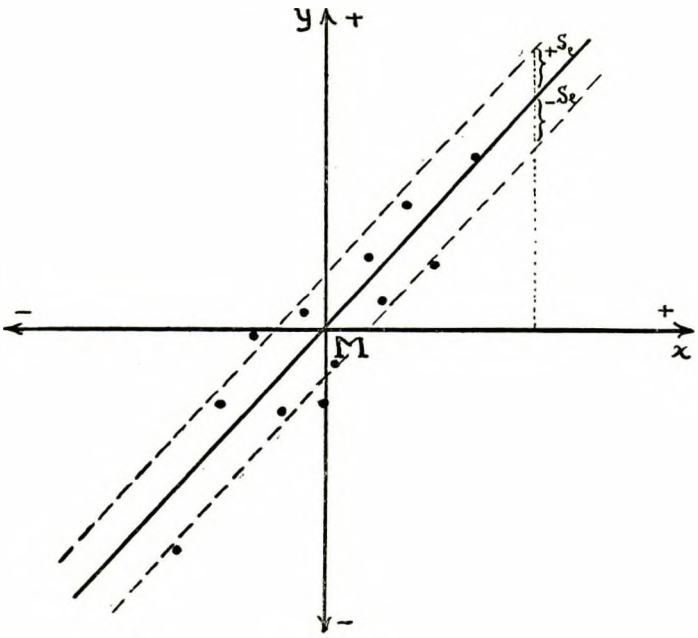


Fig. 5.

Binnen de aldus getekende band moeten de meeste punten vallen, zodat men hiermede tevens een berekening heeft van de beide *tolerantiegrenzen* in de berekening.

Doorgaans zal men e niet berekenen, zodat dan formule (III 7) onpractisch is. Daarom wordt deze formule als volgt herleid:

$$\Sigma e^2 = \Sigma (y - bx)^2 = \Sigma (y^2 - 2bxy + b^2x^2) = \Sigma y^2 - 2b \Sigma xy + b^2 \Sigma x^2 =$$

$$= \Sigma y^2 - 2 \frac{(\Sigma xy)^2}{\Sigma x^2} + \frac{(\Sigma xy)^2}{(\Sigma x^2)^2} . \Sigma x^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} =$$

$$= \Sigma y^2 \left(1 - \frac{(\Sigma xy)^2}{\Sigma x^2 . \Sigma y^2} \right) = \Sigma y^2 . (1 - r^2)$$

Derhalve vinden wij: $s_e = \sqrt{\frac{\Sigma y^2 (1 - r^2)}{N - 2}}$ ..... (III 8)

Voor ons voorbeeld vinden wij dan: $s_e = \sqrt{903{,}4 \times 0{,}221} = 14{,}13$, zoals wij reeds eerder vonden. Uitgedrukt in de gemiddelde waarde van Y, bedraagt de standaardfout $14{,}13 : 262 = 5{,}4\ \%$.

Correlatie tussen waarneming en benadering.

Ter oefening van het voorstellingsvermogen en van de theorie geven wij hieronder nog twee nadere analyses. Allereerst kunnen wij ons afvragen, wat het verband wel zal zijn tussen de waarneming en de benadering hiervan door de regressielijn, m.a.w. wat de correlatie tussen

Y en Y' is. Onderstaande figuur geeft hiervan een beeld, ontleend aan het cijfermateriaal van het tot dusver gevolgde voorbeeld.

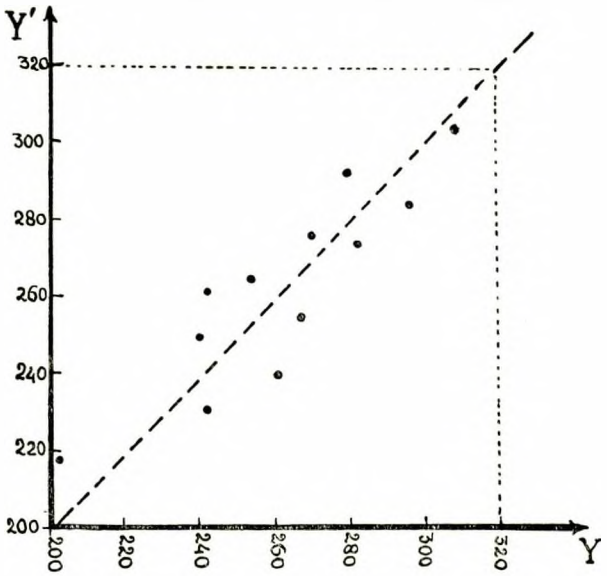


Fig. 6.

Noemen wij voor het verband tussen Y en Y' de correlatie-coëfficient $r_{yy'}$, dan moet gelden:

$r^2_{yy'} = \frac{(\Sigma\, yy')^2}{\Sigma\, y^2 . \Sigma\, (y')^2}$. Aangezien $y' = bx$ is, kunnen wij dit herleiden als volgt:

$$\Sigma\, yy' = \Sigma\, y.bx = b\,\Sigma\, xy = \frac{(\Sigma\, xy)^2}{\Sigma\, x^2}$$

$$\Sigma\, (y')^2 = \Sigma\, (bx)^2 = b^2\,\Sigma\, x^2 = \frac{(\Sigma\, xy)^2}{\Sigma\, x^2}$$

Hieruit volgt, dat $\Sigma\, yy' = \Sigma\, (y')^2$. Wij vinden nu:

$$r^2_{yy'} = \frac{\Sigma\, yy'}{\Sigma\, y^2} = \frac{\Sigma\, (y')^2}{\Sigma\, y^2} \qquad \text{(III 9)}$$

$$= \frac{(\Sigma\, xy)^2}{\Sigma\, x^2 . \Sigma\, y^2}$$

Hieruit volgt dus: $r^2_{yy'} = r^2_{xy}$ (III 10)

m.a.w. de correlatie tussen waarneming en benadering is gelijk aan de correlatie tussen de beide waarnemingen bij enkelvoudige lineaire correlatie.

Formule (III 9) zullen wij later ook tegenkomen bij de multipele correlatierekening.

Correlatie tussen waarneming en waarnemingsfout.

Ook deze correlatie tussen Y en e is op zichzelf vanuit theoretisch standpunt interessant.

Noemen wij hier de correlatiecoëfficient r_{ye}, dan geldt:

$r^2_{ye} = \frac{(\Sigma\, ye)^2}{\Sigma\, y^2 . \Sigma\, e^2}$. Deze vorm kan als volgt vereenvoudigd worden:

$$\Sigma\, ye = \Sigma\, y(y - bx) = \Sigma\, (y^2 - bxy) = \Sigma\, y^2 - b\,\Sigma\, xy =$$
$$= \Sigma y^2 - \frac{(\Sigma\, xy)^2}{\Sigma\, x^2} = \Sigma\, y^2\, (1 - r^2_{xy}).$$

Hieruit volgt, dat $\Sigma\, ye = \Sigma\, e^2$, daar wij immers reeds gezien hebben, dat ook $\Sigma\, e^2 = \Sigma\, y^2\, (1 - r^2_{xy})$ is.

Vullen wij deze uitkomst in, dan resulteert:

$$r^2_{ye} = \frac{\Sigma\, ye}{\Sigma\, y^2} = \frac{\Sigma\, y^2\, (1 - r^2_{xy})}{\Sigma\, y^2} = 1 - r^2_{xy}.$$

Wij vinden dus, dat r^2_{ye} aangeeft de afwezigheid van correlatie. Hierdoor kan men aan de waarde r^2_{xy} eerder de betekenis van maat voor de correlatie geven dan aan r_{xy} [5]).

De tweede regressielijn.

Bij de methode van berekening van de eerste regressielijn gingen wij uit van de veronderstelling, dat x en e niet gecorreleerd waren. Doch ook de andere veronderstelling is onder bepaalde omstandigheden mogelijk, t.w. dat y en e niet gecorreleerd zijn, d.w.z. dat de oorzaak der waarnemingsfout bij x ligt. Grafisch kan dit voorgesteld worden door de volgende figuur:

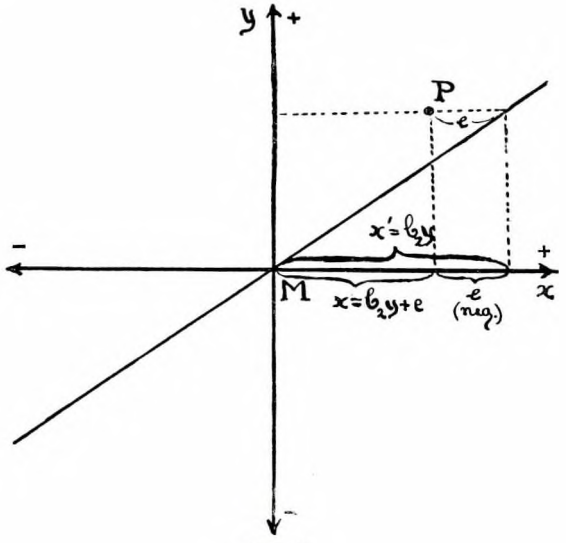


Fig. 7.

[5]) In verband hiermede spreekt men wel van *determinatiecoëfficiënt* voor r^2_{xy} en van *non-determinatiecoëfficiënt* voor $r^2_{ye} = k^2$. Men noemt dan $r_{ye} = k$ de *non-correlatiecoëfficiënt (coëfficiënt of alienation)*. Aangezien r de mate van samenhang aangeeft, kan zij tevens gebruikt worden als een maatstaf van nauwkeurigheid voor een voorspelling met betrekking tot nog niet waargenomen getallenparen; om statistische redenen is echter $1-k$ als *index* of *prediction* beter.

De normaalvergelijking is nu $\Sigma\, ye = 0$, terwijl $e = x - b_2 y$ (men kan immers x en y verwisseld denken), zodat wij vinden:
$\Sigma\, ye = \Sigma\, y(x - b_2 y) = \Sigma\, xy - b_2 \Sigma\, y^2 = 0$, waaruit volgt:

$$b_2 = \frac{\Sigma\, xy}{\Sigma\, y^2}$$

terwijl $a_2 = \overline{X} - b_2 \overline{Y}$ zal zijn.

Voor ons voorbeeld vinden wij dan $b_2 = \frac{6530}{9034} = 0{,}723$ en voor $a_2 = 150 - 0{,}723 \times 262 = -39{,}43$, zodat wij vinden:
$X' = -39{,}43 + 0{,}723\ Y$, waaruit volgt $Y = 54{,}5 + 1{,}383\ X'$

Deze regressielijn wijkt nogal sterk af van de eerste. De betekenis ervan is voor ons voorbeeld echter gering, daar men bij de bepaling van de productiehoeveelheid wel nauwelijks fouten zal maken.

Het is echter in dit verband van belang erop te wijzen, dat men bij correlaties van andere aard steeds als X moet kiezen de grootheid, welker waarnemingsfouten te verwaarlozen zijn. Is dit niet mogelijk, dan kan men het beste doen, om de beide regressielijnen te berekenen en daarna een lijn te trekken, die b.v. de hoek tussen beide regressielijnen middendoor deelt; deze lijn houdt dan rekening met de waarnemingsfouten van zowel de x als de y.

Daar $b_1 = \frac{\Sigma\, xy}{\Sigma\, x^2}$ en $b_2 = \frac{\Sigma\, xy}{\Sigma\, y^2}$ is, moet $b_1 b_2 = r^2$ zijn.

In bovenstaande vergelijking voor de tweede regressielijn is de regressiecoëfficiënt $1{,}383 = \frac{1}{0{,}723} = \frac{1}{b_2}$.

Neemt men nu het meetkundige gemiddelde tussen b_1 en $\frac{1}{b_2}$, d.w.z. $b_3 = \sqrt{\frac{b_1}{b_2}} = \frac{r}{b_2} = \frac{0{,}883}{0{,}723} = 1{,}221$, dan krijgt men een bruikbare regressielijn, die tussen de beide andere regressielijnen in loopt. Het rekenkundige gemiddelde komt op

$$b_4 = \frac{b_1 + \frac{1}{b_2}}{2} = \frac{r + 1}{2b_2} = 1{,}230$$

en wijkt dus van de vorige waarde slechts weinig af, hetgeen het gevolg is van de vrij hoge waarde van r (bij $r = 1$ is $b_3 = b_4$, hetgeen vanzelf spreekt, daar dan de beide regressielijnen moeten samenvallen). De lijn, die de hoek tussen de beide regressielijnen midden door deelt, kan men berekenen met behulp van de regressiecoëfficient

$$b_5 = \frac{b_1 - b_2}{r^2 + 1} + \sqrt{1 + \left(\frac{b_1 - b_2}{r^2 + 1}\right)^2} = 1{,}219$$

De practische waarde van dergelijke constructies achten wij zeer gering. Zoals gezegd zal het in de regel wel mogelijk zijn één variabele te vinden, van welker waarde wij kunnen aannemen, dat zij practisch juist is.

Regressielijn zonder constante factor.

Natuurlijk had men de normaalvergelijking (III 1) óók kunnen schrij-

ven in de vorm $\Sigma Xe = 0$, d.w.z. in plaats van de afwijking van het gemiddelde wordt het oorspronkelijk getal genomen. Men krijgt dan:

$$\Sigma Xe = \Sigma X\,(Y - bX - a) = \Sigma XY - b\Sigma X^2 - a\Sigma X = 0.$$

Maar, zoals men ziet, krijgt men dan één vergelijking met twee onbekenden, die echter tezamen met de formule voor a, (III 3), dezelfde uitkomsten moet opleveren. Men zou deze methode kunnen volgen bij de methode der oorspronkelijke getallen:

$478.130 - 276.058\,b - 1800\,a = 0$. Substitueert men hierin de waarde voor $a = 262 - 150\,b$, dan krijgt men:

$$\begin{aligned} 478.130 - 276.058\,b - 471.600 + 270.000\,b &= 0 \\ 6.530 &= 6058\,b \\ b &= 1{,}078 \end{aligned}$$

Dat de berekening op hetzelfde neer moet komen als de vroegere berekening is duidelijk, als men opmerkt, dat $478.130 - 471.600 = \Sigma\,xy$ is en dat $276.058 - 270.000 = \Sigma\,x_2$. Aangezien men deze waarden toch moet berekenen voor de correlatiecoëfficiënt, heeft de eerder gegeven methode de voorkeur.

In één geval echter lijkt dit niet zo te zijn, n.l. als men a priori ervan overtuigd is, dat de regressielijn door de oorsprong der figuur moet gaan, b.v. als men theoretisch ervan overtuigd is, dat er geen vaste kosten kunnen zijn. In dat geval immers is $a = 0$, zodat uit bovenstaande normaalvergelijking direct resulteert:

$b = \dfrac{\Sigma XY}{\Sigma X^2} = \dfrac{478.130}{276.058} = 1{,}7320$, zodat onze regressielijn dan zou luiden: $Y' = 1{,}732\,X$. Opgemerkt zij echter, dat hoewel de berekening mathematisch juist is, de uitkomst ons niet bevredigt, omdat $\Sigma\,Y' = 1{,}732 \times 1800 = 3117{,}6$, d.i. minder dan $\Sigma\,Y = 3144$, zodat de gelijkheid $\Sigma\,Y' = \Sigma\,Y$ nu niet meer opgaat.

Deze gelijkheid gaat wel op, als men uitgaat van de formule voor a, welke nu wordt $0 = \bar{Y} - b\bar{X}$, waaruit volgt.

$$b = \frac{\Sigma\,Y}{\Sigma\,X} = \frac{1800}{3144} = 1{,}7467.$$

Om deze reden, geven wij aan de laatste uitkomst de voorkeur. In dit verband zij opgemerkt, dat de mate, waarin de variabiliteit (der kosten) optreedt, blijkbaar afhangt van de mate waarin de waarde $\dfrac{\Sigma\,Y}{\Sigma\,X}$ benaderd wordt door de waarde $\dfrac{\Sigma\,xy}{\Sigma\,x^2}$ Om deze reden zullen wij de waarde voor $V = \dfrac{\Sigma\,xy}{\Sigma\,x^2} \cdot \dfrac{\Sigma\,X}{\Sigma\,Y}$ de *variabiliteitsgraad* noemen, welke in ons geval draagt $\dfrac{1{,}0779}{1{,}7467} = 61{,}7\ \%$. Zoals men gemakkelijk zal verifiëren, kan men voor de variabiliteitsgraad schrijven $V = 1 - \dfrac{a}{\bar{Y}}$ welke vorm ook voor de multipele correlaties kan worden gebruikt en welke men zeer gemakkelijk kan uitrekenen.

Men zou nu kunnen verwachten, dat men de correlaties tussen waarneming en benadering voor de gevallen, dat er al dan niet een constante

factor in de uitgangsformule is aangenomen, zou kunnen vergelijken door middel van de correlatiecoëfficiënt $r_{yy'}$. In deze verwachting wordt men echter teleurgesteld, daar ook voor het geval, dat in de uitgangsformule geen constante factor is opgenomen, $r^2_{yy'} = r^2_{xy}$ is. Dit komt, doordat voor dit geval ook geldt $y' = bx$, waaruit wederom volgt: $\Sigma\, yy' = b\,\Sigma\, xy$ en $\Sigma\,(y')^2 = b^2\,\Sigma\, x^2$, als gevolg waarvan de factor b uit de berekening wegvalt.

Met de correlatiecoëfficiënt kan men dus niet opsporen, of men al dan niet een constante factor in de uitgangsformule moet aannemen. Men doet daarom verstandig deze aanname steeds te doen; pas na de berekening van de regressielijn komt dan uit, of deze al dan niet door de oorsprong gaat.

(Wordt vervolgd)