M.O. 273g.

FOR OFFICIAL USE.

AIR MINISTRY

METEOROLOGICAL OFFICE

PROFESSIONAL NOTES No. 47.

(*Seventh Number of Volume IV.*)

REGRESSION EQUATIONS

WITH

MANY VARIATES

BY

C. E. P. BROOKS, D.Sc.

Published by the Authority of the Meteorological Committee.

LONDON:
PRINTED AND PUBLISHED BY HIS MAJESTY'S STATIONERY OFFICE
To be purchased directly from H.M. STATIONERY OFFICE at the following addresses:
Adastral House, Kingsway, London, W.C.2; 120, George Street, Edinburgh;
York Street, Manchester; 1, St. Andrew's Crescent, Cardiff;
15, Donegall Square West, Belfast;
or through any Bookseller.
1927.

Price 3*d*. Net.

40-57-47

REGRESSION EQUATIONS WITH MANY VARIATES

By C. E. P. Brooks, D.Sc.

The usual method of calculating regression equations from partial correlation coefficients in the form :

$$x_1 = b_{12\cdot 34 \ldots n}\, x_2 + b_{13\cdot 24 \ldots n}\, x_3 + \ldots + b_{1n\cdot 23 \ldots (n-1)}\, x_n \qquad (1)$$

is rapid when n is not greater than 5, and is practicable though somewhat laborious when $n = 6$, but with more than six variates the amount of arithmetic involved rapidly becomes prohibitive. The following method greatly lessens the arithmetic and gives practicable results.

Let the variates be expressed in the following notation :

$$x,\ y_0,\ y_1,\ y_2 \ldots y_{n-2}$$

where x is the variate which is to be represented by the regression series. Further, for convenience of printing, write r_{xy} as $[xy]$. Let m_1 represent the series remaining after the total effect of the y_0 series has been eliminated from the x series. Then :

$$m_1 = x - [xy_0]\frac{\sigma_x}{\sigma_{y_0}} y_0 \qquad (2)$$

$$\sigma_{m_1} = \sigma_x\ (1 - [xy_0]^2)^{\frac{1}{2}} \qquad (3)$$

Correlating m_1 with y_1, we have :

$$[m_1 y_1] = \frac{\Sigma x y_1 - [xy_0]\frac{\sigma_x}{\sigma_{y_0}}\ \Sigma y_0 y_1}{N\ \sigma_{m_1}\ \sigma_{y_1}}$$

$$= \frac{[xy_1] - [xy_0] \times [y_0\, y_1]}{(1 - [xy_0]^2)^{\frac{1}{2}}}, \qquad (4)$$

N being the number of observations.

This can be written :

$$[m_1\, y_1] = [xy_1 \cdot y_0]\ (1 - [y_0\, y_1]^2)^{\frac{1}{2}}$$

but as it is proposed to work entirely with coefficients of zero order, this complication is not required. Further, the regression coefficient $b(m_1\, y_1)$ is given by :

$$b\ (m_1\, y_1) = [m_1\, y_1]\ \sigma_{m_1}/\sigma_{y_1} \qquad (5)$$

We can also correlate m_1 with y_2 and with y_3, giving :

$$[m_1\,y_2] = \frac{[xy_2] - [xy_0] \times [y_0\,y_2]}{(1 - [xy_0]^2)^{\frac{1}{2}}} \tag{6}$$

$$[m_1\,y_3] = \frac{[xy_3] - [xy_0] \times [y_0\,y_3]}{(1 - [xy_0]^2)^{\frac{1}{2}}} \tag{7}$$

Let m_2 represent the series remaining after the total effect of the y_1 series has been eliminated from the m_1 series. Then :

$$m_2 = m_1 - [m_1\,y_1]\frac{\sigma_{m_1}}{\sigma_{y_1}}\,y_1 \tag{8}$$

This is identical in form with (2) and hence by (4)

$$[m_2\,y_2] = \frac{[m_1\,y_2] - [m_1\,y_1] \times [y_1\,y_2]}{(1 - [m_1\,y_1]^2)^{\frac{1}{2}}} \tag{9}$$

Similarly

$$[m_2\,y_3] = \frac{[m_1\,y_3] - [m_1\,y_1] \times [y_1\,y_3]}{(1 - [m_1\,y_1]^2)^{\frac{1}{2}}} \tag{10}$$

Generalising

$$[m_p\,y_q] = \frac{[m_{p-1}\,y_q] - [m_{p-1}\,y_{p-1}] \times [y_{p-1}\,y_q]}{(1 - [m_{p-1}\,y_{p-1}]^2)^{\frac{1}{2}}} \tag{11}$$

where $m_0 = x$; further,

$$b\,(m_p\,y_p) = [m_p\,y_p]\,\frac{\sigma_{m_p}}{\sigma_{y_p}} \tag{12}$$

Proceeding step by step in this way we obtain the regression equation :

$$x = b\,(xy_0)\,y_0 + b\,(m_1\,y_1)\,y_1 + \ldots b\,(m_p\,y_p)\,y_p \tag{13}$$

using no coefficients of higher than zero order. It should be remarked that this is not the true regression equation, which can only be obtained by the full method of partial correlation. It represents a method of successive approximation, which is equivalent to the following process :

(1) Find the regression coefficient of y_0 on x.

(2) Calculate the residuals after allowing for the effect of y_0 on x.

(3) Form a new regression coefficient of y_1 on these residuals, and calculate a second set of residuals, and so on.

As an example, it is required to find a formula to express the velocity of the north-east trade wind in terms of barometric pressure at Ponta Delgada, Gibraltar, Bermuda and Sierra Leone.

The order in which the independent variates are taken should not make any difference to the accuracy of the final result, though it alters the values of the various regression coefficients. In general it would probably be best to take them in the order of magnitude of their coefficients of correlation with the m series, as some of the later members of the regression equation will then stand a good chance of being negligible, thus simplifying the equation. Thus for y_0 is taken the variate which has the highest correlation with x, for y_1 that which has the highest correlation with m_1, and so on. As an example we have the departures from normal of the 30 monthly means for October to March 1902 to 1906.

x = wind velocity, in m.p.h.
y_0 = pressure at Sierra Leone.
y_1 = pressure at Ponta Delgada.
y_2 = pressure at Gibraltar.
y_3 = pressure at Bermuda.

The calculation is as follows :

TABLE I.—COEFFICIENTS OF ZERO ORDER.

Variates ..	xy_0	xy_1	xy_2	xy_3	$y_0 y_1$	$y_0 y_2$	$y_0 y_3$	$y_1 y_2$	$y_1 y_3$	$y_2 y_3$
Coefficient ..	−·4783	+·3895	−·1786	+·0216	−·0938	+·1338	−·0431	+·3554	+·5240	+·6720

We also require $(1-[xy_0]^2)^{\frac{1}{2}}=0\cdot8782$.

TABLE II.—CALCULATION OF COEFFICIENTS $[m_p y_q]$ by (11).

Co-efficient required.	Numerator.					Denominator.	Coefficient required.	
	1st Term.		Product.		Difference.			
$[m_1 y_1]$	$[xy_1]$	+·3895	$[xy_0]\times[y_0 y_1]$	+·0448	+·3447	·8782	$[m_1 y_1]$	+·3924
$[m_1 y_2]$	$[xy_2]$	−·1786	$[xy_0]\times[y_0 y_2]$	−·0640	−·1146		$[m_1 y_2]$	−·1305
$[m_1 y_3]$	$[xy_3]$	+·0216	$[xy_0]\times[y_0 y_3]$	+·0206	+·0010		$[m_1 y_3]$	+·0011
$[m_2 y_2]$	$[m_1 y_2]$	−·1305	$[m_1 y_1]\times[y_1 y_2]$	+·1394	−·2699	·9198	$[m_2 y_2]$	−·2934
$[m_2 y_3]$	$[m_1 y_3]$	+·0011	$[m_1 y_1]\times[y_1 y_3]$	+·2056	−·2045		$[m_2 y_3]$	−·2224
$[m_3 y_3]$	$[m_2 y_3]$	−·2224	$[m_2 y_2]\times[y_2 y_3]$	−·1971	−·0253	·9560	$[m_3 y_3]$	−·0265

The calculation was carried through on a pocket calculator and is subject to an error of 2 or 3 in the fourth figure, which has, however, been retained for working purposes.

The final step is the calculation of the regression coefficients. For this purpose we require the standard deviations of the m's, given by the expression :

$$\sigma_{m_{p+1}} = \sigma_{m_p}\,(1-[m_p y_p]^2)^{\frac{1}{2}} \qquad (14)$$

The actual coefficients are calculated by (12).

TABLE III.—CALCULATION OF REGRESSION COEFFICIENTS.

Correlation Coefficients.		Standard Deviations.				Regression Coefficients (by 12).	
		m's (by 14).		*y*'s			
$[xy_0]$	$-\cdot 4783$	x	$9\cdot 120$	y_0	$3\cdot 815$	xy_0	$-1\cdot 14$
$[m_1y_1]$	$+\cdot 3924$	m_1	$8\cdot 009$	y_1	$4\cdot 218$	m_1y_1	$+0\cdot 75$
$[m_2y_2]$	$-\cdot 2934$	m_2	$7\cdot 367$	y_2	$4\cdot 076$	m_2y_2	$-0\cdot 53$

The coefficient $[m_3y_3]$ being only $\cdot 026$, it is not necessary to calculate the regression coefficient m_3y_3.

The regression equation therefore has the form :

$$(x) = -1\cdot 14y_0 + 0\cdot 75\, y_1 - 0\cdot 53\, y_2 \qquad (15)$$

The values of (x) calculated by this expression have a correlation R with the original series of x which is given by :

$$(1 - R^2) = (1 - [xy_0]^2)\ (1 - [m_1\, y_1]^2)\ (1 - [m_2\, y_2]^2)$$

This gives $R = 0\cdot 638$. The regression equation calculated from the partial coefficients in the usual way is :

$$x' = -0\cdot 97\, y_0 + 1\cdot 04\, y_1 - 0\cdot 53\, y_2 - 0\cdot 29\, y_3$$

and the correlation between x and x' is $0\cdot 651$. Hence the gain in accuracy by the full method is not large. Against it may be set the fact that the calculation of equation (15), in spite of the unfamiliarity of the method, occupied barely 45 minutes, while the full method, although very familiar, required several pages of foolscap and nearly four hours of time. It must be pointed out that this short method should not be used when it is intended to deduce physical connexions from the results, but only as a means of deducing formulae for interpolation or forecasting.