MISCELLANEA.

I. On a General Method of determining the successive terms in a Skew Regression Line*.

By KARL PEARSON, F.R.S.

Let x and y be the two variates. Suppose the x-range divided up into any not necessarily equal intervals h_x giving arrays of y of which the mean of those which centre at x (lying somewhere in h_x) is $\bar{y}_x$ and the array total n_x. Let $\bar{x}$, $\bar{y}$, σ_x, σ_y be the means and standard deviations of the total population N of the two variates. Assume the form of the regression line of y on x to be:

$$\frac{\bar{y}_x - \bar{y}}{\sigma_y} = a_0 \psi_0 + a_1 \psi_1 + \ldots + a_n \psi_n \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(i)},$$

where a_0, $a_1 \ldots a_n$ are absolute constants to be determined and ψ_s is an orthogonal function of $\frac{x-\bar{x}}{\sigma_x}$, i.e. is subject to the condition that:

$$S(n_x \psi_s \psi_{s'}) = 0, \quad s \text{ and } s' \text{ unequal} \ldots\ldots\ldots\ldots\ldots\ldots\text{(ii)},$$

if the summation S be taken for all values of x corresponding to the arbitrary system of arrays.

Further let us suppose† that:

$$q_{1t} = \frac{S\{n_x(\bar{y}_x - \bar{y})(x-\bar{x})^t\}}{N\sigma_y \sigma_x{}^t} = \frac{\Sigma\{n_{xy}(y-\bar{y})(x-\bar{x})^t\}}{N\sigma_y\sigma_x{}^t} = p_{1t}/(\sigma_y \sigma_x{}^t)$$

in the usual product-moment notation.

Clearly if the ψ_s's can be determined we must have by virtue of (ii) from (i)

$$S\left(\frac{n_x(\bar{y}_x - \bar{y})}{N\sigma_y}\psi_n\right) = a_n \frac{S(n_x \psi_n{}^2)}{N} \ldots\ldots\ldots\ldots\ldots\ldots\text{(iii)}.$$

Now if ψ_n be determined as an integral function of $(x-\bar{x})/\sigma_x$ of the nth degree, the left-hand side of (iii) is expressible in terms of the q_{1t}'s, or the product-moments of the correlation distribution. Thus a_n will be determined. Let us write:

$$\left.\begin{aligned} \kappa_n &= \frac{S\{n_x(\bar{y}_x - \bar{y})\psi_n\}}{N\sigma_y} \\ \text{and} \qquad \lambda_n &= \frac{S(n_x\psi_n{}^2)}{N} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(iv)}.$$

$$\text{Then} \qquad a_n = \kappa_n/\lambda_n \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(iii) bis}$$

is known, if the ψ_s's have once been determined, from the product-moments of the system of the correlated variates, and the moment coefficients of the x-variate.

Square (i), multiply by n_x/N and sum for all arrays, we have:

$$S\left(\frac{n_x(\bar{y}_x - \bar{y})^2}{N\sigma_y{}^2}\right) = \eta^2{}_{y,x} = a_0{}^2\lambda_0 + a_1{}^2\lambda_1 + \ldots + a_n{}^2\lambda_n \ldots\ldots\ldots\ldots\ldots\ldots\text{(v)}.$$

Here $\eta_{y,x}$ is the well-known correlation ratio of y-variate on x-variate, and must always lie between 0 and 1. The series on the right-hand side of (v) consists of a system of squares, and accordingly every additional term we take in our series for the regression line must carry us nearer this value of $\eta^2{}_{y,x}$. Unless λ_n tends to zero the a_n's must grow smaller and smaller, or we have considerable anticipation of the convergency of the series, but this does not amount of course to definite proof.

* Reproduced from lecture notes.

† S denotes summation of all arrays, Σ denotes summation for each individual point.

Now suppose we had to determine our a_s's from (i) by the method of least squares, each array being weighted by its frequency. We should have to make

$$u^2 = S\left\{\frac{n_x}{N}\left(\frac{\bar{y}_x - \bar{y}}{\sigma_y} - a_0\psi_0 - a_1\psi_1 - \dots - a_n\psi_n\right)^2\right\}$$

a minimum.

But $$\frac{du^2}{da_s} = 0 = 2S\left(\frac{n_x(\bar{y}_x - \bar{y})}{N\sigma_y}\psi_s\right) - 2a_s\frac{S(n_x\psi_s^2)}{N},$$

or $$a_s = \kappa_s/\lambda_s,$$

in agreement with the result obtained in (iii) bis above as a result of the ψ_s's being orthogonal functions. Now the fit by least squares to the means of the arrays is precisely the same as the fit by least squares to the whole swarm of points. In other words if we fit our regression line as above indicated to the whole population it will be "the best" fitting curve, if we make the assumption that least squares does give the best fit*. Anyhow it is likely to be a good fit, and that suffices for our present purposes.

We will now write $X = (x - \bar{x})/\sigma_x$ and $Y_x = (\bar{y}_x - \bar{y})/\sigma_y$, whence it follows that $q_{1t} = S(n_x Y_x X^t)/N$. We shall further write:

$$\mu_s = S(n_x X^s)\,\sigma_x{}^s/N,$$

and $$\beta_{2s} = \mu_{2s+2}/\sigma_x{}^{2s+2} = \frac{S\{n_x(x-\bar{x})^{2s+2}\}}{N\sigma_x{}^{2s+2}} = \frac{S(n_x X^{2s+2})}{N},$$

$$\beta_{2s-1} = \frac{\mu_{2s+1}\mu_3}{\sigma^{2s+4}}.$$

We may rewrite (i) $$Y = a_0\psi_0 + a_1\psi_1 + \dots + a_n\psi_n \quad \text{..............................(i) bis.}$$

Multiply both sides by $n_x\psi_0/N$ and sum:

$$\frac{S(n_x Y\psi_0)}{N} = a_0\frac{S(n_x)}{N}\psi_0{}^2 = a_0,$$

if we assume $\psi_0 = 1$ as we are at liberty to do. But the left-hand side is zero. Therefore $a_0 = 0$.

Now assume $\psi_1 = X - c_{10}\psi_0$ and multiply both sides by $n_x\psi_0/N$ and sum:

$$\frac{S(n_x\psi_0\psi_1)}{N} = 0 \text{ by hypothesis} = \frac{S(n_x X\psi_0)}{N} - c_{10}\frac{S(n_x\psi_0{}^2)}{N}.$$

But $$\frac{S(n_x X\psi_0)}{N} = 0, \therefore c_{10} = 0, \text{ and } \psi_1 = X.$$

Multiply both sides by $n_x X/N$ and sum:

$$\frac{S(n_x YX)}{N} = r = a_1\frac{S(n_x X\psi_1)}{N} = a_1\frac{S(n_x X^2)}{N} = a_1.$$

Hence $a_1 = r$.

Now assume $\psi_2 = X^2 - c_{21}\psi_1 - c_{20}\psi_0$. Multiply by $n_x\psi_0/N$ and sum:

$$0 = \frac{S(n_x X^2)}{N}\psi_0 - 0 - c_{20} = 1 - c_{20}.$$

Thus $c_{20} = 1$.

Multiply by $n_x\psi_1/N$ and sum:

$$0 = \frac{S(n_x\psi_1 X^2)}{N} - c_{21}\frac{S(n_x\psi_1{}^2)}{N},$$

$$\frac{S(n_x X^3)}{N} = \sqrt{\beta_1} = c_{21}\frac{S(n_x X^2)}{N} = c_{21}.$$

Thus $c_{21} = \sqrt{\beta_1}$.

* In order that this statement should be absolutely true our arrays would have to follow the normal or Gaussian distribution. This produces, however, an unnecessary limitation. It is known, but possibly not well known, that expansions in orthogonal functions as defined by (ii) give least square fits. I am rather inclined to think that this is an argument in favour of least square fits, rather than a justification of the expansion, i.e. that the method of least squares has a wider validity than Gauss' proof provides.

$\sqrt{\beta_1}$ must of course be given the same sign as μ_3. Accordingly we have:

$$\psi_2 = X^2 - \sqrt{\beta_1}X - 1.$$

Square, multiply by n_x/N and sum and we have:

$$\frac{S(n_x\psi_2^2)}{N} = \frac{S(n_x\psi_2 X^2)}{N} = \beta_2 - \beta_1 - 1.$$

It remains to find:

$$\frac{S(n_x Y\psi_2)}{N} = S\{n_x Y(X^2 - \sqrt{\beta_1}X - 1)/N\} = q_{12} - r\sqrt{\beta_1}.$$

Thus it follows from (iii) that $\quad a_2 = (q_{12} - r\sqrt{\beta_1})/(\beta_2 - \beta_1 - 1)$.

We now proceed to find ψ_3:

$$\psi_3 = X^3 - c_{32}\psi_2 - c_{31}\psi_1 - c_{30}\psi_0.$$

Multiply by $n_x\psi_0/N$ and sum:

$$0 = \frac{S(n_x X^3)}{N} - c_{30}, \quad \text{or} \quad c_{30} = \sqrt{\beta_1}.$$

Multiply by $n_x\psi_1/N$ and sum:

$$0 = \frac{S(n_x X^4)}{N} - c_{31}\frac{S(n_x X^2)}{N}, \quad \text{or} \quad c_{31} = \beta_2.$$

Multiply by $n_x\psi_2/N$ and sum:

$$0 = \frac{S\{n_x X^3(X^2 - \sqrt{\beta_1}X - 1)\}}{N} - c_{32}\frac{S(n_x\psi_2^2)}{N},$$

$$\frac{\mu_5}{\sigma_x^5} - \sqrt{\beta_1}\beta_2 - \sqrt{\beta_1} = c_{32}(\beta_2 - \beta_1 - 1),$$

$$\frac{\beta_3 - \beta_1\beta_2 - \beta_1}{\sqrt{\beta_1}} = c_{32}(\beta_2 - \beta_1 - 1),$$

$$c_{32} = \frac{\beta_3 - \beta_1\beta_2 - \beta_1}{\sqrt{\beta_1}(\beta_2 - \beta_1 - 1)}.$$

Accordingly:

$$\psi_3 = X^3 - \frac{\beta_3 - \beta_1\beta_2 - \beta_1}{\sqrt{\beta_1}(\beta_2 - \beta_1 - 1)}\psi_2 - \beta_2\psi_1 - \sqrt{\beta_1}\psi_0,$$

or

$$\psi_3 = X^3 - \frac{\beta_3 - \beta_1\beta_2 - \beta_1}{\sqrt{\beta_1}(\beta_2 - \beta_1 - 1)}X^2 + \frac{\beta_3 - \beta_2^2 + \beta_2 - \beta_1}{\beta_2 - \beta_1 - 1}X + \frac{\beta_3 - 2\beta_1\beta_2 + \beta_1^2}{\sqrt{\beta_1}(\beta_2 - \beta_1 - 1}.$$

We have next to determine κ_3:

$$\begin{aligned}\kappa_3 &= \frac{S(n_x Y\psi_3)}{N}\\ &= q_{13} - \frac{\beta_3 - \beta_1\beta_2 - \beta_1}{\sqrt{\beta_1}(\beta_2 - \beta_1 - 1)}q_{12} + \frac{\beta_3 - \beta_2^2 + \beta_2 - \beta_1}{\beta_2 - \beta_1 - 1}r\\ &= q_{13} - \beta_2 r - \frac{\beta_3 - \beta_1\beta_2 - \beta_1}{\sqrt{\beta_1}(\beta_2 - \beta_1 - 1)}(q_{12} - r\sqrt{\beta_1}).\end{aligned}$$

Let us write:

$$\epsilon_{12} = q_{12} - r\sqrt{\beta_1}, \quad \epsilon_{13} = q_{13} - \beta_2 r,$$

$$\phi_2 = \beta_2 - \beta_1 - 1, \quad \phi_3 = (\beta_3 - \beta_1\beta_2 - \beta_1)/\sqrt{\beta_1}.$$

Thus

$$\kappa_3 = \epsilon_{13} - \frac{\phi_3}{\phi_2}\epsilon_{12}.$$

We have next to find λ_3:

$$\begin{aligned}\lambda_3 &= \frac{S(n_x\psi_3^2)}{N} = S\left\{\frac{n_x X^3}{N}\psi_3\right\}\\ &= \beta_4 - \frac{(\beta_3 - \beta_1\beta_2 - \beta_1)\beta_3}{\beta_1(\beta_2 - \beta_1 - 1)} + \frac{(\beta_3 - \beta_2^2 + \beta_2 - \beta_1)\beta_2}{\beta_2 - \beta_1 - 1} + \frac{\beta_3 - 2\beta_1\beta_2 + \beta_1^2}{\beta_2 - \beta_1 - 1}\\ &= \beta_4 - \beta_2^2 - \beta_1 - \frac{(\beta_3 - \beta_1\beta_2 - \beta_1)^2}{\beta_1(\beta_2 - \beta_1 - 1)}\\ &= \phi_4 - \phi_3^2/\phi_2 \quad \text{f} \ \phi_4 = \beta_4 - \beta_2^2 - \beta_1.\end{aligned}$$

Accordingly: $$a_3 = \frac{\kappa_3}{\lambda_3} = \frac{\epsilon_{13}\phi_2 - \epsilon_{12}\phi_3}{\phi_2\phi_4 - \phi_3^2}.$$

We now pass to the fourth order regression line function:

$$\psi_4 = X^4 - c_{43}\psi_3 - c_{42}\psi_2 - c_{41}\psi_1 - c_{40}\psi_0,$$

and we find:

$$\frac{S(n_x X^4 \psi_3)}{N} = c_{43}\frac{S(n_x \psi_3^2)}{N} = c_{43}\lambda_3$$

$$= \frac{\beta_5}{\sqrt{\beta_1}} - \frac{(\beta_3 - \beta_1\beta_2 - \beta_1)\beta_4}{\sqrt{\beta_1}(\beta_2 - \beta_1 - 1)} + \frac{(\beta_3 - \beta_2^2 + \beta_2 - \beta_1)\beta_3}{\sqrt{\beta_1}(\beta_2 - \beta_1 - 1)} + \frac{(\beta_3 - 2\beta_1\beta_2 + \beta_1^2)\beta_2}{\sqrt{\beta_1}(\beta_2 - \beta_1 - 1)}$$

$$= \{\beta_5(\beta_2 - \beta_1 - 1) - \beta_4(\beta_3 - \beta_1\beta_2 - \beta_1) + \beta_3(\beta_3 - \beta_2^2 + 2\beta_2 - \beta_1) + \beta_1\beta_2(\beta_1 - 2\beta_2)\} \div \sqrt{\beta_1}(\beta_2 - \beta_1 - 1).$$

Let $$\phi_5 = \frac{\beta_5(\beta_2 - \beta_1 - 1) - \beta_4(\beta_3 - \beta_1\beta_2 - \beta_1) + \beta_3(\beta_3 - \beta_2^2 + 2\beta_2 - \beta_1) + \beta_1\beta_2(\beta_1 - 2\beta_2)}{\sqrt{\beta_1}}.$$

Thus: $$c_{43} = \frac{\phi_5}{\phi_4\phi_2 - \phi_3^2}.$$

Again: $$\frac{S(n_x X^4 \psi_2)}{N} = c_{42}\lambda_2,$$

$$\beta_4 - \beta_3 - \beta_2 = c_{42}(\beta_2 - \beta_2 - 1),$$

or $$\phi_4 - \phi_3\sqrt{\beta_1} + \phi_2\beta_2 = \phi_4', \text{ say, } = c_{42}\phi_2.$$

Thus: $$c_{42} = \frac{\phi_4'}{\phi_2}.$$

Further: $$\frac{\beta_3}{\sqrt{\beta_1}} = c_{41},$$

and $$\beta_2 = c_{40}.$$

Thus: $$\psi_4 = X^4 - \frac{\phi_5}{\phi_4\phi_2 - \phi_3^2}\psi_3 - \frac{\phi_4'}{\phi_2}\psi_2 - \frac{\beta_3}{\sqrt{\beta_1}}\psi_1 - \beta_2\psi_0.$$

To find a_4 we must determine κ_4 and λ_4:

$$\lambda_4 = \frac{S(n_x X^4 \psi_4)}{N}$$

$$= \beta_6 - \frac{\phi_5^2}{(\phi_4\phi_2 - \phi_3^2)\phi_2} - \frac{\phi_4'}{\phi_2}(\beta_4 - \beta_3 - \beta_2) - \frac{\beta_3^2}{\beta_1} - \beta_2^2$$

$$= \phi_6 - \frac{\phi_5^2}{\phi_2(\phi_4\phi_2 - \phi_3^2)} - \frac{\phi_4'^2}{\phi_2},$$

where $$\phi_6 = \beta_6 - \frac{\beta_3^2}{\beta_1} - \beta_2^2,$$

$$\kappa_4 = \frac{S(n_x Y \psi_4)}{N} = q_{14} - \frac{\phi_5}{\phi_4\phi_2 - \phi_3^2}\left(\epsilon_{13} - \frac{\phi_3}{\phi_2}\epsilon_{12}\right) - \frac{\phi_4'}{\phi_2}(q_{12} - r\sqrt{\beta_1}) - \frac{\beta_3}{\sqrt{\beta_1}}r.$$

Let us write $q_{14} - \dfrac{\beta_3}{\sqrt{\beta_1}}r = \epsilon_{14}$, then

$$a_4 = \frac{\epsilon_{14} - \dfrac{\phi_5}{\phi_4\phi_2 - \phi_3^2}\left(\epsilon_{13} - \dfrac{\phi_3}{\phi_2}\epsilon_{12}\right) - \dfrac{\phi_4'}{\phi_2}\epsilon_{12}}{\phi_6 - \dfrac{\phi_5^2}{\phi_2(\phi_4\phi_2 - \phi_3^2)} - \dfrac{\phi_4'^2}{\phi_2}}.$$

We have thus obtained the regression orthogonal functions up to the fourth order. Higher order terms can also be found, but their expressions become very complicated and such expressions involving fifth product-moments and eighth marginal total moments will be subject to very large probable errors.

We can now by aid of (v) express $\eta_{y,x}$. We have:

$$\eta^2_{y,x}=r^2+\epsilon_{12}^2/\phi_2+\left(\epsilon_{13}-\frac{\phi_3}{\phi_2}\epsilon_{12}\right)^2\Big/(\phi_4-\phi_3^2/\phi_2)$$

$$+\frac{\left\{\epsilon_{14}-\frac{\phi_4'}{\phi_2}\epsilon_{12}-\frac{\phi_5}{\phi_2^2}\left(\epsilon_{13}-\frac{\phi_3}{\phi_2}\epsilon_{12}\right)\Big/(\phi_4-\phi_3^2/\phi_2)\right\}^2}{\phi_6-\phi_4'^2/\phi_2-\frac{\phi_5^2}{\phi_2^2}\Big/(\phi_4-\phi_3^2/\phi_2)}+\text{etc.}\dots\dots$$

The conditions therefore for linear regression, or $\eta_{rx}=r$, are:

$$\epsilon_{12}=\epsilon_{13}=\epsilon_{14}=\text{etc.}=0\dots\dots$$

That is: $$q_{12}=r\sqrt{\beta_1},\quad q_{13}=r\beta_2,\quad q_{14}=r\frac{\beta_3}{\sqrt{\beta_1}},\ \text{etc.}\dots\dots$$

For parabolic regression: $$\epsilon_{13}=\frac{\phi_3}{\phi_2}\epsilon_{12},\quad \epsilon_{14}=\frac{\phi_4'}{\phi_2}\epsilon_{12},\ \text{etc.}\dots\dots,$$

or $$q_{13}-\frac{\phi_3}{\phi_2}q_{12}=r\left(\beta_2-\frac{\phi_3}{\phi_2}\sqrt{\beta_1}\right),$$

$$q_{14}-\frac{\phi_4'}{\phi_2}q_{12}=r\left(\frac{\beta_3}{\sqrt{\beta_1}}-\frac{\phi_4'}{\phi_2}\sqrt{\beta_1}\right),\ \text{etc.}\dots\dots$$

And lastly for cubical regression:

$$\epsilon_{14}-\frac{\phi_4'}{\phi_2}\epsilon_{12}-\frac{\phi_5}{\phi_2^2}\left(\epsilon_{13}-\frac{\phi_3}{\phi_2}\epsilon_{12}\right)\Big/\left(\phi_4-\frac{\phi_3^2}{\phi_2}\right)=0,\ \text{etc.}\dots\dots$$

Such conditions, especially with regard to their probable errors, become less and less manageable as we proceed.

The general principle involved in the present paper has been discussed by Tchebycheff*, and more adequately by J. P. Gram†, but the former had in view the fitting or graduating of curves. He calculated quantities which correspond to our μ_s s on the assumption that $n_x=1$, i.e. that the weight of the $\bar{y}_x$'s are all the same or that the marginal total is a *rectangle.* He was thinking of fitting a curve to a curve and not fitting a curve to a swarm of points. In his case each μ_s and accordingly each β and each ϕ is expressible in terms of the total number m of subranges which he takes of equal length. There are I think simpler methods of calculating the equation to a higher order parabola in such cases‡. As far as I am aware these orthogonal regression functions have not hitherto been dealt with and they throw a good deal of light on the original equations I provided in 1905 for skew regression. I had not recognised at that time that my expressions of each order were true orthogonal functions. It will be seen that my solution does not involve equality of subranges and is not limited to any special frequency distribution.

II. Note on the "Fundamental Problem of Practical Statistics."

(*Biometrika*, Vol. XIII, p. 1.)

Some misunderstanding has arisen with regard to my paper under the above title in the last issue of this *Journal.* I believe it to be due to the critics not having read Bayes' original theorem as given by Price in the *Phil. Trans.*, Vol. LIII. Bayes takes a ball and places it at random on a table, say of breadth unity, and its distance from one side being x, its chance of falling between x and $x+\delta x$ is δx. x is thus not a chance, but a *variate.* He now calls a "success," the chance that any other ball placed at random on the table will be nearer to the same side than the first

* *Mémoires de l'Académie de Saint-Pétersbourg.* Memoirs in 1854 and 1859. A *résumé* by R. Radau: *Bulletin Astronomique*, T. VIII, Paris, 1891, pp. 350, 376 *et seq.* See also Liouville's *Journal*, 2ᵉ Série, T. III (1858), p. 289 *et seq.*

† *Thesis*: "Om Rækkeudviklinger bestemte ved Hjaelp af de mindste Kradvaters Methöde." Kjøbenhavn, 1879.

‡ *Biometrika*, Vol. II, pp. 12—16.