NOTE ON REGRESSION FUNCTIONS IN THE CASE OF THREE SECOND ORDER RANDOM VARIABLES

By Clyde A. Bridger

The study of the correlation of two second-order random variables has received the attention of several authors, among them Yule [1], Charlier [2], Wicksell [3, 4], and Tschuprow [5]. Yule writes of them under the guise of "attributes." The study of three or more second order random variables has lagged behind. In this note we shall examine the regression function of one second order random variable on two others by considering the problem from the point of view of Tschuprow's [6] paper on the correlation of three random variables.

A variable X that takes on m values $x_1, \cdots, x_m$ with corresponding probabilities $p_1, \cdots, p_m$ subject to the condition $\sum_i p_i = 1$ is defined as a random variable of order m. (In particular, if X takes on only two values, x and x' with probabilities p and q, where $p + q = 1$, X is a random variable of second order.) The system of values x and probabilities p constitute the law of distribution of X. In the case of two random variables, X and Y, there exists a joint distribution law, covering all possible combinations of X and Y, together with their associated probabilities $p_{11}, \cdots, p_{mn}$ the joint distribution law contains all of the information regarding the stochastical dependence of X and Y.

The extension to more than two variables is immediate. Let p_{ijk} represent the probability of the simultaneous occurrence of the set of values x_i, y_j, z_k of three random variables X, Y, and Z; p_{ij}, that of the simultaneous occurrence of x_i, y_j together without reference to Z; p_i, that of the occurrence of x_i without reference to Y or Z; etc. Then, we have relationships of the types $\sum_i \sum_j \sum_k p_{ijk} = \sum_i \sum_j p_{ij} = \sum_i p_i = 1$; $\sum_i p_{ijk} = p_{jk}$; $\sum_i \sum_j p_{ijk} = \sum_j p_{jk} = \sum_i p_{ik} = p_k$. Similarly, let $p_{jk}^{(i)}$ be the probability of the simultaneous occurrence of y_j and z_k on the condition that X takes on the value x_i; $p_j^{(i)}$, that of the occurrence of y_j without reference to Z, on the same condition; etc. Then

$$\sum_k p_k^{(i)} = \sum_k p_k^{(ij)} = \sum_j \sum_k p_{jk}^{(i)} = 1; \qquad \sum_j p_{jk}^{(i)} = p_k^{(i)}; \qquad p_i p_j^{(i)} = p_{ij};$$

$$p_{ij} p_k^{(j)} = p_i p_{jk}^{(i)} = p_i p_j^{(i)} p_k^{(ij)} = p_{ijk}; \qquad \sum_i p_i p_j^{(i)} = p_j; \text{ etc.}$$

Denoting by $E(x)$ or simply Ex the expression "the mean value or mathematical expectation of x," we have $m_{fgh} = EX^f Y^g Z^h = \sum_i \sum_j \sum_k p_{ijk} x_i^f y_j^g z_k^h$. In particular, the mean values of the distributions are given by $m_X = EX$

$= \sum_i p_i x_i$, $m_Y = EY = \sum_j p_j y_j$, $m_Z = EZ = \sum_k p_k y_k$. Then we may write $\mu_{fgh} = E(X - m_X)^f (Y - m_Y)^g (Z - m_Z)^h = Eu^f v^g w^h = \sum_i \sum_j \sum_k p_{ijk}(x_i - m_X)^f (y_j - m_Y)^g (z_k - m_Z)^h$. The quantities μ may be identified as terms in the expression for the moments for the sum of three variables as follows: $E(u + v + w)^n = Eu^n + nEu^{n-1}v + \cdots + kEu^f v^g w^h + \cdots + Ew^n$, where $f + g + h = n$. If $n = 2$, we have the variance of the sum of three variables given by $\mu_{2..} + 2\mu_{11.} + \mu_{.2.} + 2\mu_{.11} + 2\mu_{1.1} + \mu_{..2}$, where the dots in the subscripts indicate variables not considered. Thus $\mu_{2..}$ refers to the second moment of the distribution of the variable X about its mean, m_X, without consideration of the distributions of Y or Z. If every term of the expansion of the n-th moment of the sum of three variables is divided by the quantity $\sqrt{\mu_{2..}^f \mu_{.2.}^g \mu_{..2}^h}$, the expansion takes the "normal form." The type term is $r_{fgh} = \mu_{fgh}/\sqrt{\mu_{2..}^f \mu_{.2.}^g \mu_{..2}^h}$. In the case of one variable, $r_f = \mu_f/\sqrt{\mu_2^f}$, so $r_1 = 0$, $r_2 = 1$, $r_3 = \sqrt{\beta_1}$, $r_4 = \beta_2$, etc. In the case of two variables, $r_{1.} = r_{.1} = 0$, $r_{2.} = r_{.2} = 1$, r_{11} = Pearson's product-moment coefficient of correlation, etc. Functions of parameters r will serve to characterize the law of correlation among the variables.

By writing the expressions with superscript (i) to denote that the values of the distributions of Y and Z are those which correspond to a fixed value x_i of the distribution of X, we have $m_Y^{(i)} = (EY)^{(i)}$, $m_Z^{(i)} = (EZ)^{(i)}$, $\mu_{gh}^{(i)} = E(Y - m_Y^{(i)})$ times $(Z - m_Z^{(i)})$, $r_{gh}^{(i)} = \mu_{gh}^{(i)}/\sqrt{\mu_{.2.}^{(i)g} \mu_{..2}^{(i)h}}$. (For $g = h = 1$, $r_{gh}^{(i)}$ becomes the conditional coefficient of correlation between Y and Z for $X = x_i$.) Thus it follows that we can study the correlation between Y and Z for each value of X separately.

For second order random variables, some changes in notation can be made. Let p_x and $p_{x'}$ be the probabilities corresponding to the values x and x', respectively, of X; p_y and $p_{y'}$ correspond to y and y', respectively; p_z and $p_{z'}$ correspond to z and z', respectively. Also, let p_{xy} represent the probability of the simultaneous occurrence of x and y together without reference to the distribution of Z, etc., and p_{xyz} represent the probability of the simultaneous occurrence of all three values, x, y, z, of their respective distributions, etc. Then, $p_x + p_{x'} = p_y + p_{y'} = p_z + p_{z'} = 1$; $p_{xy} + p_{xy'} = p_x$; $p_{xyz} + p_{xyz'} + p_{xy'z} + p_{xy'z'} = p_x$; etc.

Let us set up a system of normal coordinates in which the values U_i along the U-axis are defined by $U_i = \frac{x_i - m_X}{\sqrt{\mu_{2..}}}$, those along the V-axis by $V_j = \frac{y_j - m_Y}{\sqrt{\mu_{.2.}}}$, and those along the W-axis by $W_k = \frac{z_k - m_Z}{\sqrt{\mu_{..2}}}$. Let $m_Z^{(ij)}$ represent the mean of the set of values of the Z distribution which correspond to the fixed pair of values, (x_i, y_j), of the X and Y distributions. Then, in the new coordinate system, the same thing is given by $M_W^{(ij)} = \frac{m_Z^{(ij)} - m_Z}{\sqrt{\mu_{..2}}}$. Now, the series of values $M_W^{(ij)}$ obtained by giving i and j different values for the pair (U_i, V_j) determine what is called the regression function of W on U and V (or, in the

original notation, the surface of regression of the distribution of Z on the distributions of X and Y). Similarly, the values of $[M_W^{(ij)}]^{(i)} = \dfrac{m_Z^{(ij)} - m_Z^{(i)}}{\sqrt{\mu_{..2}}}$ obtained by fixing U and varying V in the set (U_i, V_j) determine what is called the conditional line of regression of W on V for a fixed value of U. With these definitions we shall consider the problem of finding a regression function of W on U and V for three second order random variables.

For convenience, write $\delta_{xy} = p_{xy} - p_x p_y$, $\delta_{xz} = p_{xz} - p_x p_z$, $\delta_{yz} = p_{yz} - p_y p_z$, $\alpha_{yz} = p_x p_{xyz} - p_{xy} - p_{xz}$, $\epsilon_z = p_{xyz} - p_{xy} p_z$, $\beta_{yz} = p_{x'} p_{x'yz} - p_{x'y} p_{x'z}$, $\theta_{xyz} = \epsilon_z - p_y \delta_{xz} - p_x \delta_{yz} = \epsilon_y - p_x \delta_{yz} - p_z \delta_{xy} = \epsilon_x - p_y \delta_{xz} - p_z \delta_{xy}$. Direct substitutions into the several formulas developed above then gives us the representative forms to be used in subsequent calculations:

$$x - m_X = p_{x'}(x - x'), \quad x' - m_X = -p_x(x - x').$$

$$m_X = p_x x + p_{x'} x', \quad r_{1..} = 0, \quad r_{2..} = 1, \quad r_{3..} = \frac{p_{x'} - p_x}{\sqrt{p_x p_{x'}}},$$

$$r_{4..} = \frac{1}{p_x p_{x'}} - 3, \quad r_{11.} = \frac{\delta_{xy}}{\sqrt{p_x p_{x'} p_y p_{y'}}}, \quad r_{21.} = r_{11.} r_{3..},$$

$$r_{12.} = r_{11.} r_{.3.}, \quad r_{13.} = r_{11.} r_{.4.}, \quad r_{22.} = r_{11.} r_{3..} r_{.3.} + 1,$$

$$r_{12.} r_{21.} = r_{11.}(r_{22.} - 1), \quad r_{211} = r_{3..} r_{111} + r_{.11},$$

$$r_{121} = r_{.3.} r_{111} + r_{1.1}, \quad r_{112} = r_{..3} r_{111} + r_{11.},$$

$$r_{111} = \frac{\theta_{xyz}}{\sqrt{p_x p_{x'} p_y p_{y'} p_z p_{z'}}}, \quad U_1 = \frac{p_{x'}}{\sqrt{p_x p_{x'}}}, \quad U_2 = \frac{-p_x}{\sqrt{p_x p_{x'}}},$$

$$M_W^{(11.)} = \frac{\epsilon_z}{p_{xy}\sqrt{p_z p_{z'}}}, \quad M_W^{(12.)} = \frac{\delta_{xz} - \epsilon_z}{p_{xy'}\sqrt{p_z p_{z'}}},$$

$$M_W^{(21.)} = \frac{\delta_{yz} - \epsilon_z}{p_{x'y}\sqrt{p_z p_{z'}}}, \quad M_W^{(22.)} = \frac{\epsilon_z - \delta_{xz} - \delta_{yz}}{p_{x'y'}\sqrt{p_z p_{z'}}},$$

$$[M_W^{(11.)}]^{(1..)} = \frac{\alpha_{yz}}{p_{xy}\sqrt{p_{xz} p_{xz'}}}, \quad [M_W^{(21.)}]^{(2..)} = \frac{\beta_{yz}}{p_{x'y}\sqrt{p_{x'z} p_{x'z'}}},$$

$$[M_W^{(12.)}]^{(1..)} = \frac{-\alpha_{yz}}{p_{xy'}\sqrt{p_{xz} p_{xz'}}}, \quad [M_W^{(22.)}]^{(2..)} = \frac{-\beta_{yz}}{p_{x'y'}\sqrt{p_{x'z} p_{x'z'}}}.$$

In the case of correlation of two second order random variables, a linear regression function can always be found [3, 5]. Similarly, the conditional regression functions in the case of three second order random variables can always be taken as linear. If we take as the form of the regression function of W on U and V the form $M_W^{(ij)} = aU_i + bV_j + cU_iV_j + d$, where a, b, c, d are constants to be determined by direct substitution for U_i and V_j from the distributions of X and Y, it is seen that linearity of all total and conditional

regression functions is preserved. By total regression function, we mean the regression of W on U or W on V.

Now consider the problem of finding a, b, c, d. The direct substitution provides us with four linearly dependent equations in four unknowns. Linear combinations reduce the set to three, from which the relationship $d = -cr_{11}.$ is obtained. By building up the various terms in the equations through dividing by the necessary values of p, the parameters r can be made to appear. Further combinations now reduce the set to the following three:

$$r_{111} = ar_{21}. + br_{12}. + c(r_{22}. - r_{11}^2.)$$

$$r_{.11} = ar_{11}. + b \quad + cr_{12}.$$

$$r_{1.1} = a \quad + br_{11}. + cr_{21}.$$

The solution gives

$$a = \frac{r_{1.1} - r_{11}.r_{.11}}{1 - r_{11}.^2} - \frac{r_{21}. - r_{11}.r_{12}.}{1 - r_{11}.^2} c = a' - a''c$$

$$b = \frac{r_{.11} - r_{11}.r_{1.1}}{1 - r_{11}.^2} - \frac{r_{12}. - r_{11}.r_{21}.}{1 - r_{11}.^2} c = b' - b''c$$

$$c = (1 - r_{11}.^2)(r_{111} - a'r_{12}. - b'r_{21}.) \div \Delta, \qquad \text{where}$$

$$\Delta = \begin{vmatrix} 1 & r_{11}. & r_{21}. \\ r_{11}. & 1 & r_{12}. \\ r_{21}. & r_{12}. & r_{22}. - r_{11}.^2 \end{vmatrix}$$

The regression function becomes
$M_W^{(ij)} = a'U_i + b'V_j - c(r_{111} + a''U_i + b''V_j - U_iV_j)$. If $c = 0$ the surface is a plane. Examination of the characteristics of r_{111} shows that generally c cannot be zero. The vanishing of c implies that special relations must exist between p_{ijk} and p_{ij} , p_{ik} , p_{jk} .

Two constants of considerable importance in the theory of correlation are the multiple correlation coefficient and the multiple correlation ratio. For the regression of W on U and V, the former is defined as $R_{11}.^2 = a'r_{1.1} + b'r_{.11}$ and the latter as $\eta_{--2} = \sum_i \sum_j p_{ij}[M_W^{(ij)}]^2$. For planar regression, the difference $\eta_{--2} - R_{11}.^2$ must vanish. For others, the difference takes on values characteristic of the regression function. To find the value it takes for our case, we set up the value of η_{--2} from the regression function just given and subtract $R_{11}.^2$.

By direct substitution, we have $\eta_{--2} - R_{11}.^2 = \sum_i \sum_j p_{ij}(aU_i + bV_j - cU_iV_j + cr_{111})^2 - a'r_{1.1} - b'r_{.11}$. Since $\sum_i \sum_j p_{ij}U_i^2 = 1$, $\sum_i \sum_j p_{ij}(U_iV_j)^2 = r_{22}.$, etc., we find rather easily that

$$\eta_{--2} - R_{11}.^2 = c^2(r_{22}. - r_{11}.^2) - a''r_{21}. - b''r_{12}..$$

We can also obtain the same value of $\eta_{--2} - R_{11.}^2$ by direct substitution for the four values of $M_W^{(ij)}$ in η_{--2} and subtracting $R_{11.}^2$. To actually obtain this is a long laborious process complicated by the fact that so many alternate forms for the answer are possible, of which only one is comparable with the value previously found. The general procedure is first to set up from the definition the expression $K = p_z p_{z'} \eta_{--2} =$

$$p_{xy}\left(\frac{\epsilon_z}{p_{xy}}\right)^2 + p_{xy'}\left(\frac{\delta_{xz} - \epsilon_z}{p_{xy'}}\right)^2 + p_{x'y}\left(\frac{\delta_{yz} - \epsilon_z}{p_{x'y}}\right)^2 + p_{x'y'}\left(\frac{\epsilon_z - \delta_{xz} - \delta_{yz}}{p_{x'y'}}\right)^2.$$

Then we build up each square by addition and subtraction so that it will contain a θ_{xyz} term. At the close of the process, we convert the whole expression into the parameters r by dividing through by $p_z p_{z'}(p_x p_{x'} p_y p_{y'})^2$ and substituting from the list of representative forms given at the beginning of the paper. A matter of rearrangement now gives the same result as before.

From the symmetry involved, we can say that, in the case of the correlation of three second order random variables, the function representing the regression of one on the other two has an equation in normal coordinates of the form $M_W^{(ij)} = aU_i + bV_j - cU_iV_j - cr_{111}$, where a, b, and c satisfy equations of type

$$\begin{aligned} r_{111} &= ar_{21.} + br_{12.} + c(r_{22.} - r_{11.}^2) \\ r_{.11} &= ar_{11.} + b \quad + cr_{12.} \\ r_{1.1} &= a \quad + br_{11.} + cr_{21.} \end{aligned}$$

STATE DIVISION OF PUBLIC HEALTH,
BOISE, IDAHO.

BIBLIOGRAPHY

[1] G. UDNY YULE. *An Introduction to the Theory of Statistics.* London: Charles Griffin & Co., Ltd., 1922. 6th Ed.

[2] C. V. L. CHARLIER. "Om korrelation mellan eganskaper inom den homograda statistiken." *Svenska Aktuarieföreningens Tidskrift.* Vol. I (1914), pp. 21–35.

[3] S. D. WICKSELL. "Some theorems in the theory of probability, with special reference to their importance in the theory of homograde correlation." *Svenska Aktuarieföreningens Tidskrift.* Vol. III (1916), pp. 165–213.

[4] S. D. WICKSELL. "On the correlation of acting probabilities." *Skandinavisk Aktuarietidskrift.* Vol. I (1918), pp. 98–135.

[5] A. A. TSCHUPROW. *Grundbegriffe und Grundprobleme der Korrelationstheorie.* Leipzig: B. G. Teubner, 1925.

[6] A. A. TSCHUPROW. (Translation into English by L. Isserlis.) "The Mathematical Theory of the Statistical Methods Employed in the Study of Correlation in the Case of Three Variables." *Transactions of the Cambridge Philosophical Society.* Vol. XXIII, no. 12 (1928), pp. 337–382.