A METHOD OF DETERMINING THE REGRESSION CURVE WHEN THE MARGINAL DISTRIBUTION IS OF THE NORMAL LOGARITHMIC TYPE

By Carl-Erik Quensel

Assistant at the Statistical Institute of the University of Lund, Sweden

In a paper[1] in this Journal Professor S. D. Wicksell gave the general outlines of a new method of calculating the regression lines. This problem was later on treated in detail by Dr. Walter Andersson.[2] His method was to develop the formulas for the regression lines into a series of orthogonal polynomials under the assumption that the marginal distribution of the independent variate belonged to certain mathematically defined distributions, and to determine the constants with the aid of the method of the least squares.

Among other cases he treated also the case where the marginal distribution was of the normal logarithmic type:

$$F(x) = \frac{\log e}{\sigma_l \sqrt{2\pi}\,(x-a)}\, e^{-\frac{1}{2}\left[\frac{\log (x-a) - l}{\sigma_l}\right]^2}. \tag{1}$$

But as his method is entirely different from the method I shall give here, I will not go any further into the method used by Dr. Andersson.

When the correlation surface $F(x, y)$ of the variates x and y is given and then of course also the marginal distribution of x, $F(x)$, it is known that the mean y_x of the dependent variate y in an infinitely small array with the value of x between x and $x + dx$ is given as a function of the independent variate x by the following formula (2)

$$y_x = \frac{\int yF(x, y)\, dy}{\int F(x, y)\, dy}. \tag{2}$$

In this formula the integrals are to be extended over the whole domain of the variation of y.

If now we make any transformation of x by introducing a new variate u, related to x by the formula $u = \psi(x)$, where we must suppose that u is a one-valued function of x and contrary, the distribution $f(u, y)$ of the variate u and y is given by the relation

$$f(u, y)\, du\, dy = F(x, y)\, dx\, dy \tag{3}$$

[1] S. D. Wicksell. Remarks on Regression. Annals of Mathematical Statistics, 1930.

[2] Walter Andersson. Researches into the theory of Regression. Meddelande från Lunds Astronomiska Observatorium. Ser II. N:r 64.

Writing the formula (2) in the following form:

$$y_x = \frac{\int yF(x, y)\,dx\,dy}{\int F(x, y)\,dx\,dy};$$

we see at once that the mean y_x can be given as the following function of u:

$$y_x = \frac{\int yf(u, y)\,dy}{\int f(u, y)\,dy}. \tag{4}$$

This relation, of course, is self-evident. The mean of the dependent variate in an array of the independent variate will be unchanged, when we change the variate x for another variate u, related to x by a one-valued function.

The problem of finding the regression line of the mean y_x can in such a way be much simplified, if it is possible to make a favorable transformation of the independent variate x.

As shown by Professor Wicksell[3] we may, under certain conditions concerning the marginal distribution $f(u)$, write the expression of the regression line in the following form:

$$y_x = \sum_0^\infty (-1)^n \frac{\lambda_{n,1}}{n!} \frac{f^{(n)}(u)}{f(u)}; \tag{5}$$

where the $\lambda_{n,1}$ coefficients are the seminvariants of the distribution of u and y.

The conditions which the function $f(u)$ must satisfy are among others that the function and all its derivates are continuous in the domain of variation and that the function and its derivates disappear in the limits of that domain. These conditions are satisfied by the normal curve of error.

In the case where the distribution of u is normal, the derivates $f^{(n)}(u)$ take the following form:

$$f^{(n)}(u) = (-1)^n H_n(u) f(u); \tag{6}$$

where the expressions $H_n(u)$ are the well known Hermitian polynomials.

The formula (5) takes the following simple form.

$$y_x = \sum_0^\infty \frac{\lambda_{n,1}}{n!} H_n(u) \tag{7}$$

If we can change the given marginal distribution $F(x)$ by a favorable substitution $u = \psi(x)$ into a normal curve, and if, this substitution made, we can

[3] S. D. Wicksell. Analytical Theory of Regression. Meddelande från Lunds Astronomiska Observatorium. Ser II. N:r 69.

calculate the coefficients $\lambda_{n,1}$ from the moments or other known characteristics of the given correlation distribution, $F(x, y)$, it is possible to express the regression line as the formula (8) shows:

$$y_x = \sum_0^\infty \frac{\lambda_{n,1}}{n!} H_n[\psi(x)] \tag{8}$$

It must be observed that the polynomials $H_n[\psi(x)]$ are orthogonal with regard to the distribution $F(x)$ of the independent variate x. We have

$$\int H_i[\psi(x)] H_j[\psi(x)] F(x)\, dx = \int H_i(u) H_j(u) f(u)\, du = 0 \qquad i \neq j$$

Not in all cases it will perhaps be possible to calculate the $\lambda_{n,1}$ coefficients, when we have transformed the marginal distribution into the normal curve, but in one case it is rather simple to calculate these coefficients from the moments given.

The case alluded to is the one, where the variate u is given from x by the relation $u = \log(x - a)$, that is that the marginal distribution is of the so called normal logarithmic type (1).

In that case it is possible to calculate the $\lambda_{n,1}$ coefficients from the marginal moments $V_{n,0}$ and from the correlation moments of the type $V_{n,1}$.

We suppose that the marginal distribution is of the logarithmic type and that from the moments of the x distribution we have determined the three constants a, σ_l and l in the usual manner.[4]

Then we calculate from the given correlation distribution the moments $V'_{n,0}$ about the point $x = a$ and the correlation moments $V'_{n,1}$ about the point $x = a$ and $y = m_y$ (the mean value of the y-variate).

From these moments it is possible to calculate the $\lambda_{n,1}$ coefficients in the following way.

The characteristic function of u and y is given by the following relation:

$$U(t_1 t_2) = e^{\sum \frac{\lambda_{kl}}{k!\,l!} t_1^k t_2^l} = \int\!\!\int e^{t_1 u + t_2 y} f(u, y)\, du\, dy \tag{9}$$

where the integrals are extended over the whole domain of variation.

If the distribution of u is according to the normal law, we have $\lambda_{k,0} = 0$ for $k \geqq 3$, but in the calculations here it is not at all necessary to suppose anything about these higher seminvariants. On the other side, the correlation distribution $f(u, y)$ is obtained from the characteristic function by the inversion theorem.

$$f(u, y) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{\sum \frac{\lambda_{kl}}{k!\,l!} (i\,w_1)^k (i\,w_2)^l} e^{-iw_1 u - iw_2 y}\, dw_1\, dw_2 \tag{10}$$

[4] How these are to be determined is shown in Pae- Tsi Yuan. On the logarithmic Frequency distribution and the Semi-Logarithmic Correlation Surface. Annals of Mathematical Statistics, 1933.

But we can also get the following relation

$$\int e^{t_1 u} f(u, y)\, du = \frac{1}{2\pi} \int e^{-iw_2 y} e^{\Sigma \frac{\lambda_{kl}}{k!\,l!} t_1^k (i w_2)^l} dw_2 \tag{11}$$

Of this last expression (11) between the characteristic function and the distribution function I will make use in the following.

The moments V'_{ij} of the distribution $F(x, y)$ about the point $x = a$, $y = m_y$ are given by the formula

$$V'_{ij} = \int\int (x - a)^i (y - m_y)^j F(x, y)\, dx\, dy. \tag{12}$$

If we write y instead of $y - m_y$ and instead of $x - a$ we write e^{bu} ($b = {}_{10}\log e$) the expression (12) takes the following form:

$$V'_{ij} = \int\int e^{ibu} y^j f(u, y)\, du\, dy \tag{13}$$

For the marginal moments of x about the point $x = a$ we get

$$V'_{n,0} = \int_a^\infty (x - a)^n F(x)\, dx = \int_{-\infty}^\infty e^{nbu} f(u)\, du \tag{14}$$

Comparing this formula (14) with the expression for the characteristic function of the distribution $f(u)$

$$U(t_1) = \int_{-\infty}^\infty e^{t_1 u} f(u)\, du = e^{\Sigma \frac{\lambda_{k,0}}{k!} t_1^k}; \tag{15}$$

we find the following simple relation

$$V'_{n,0} = e^{\Sigma \frac{\lambda_{k,0}}{k!} (n\,b)^k} \tag{16}$$

For the moments of the type $V'_{n,1}$ we get

$$V'_{n,1} = \int\int e^{nbu} y f(u, y)\, du\, dy = \int y dy \int e^{nbu} f(u, y)\, du. \tag{17}$$

If we compare the last integral in the formula (17) $\int e^{nbu} f(u, y) du$ with the formula (11) we see that we can write (17) as follows:

$$V'_{n,1} = \frac{1}{2\pi} \int y dy \int e^{-iw_2 y} e^{\Sigma \frac{\lambda_{kl}}{k!\,l!} (n\,b)^k (i\,w_2)} dw_2 \tag{18}$$

From the sum $\sum \frac{\lambda_{kl}}{k!l!} (nb)^k (iw_2)^l$ we may take out the part $\sum \frac{\lambda_{k,0}}{k!} (nb)^k$, where l is zero and which therefore does not contain any dignity of w_2, and write the remainder in the following form:

$$\sum \frac{\lambda_l'}{l!} (iw_2)^l$$

where we have

$$\lambda_1' = \lambda_{11} nb + \frac{\lambda_{21}}{2!} (nb)^2 + \frac{\lambda_{31}}{3!} (nb)^3 \dots$$

$$\frac{\lambda_2'}{2!} = \frac{\lambda_{02}}{2!} + \frac{3\lambda_{12}}{3!} nb + \frac{6\lambda_{22}}{4!} (nb)^2 \dots$$

The integral $\frac{1}{2\pi} \int e^{-iw_2 y} e^{\Sigma \frac{\lambda_l'}{l!} (i w_2)^l}$ may be considered as a frequency distribution $\varphi(y)$ with the seminvariants λ_l'.

The formula (18) will thus be written

$$V_{n,1}' = e^{\Sigma \frac{\lambda_{k0}}{k!} (n b)^k} \int ydy \, \varphi(y) \tag{19}$$

According to (16) we have

$$e^{\Sigma \frac{\lambda_{k0}}{k!} (n b)^k} = V_{n,0}'$$

and as

$$\int ydy \, \varphi(y) = \lambda_1' = \lambda_{11} nb + \frac{\lambda_{21}}{2!} (nb)^2 + \frac{\lambda_{31}}{3!} (nb)^3 \dots$$

we get

$$V_{n,1}' = V_{n,0}' \cdot \lambda_1' \tag{20}$$

or

$$\frac{V_{n,1}'}{V_{n,0}'} = \lambda_{11} nb + \frac{\lambda_{21}}{2!} (nb)^2 + \frac{\lambda_{31}}{3!} (nb)^3 \dots \tag{21}$$

We see that in the formulas for $V_{n,1}'$ we have all the seminvariants $\lambda_{n,1}$ involved. A successive determination of the seminvariants $\lambda_{n,1}$ with the aid of the moments of the same and lower degree is therefore not possible.

However, when we use the formula (8) for the regression, we must suppose that the seminvariants $\lambda_{n,1}$ with growing n converge rather soon towards zero. If the successive differences $\Delta^n \left(\frac{V_{n,1}'}{V_{n,0}'}\right)$ of the quotients $\frac{V_{n,1}'}{V_{n,0}'}$ are calculated, it may be possible to judge, how far it is possible to go with success. These differences will in most cases diminish rather soon and we shall therefore in most cases get a value of n about which we can suppose that the differences of higher order than this will all be so small that they can be neglected and as a consequence of this fact all higher seminvariants can be neglected too.

When this value of n has been determined, the n first seminvariants will all be obtained from the n first quotients $\frac{V'_{n,1}}{V'_{n,0}}$.

Thus we finally get the regression line as follows:

$$y_x = m_y + \sum_1^n \frac{\lambda_{i,1}}{i!} H_i[\log (x - a) - l]$$

or in standardized units:

$$y_x = m_y + \sum_1^n \frac{\lambda_{i,1}}{i!\, \sigma_l^i} H_i\left[\frac{\log (x - a) - l}{\sigma_l}\right]$$