# THE DETERMINATION OF WEIGHTS FOR A NON-QUANTITATIVE INDEPENDENT VARIABLE USED IN LINEAR REGRESSION

by

ESTHER JUNE ALKIRE

A THESIS

Presented to the Department of Mathematics
and the Graduate Division of the University of Oregon
in partial fulfillment
of the requirements for the degree of
Master of Arts

June 1945

APPROVED:

(Adviser for the Thesis)

(For the Committee)

The author acknowledges with gratitude her indebtedness to Dr. Carl F. Kossack of the Mathematics Department of the University of Oregon, for his kindly guidance and inspirational direction throughout this work. She also wishes to express her appreciation to Dr. Albert Poole for his helpful suggestions.

## TABLE OF CONTENTS

## LIST OF TABLES

# INTRODUCTION

In making use of the theory of linear regression to obtain an estimation of a dependent variate from the information contained in an independent variate, one frequently is faced with the problem of having the independent variable given in a non-quantitative manner. In these cases the independent variable usually is classified into ordered groups. In order to use the theory of regression one must assign a numerical weight to each of these groups. It is the purpose of this paper to consider the problem of determining these weights.

The data considered in this paper will be assumed to be bivariate with the dependent variable quantitatively measured and the independent variable classified into ordered groups. Numerical weights are to be determined such that the regression equation thereupon obtained will give the best estimate of the dependent variable.

At present the usual practice is to obtain these weights in a more or less subjective manner. In considering this problem Frank A. Pearson[1] states that since there is no numerical value given to the classes, ". . . . a unit rate of change cannot be calculated for a relationship in which the independent variable is non-numerical", while Ezekial[2], in the

[1] Frank A. Pearson and Kenneth R. Bennett, Statistical Methods, (New York: John Wiley and Sons, 1942), p. 135.

[2] Mordecai Ezekial, Methods of Correlation Analysis, (New York: John Wiley and Sons, 1941), p. 310.

recent edition of his book, makes the following statement.

> In case a non-quantitative factor is a very important one, so that ignoring it in determining the net linear regressions may seriously impair their accuracy, it may be roughly included by designating successive groups by a numerical code which approximates the expected influence of the variable.

The literature contains very little in the way of a direct reference to this problem as it arises in connection with regression, however, one finds in the literature many references to the problem of estimating the correlation coefficient from qualitative data. Among the references available, an assignment of weights may be made incidental to the estimation of the correlation coefficient. For this reason, and since the problem of correlation is so closely related to that of regression, the principle methods of determining the correlation coefficient for non-quantitative data will be given before the actual problem of this paper is considered.

Following the discussion of these methods, a method of assigning weights to the ordered classes of an independent variable, which is based on minimizing the standard error of estimate, is developed. This will be followed by a numerical example to illustrate this method for determination of weights and the result obtained compared with those obtained for several other choices of weights. A discussion of other definitions of what might be considered as the best estimate of the dependent variable together with a few summarizing remarks will conclude the paper.

# CHAPTER I

## LINEAR REGRESSION AND CORRELATION

### For a Sample Population

Consider a series of observations $x_i$ and $y_i$ that are linearly connected. These observations can be plotted on a graph to form what is called the scattered diagram. By fitting a straight line to the scattered diagram in such a way as to make the sum of the squares of the ordinate distances from the points to the line a minimum, one obtains the regression line of y on x. The regression line or the best fitting straight line thus obtained is the best estimate, in the least square sense, of the relation of the values of the dependent variable, y, to the values of the independent variable, x.

Let $y_c = mx + b$ be the line of regression of y on x. Representing the difference between the ordinate of any given point and the corresponding ordinate of the line by $e_i$, that is, $e_i = y_i - y_c = y_i - mx_i - b$. These differences are called residual errors.

Now m and b are chosen subject to the condition that the $\sum_{i=1}^{n} e_i^2$[1] is to be a minimum. Using the calculus to minimize

$$\Sigma e^2 = \Sigma(y - mx - b)^2,$$

[1] In the rest of the paper $\Sigma$ sign will be used in place of $\sum_{i=1}^{n}$.

one obtains upon differentiating with respect to m and b respectively, and setting the derivatives equal to zero, the two normal equations:

$$\Sigma y = m\Sigma x + nb$$

$$\Sigma xy = m\Sigma x^2 + b\Sigma x.$$

Solving the normal equations simultaneously, the following values for m and b are obtained:

$$m = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

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

Hence the regression line of y on x may be written as

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

If by chance all of the plotted points should fall on the regression line, the estimate of the dependent variable would be perfect. In general the points will not fall on the line. Consequently the goodness of estimate is judged in terms of the standard deviation of the residual errors. This standard deviation is known as the standard error of estimate and will be denoted by S with a subscript to indicate the variable whose deviations are being measured. It differs from an ordinary standard deviation of a single variable only in that deviations are measured from the regression line instead of the mean or arithmetic average.

At this point in the discussion we are interested in $S_y$. In symbols the computation is

$$S_y^2 = \frac{\Sigma(y - y_c)^2}{n}.$$

Substituting mx + b for $y_c$, the equation becomes

$$S_y^2 = \frac{\Sigma(y - mx - b)^2}{n} .$$

Upon further simplification and since $\Sigma\varepsilon = 0$, the equation becomes

$$S_y^2 = \frac{\Sigma y^2 - m\Sigma y - b\Sigma xy}{n} . \qquad (2)$$

The standard error of estimate is a measure of the scatter of the points from the regression line. The closer the points lie to the line, the smaller will be the value of $S_y$ and vice versa. $S_y$ is to be interpreted in the same way as any other standard deviation. It gives the range on either side of the regression line within which about 68 percent of the points can be expected to fall, provided the distribution of both variables are approximately normal.

The variability of the y-variable which can be explained by the linear association of y with x is determined from the equation of the regression line. $S_y$ is a measure of the remaining part of the variability of y that is not explained by the regression line.

Some standard is necessary in order to determine what constitutes a large or a small value of $S_y$. The largest value that $S_y$ can take is $\sigma_y$. That is if x and y were completoly independent, the regression line of y on x would be $y = \Sigma y/n$. This means that the best estimate of y for any value of x is the mean of the y distribution, hence all the values of $y_c$ would be equal. We also know that when there is a perfect relationship $S_y = 0$, and we therefore can use $\sigma_y$ as the standard for judging whether the values of $S_y$ are large or small.

In order to obtain the relationship existing between the correlation coefficient and the standard error of estimate let us replace the values of m and b in equation (2) by the values of m and b used in the regression equation (1).

$$S_y^2 = \frac{\frac{n\Sigma y^2}{n} - (\frac{\Sigma y - b\Sigma x}{n})\Sigma y - b\Sigma xy}{n}$$

$$= \frac{1}{n^2}\left[n\Sigma y^2 - (\Sigma y)^2 - (\frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2})(n\Sigma xy - \Sigma x\Sigma y)\right]$$

$$= (\frac{n\Sigma y^2 - (\Sigma y)^2}{n^2})\ (1 - \frac{(n\Sigma xy - \Sigma x\Sigma y)^2}{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]})$$

$$= \sigma_y^2(1 - r^2) \qquad (3)$$

where r denotes the correlation coefficient between the variables x and y and takes the sign of the slope of the regression line.

Therefore one can see from equation (3) that the fundamental relationship between the standard error and the correlation coefficient is such that if the two variables are unrelated, $S_y$ and $\sigma_y$ are identical and the value of r is zero. If the two variables are perfectly related, that is, all points falling on the regression line, the value of $S_y$ is zero and the value of r would be plus or minus one. The value of the correlation coefficient may therefore range from zero to plus or minus one. These limits represent perfect correlation (direct or inverse), and complete absence of correlation.

## For a Parent Population

Let us extend the discussion of the relation between regression and correlation of a sample to that of a parent population.[1] The parent population may be represented by the continuous variables x and y. We assume the variables x and y have the joint probability function f(x,y), where the double integral of f(x,y) over a region of the xy plane measures the relative frequency of occurrence of the pairs of values of x and y in the region. Hence we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\, dy\, dx = 1,$$

where f(x,y) dy dx is the probability that simultaneously x lies in the interval (x, x + dx) and y lies in the interval (y, y + dy).

The probability that x occurs in the interval (x, x + dx) for all y's, will be denoted by g(x) dx. Then integrating over all admissible values of y, we have

$$g(x)\, dx = dx \int_{-\infty}^{\infty} f(x,y)\, dy.$$

Similarly, if h(y) dy is the probability that y occurs in the interval (y, y + dy) for all assignments of x, we have

$$h(y)\, dy = dy \int_{-\infty}^{\infty} f(x,y)\, dx.$$

In accordance with convention we shall call g(x) and h(y) the marginal distributions.

The general product moment about the common origin of x and y may be defined as follows:

$$\nu_{mn} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\, x^m y^n\, dy\, dx.$$

[1] John F. Kenney, *Mathematics of Statistics*, (New York: D. Van Nostrand Company, 1939), pp. 63-70.

If m = 0 and n = 1, we have

$$\nu_{01} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ y\ dy\ dx.$$

Let f(x,y) be a function in which the order of integration may be interchanged. The $\nu_{01}$ becomes

$$\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f(x,y)\ dx\right] y\ dy = \int_{-\infty}^{\infty} h(y)y\ dy,$$

which is the mean, $\bar{y}$, of the y's. Similarly the mean of the x's is

$$\nu_{10} = \bar{x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)x\ dy\ dx = \int_{-\infty}^{\infty} g(x)x\ dx.$$

Now defining the general product moment about the means $(\bar{x},\bar{y})$ as follows:

$$\mu_{mn} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \bar{x})^m(y - \bar{y})^n f(x,y)\ dy\ dx.$$

When m = n = 1, we have

$$\mu_{11} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \bar{x})(y - \bar{y})f(x,y)\ dy\ dx,$$

which is called the co-variance of the joint distribution.

When m = 2 and n = 0, we have the variance of x,

$$\mu_{20} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \bar{x})^2 f(x,y)\ dy\ dx$$

$$= \sigma_x^2.$$

Similarly, when m = 0 and n = 2, we have the variance of y,

$$\mu_{02} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y - \bar{y})^2 f(x,y)\ dy\ dx$$

$$= \sigma_y^2.$$

Having defined the moments of the distribution function f(x,y), we are now ready to consider the regression curve.

If y has been assigned in the joint probability function f(x,y), the probability that x will lie in an infinitesimal interval is

$$\frac{f(x,y)}{h(y)}\,dx.$$

Thus, when y is fixed,

$$\int_{-\infty}^{\infty} \frac{f(x,y)}{h(y)}\,dx = 1$$

and so $f(x,y)/h(y)$ is the probability function of x for a fixed y. It may be called the probability density representing a y array of x's.

Likewise if we fix x, the probability density for an x array of y's is given by $f(x,y)/g(x)$, and

$$\int_{-\infty}^{\infty} \frac{f(x,y)}{g(x)}\,dy = 1 \qquad (4)$$

when x is fixed.

The mean of an x array of y's is

$$\bar{y}_x = \int_{-\infty}^{\infty} \frac{y\,f(x,y)}{g(x)}\,dy \qquad (5)$$

where the integration is performed over all values in the array defined by x. Similarly, the mean of a y array of x's is

$$\bar{x}_y = \int_{-\infty}^{\infty} \frac{x\,f(x,y)}{h(y)}\,dx \qquad (6)$$

integrated over all x's in an array for a fixed y.

The variance in an x array of y's is given by

$$\int_{-\infty}^{\infty} (y - \bar{y}_x)^2\,\frac{f(x,y)}{g(x)}\,dy \qquad (7)$$

integrated over all values in the array fixed by x. Similarly the variance in a y array of x's is

$$\int_{-\infty}^{\infty} (x - \bar{x}_y)^2 \frac{f(x,y)}{h(y)}\, dx. \tag{8}$$

Taking different x arrays of y's fixes the mean points $\bar{y}_x$ and as x varies continuously we get the locus of these means which is called the regression curve of y on x. Its equation is given by (5). Similarly (6) gives the regression curve of x on y.

We shall consider only the case where the regression curves are straight lines. If the equation of the regression curve of y on x is of the form

$$\bar{y}_x = mx + b$$

then the regression of y on x is said to be linear.

Consider

$$\bar{y}_x = \int_{-\infty}^{\infty} \frac{y\, f(x,y)}{g(x)}\, dy = mx + b$$

or
$$\int_{-\infty}^{\infty} y\, f(x,y)\, dy = mx\, g(x) + b\, g(x). \tag{9}$$

Integrating each side and remembering that we may change the order of integration, we obtain

$$\nu'_{01} = m\, \nu'_{10} + b. \tag{10}$$

Multiplying each side of equation (9) by x and integrating with respect to x, we have

$$\nu'_{11} = m\, \nu'_{20} + b\, \nu'_{10}. \tag{11}$$

A simultaneous solution of (10) and (11) yields

$$m = \frac{\mu_{11}}{\mu_{20}} = \frac{\mu_{11}}{\sigma_x^2}$$

$$b = \nu_{01} - \nu_{10}\frac{\mu_{11}}{\sigma_x^2} = \bar{y} - \bar{x}\frac{\mu_{11}}{\sigma_x^2}.$$

Therefore the equation of regression of y on x becomes

$$\bar{y}_x - \bar{y} = \frac{\mu_{11}}{\sigma_x^2}(x - \bar{x}). \qquad (12)$$

We shall now consider the standard error of estimate. We have seen that the probability density in an x array of y's is $f(x,y)/g(x)$, and the variance $s_{y.x}^2$ within such an array is given by (7).

The mean, over all x arrays, of values of $s_{y.x}^2$ weighted with the marginal distribution of x is denoted by $\not{s}_y^2$, and $\not{s}_y$ is called the standard error of estimate. We will now show that $\not{s}_y^2 = \sigma_y^2(1 - \rho^2)$. By definition,

$$\not{s}_y^2 = \int_{-\infty}^{\infty} g(x)s_{y.x}^2\, dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y - \bar{y}_x)^2 f(x,y)\, dy\, dx.$$

Using the value of $\bar{y}_x$ given in (12) the above expression becomes

$$\not{s}_y^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[y - \bar{y} - \frac{\mu_{11}}{\sigma_x^2}(x - \bar{x})\right]^2 f(x,y)\, dy\, dx$$

$$= \left[(y - \bar{y})^2 - 2\frac{\mu_{11}}{\sigma_x^2}(y - \bar{y})(x - \bar{x}) + \frac{\mu_{11}^2}{\sigma_x^4}(x - \bar{x})^2\right] f(x,y)\, dy\, dx.$$

The right member simplifies so that we have the result

$$\not{s}_y^2 = \sigma_y^2\left(1 - \frac{\mu_{11}^2}{\sigma_x^2\sigma_y^2}\right)$$

$$= \sigma_y^2(1 - \rho^2)$$

where $\rho$ is the correlation coefficient of the parent population.

This result is analogous to that obtained from the sample problem.

and thus we see that the relationship between regression and correlation is the same in sample and parent population.

# CHAPTER II

## THE PROBLEM OF ESTIMATING THE CORRELATION COEFFICIENT FOR NON-QUANTITATIVE DATA

### Tetrachoric Correlation

One of the first references dealing with the correlation of non measurable characters is given by Pearson in 1900.[1]

Pearson considered the case of N objects which are classified into a fourfold table as they possess one, both, or neither, of two qualitative traits or characters, which may, for convenience, be denoted by I and II. Such a classification may be represented in the following fourfold table:

Table I

| | Not II | II | Total |
|---|---|---|---|
| Not I | a | b | a + b |
| I | c | d | c + d |
| Total | a + c | b + d | N |

where N = a + b + c + d.

To measure the intensity of association between two characters in

[1] Karl Pearson, "Mathematical Contributions to the Theory of Evolution," Philosophical Transactions, Vol. 195, A, (Feb. 1900), pp. 1-6.

such a classification, it is supposed that the data can be represented by a normal correlation surface containing r, the correlation coefficient, as the parameter. The problem is to determine r so that the surface can be divided into four cells by two planes intersecting at right angles, to yield the relative frequencies observed. Then the correlation coefficient for this normal surface is called the tetrachoric r.

Let the frequency surface

$$z = \frac{N}{2\pi\,(1 - r^2)^{1/2}}\, e^{-\frac{1}{2}\frac{1}{1 - r^2}(x^2 + y^2 - 2rxy)},$$

where the variates are measured in standard deviation units, be divided into four cells by two planes x = h and y = k. The total volumes or frequencies in these parts will be represented by a, b, c, and d, in the manner indicated in Table I.

The value of h and k can readily be found since $\frac{c + d}{N}$ is the area under the normal curve between k and ∞, and $\frac{b + d}{N}$ is the area under the normal curve between h and ∞.

We see that

$$d = \frac{N}{2\pi(1 - r^2)^{1/2}} \int_h^\infty \int_k^\infty e^{-\frac{1}{2}\frac{1}{1 - r^2}(x^2 + y^2 - 2rxy)}\, dx\, dy. \quad (1)$$

From this equation the value of r is found since d, N, h, and k are known.

The solution of equation (1) is obtained by expanding the equation in terms of r by Maclaurin's theorem. After taking logarithmic differentials and differentiating n times by Leibnitz's theorem we may integrate from h to ∞ with respect to x, and from k to ∞ with respect to y, obtaining, after some reductions, the following equation:

$$\frac{ad - bc}{N^2HK} = r + \frac{r^2}{2}hk + \frac{r^3}{6}(h^2 - 1)(k^2 - 1) + \frac{r^4}{24}h(h^2 - 3)k(k^2 - 3)$$

$$+ \frac{r^5}{120}(h^4 - 6h^2 + 3)(k^4 - 6k^2 + 3)$$

$$+ \frac{r^6}{720}h(h^4 - 10h^2 + 15)k(k^4 - 10k^2 + 15) + \ldots .$$

where

$$H = \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}h^2} \quad \text{and} \quad K = \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}k^2}.$$

The numerical solution has to be obtained by approximating to the roots, and Newton's method[1] is convenient for this purpose.

To facilitate the arithmetical work of this method there are tables[2] available. These are arranged so that the equation

$$d/N = \Gamma_o\Gamma'_o + \Gamma_1\Gamma'_1 r + \Gamma_2\Gamma'_2 r^2 + \ldots .$$

can be used. $\Gamma_n$ are known to be tetrachoric functions[3] and are tabulated up to $\Gamma_6$. Further values may be obtained by a difference formula. Much work can be avoided since all that has to be done, if these tables are available, is to calculate h, k, and the ratio d/N, then interpolate in the tables so as to obtain r.

## Polychoric Correlation

[1] W. Palin Elderton, Frequency Curves and Correlation 3rd edition, (London: Cambridge University Press, 1938), pp. 175.

[2] Karl Pearson, Tables for Statisticians and Biometricians, (London: Cambridge University Press, 1914), pp. 42-57.

[3] See page (18) for further details on tetrachoric functions.

The more general problem, that of determining r when objects are classified into an m by n fold table was also considered by Pearson.[1] The procedure he followed is similar to that used when considering the specialized 2 by 2 classification, in as much as, a normal correlation surface is fitted to the table and equations for the correlation coefficient are derived employing the tetrachoric functions and the observed cell frequencies. From these equations the value of r is obtained.

Again it is supposed that there exists a normal correlation surface with a fixed r which when cut up into the m by n cells will contain in the several cells precisely the relative frequencies given. The r that pretains to such a surface is the coefficient of correlation sought. This coefficient is commonly known as the polychoric coefficient of correlation. In the cases encountered in applications, there is usually no one value of r which determines a surface whose theoretical frequencies will exactly equal all of the observed frequencies, since the observed frequencies contain sampling variations. The problem is, therefore, to determine the surface that satisfies the conditions as nearly as possible.

Before going into the details of this method of finding the polychoric coefficient of correlation, a brief discussion pretaining to the notation to be used will be given.

We start with the assumption that both the horizontal and vertical marginal totals of the polychoric table can be represented on a normal scale. Now the polychoric table is such that in the population N under

[1] Karl Pearson and Egon Pearson, "On Polychoric Coefficients of Correlation", Biometrika, XIV, pp. 127-156.

discussion, the sth category of the first variate, x, contains $n_{s.}$ individuals and the rth category of the second variate, y, contains $n_{.r}$ individuals, while the number of individuals who combine in the population N the sth category of x and the rth category of y will be denoted by $n_{sr}$.

Table II

| x<br>y | 1 | 2 | 3 |  | s |  |  |  |
|---|---|---|---|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ |  |  |  |  | $n_{.1}$ | $-\infty$ |
| 2 | $n_{21}$ |  |  |  |  |  | $n_{.2}$ | $k_1$ |
|  |  |  |  |  |  |  |  | $k_2$ |
| r |  |  |  |  | $n_{sr}$ |  | $n_{.r}$ | $k_{r-1}$ |
|  |  |  |  |  |  |  |  | $k_r$ |
|  | $n_{1.}$ | $n_{2.}$ |  |  | $n_{s.}$ |  | N |  |
| $-\infty$ | $h_1$ | $h_2$ |  | $h_{s-1}$ | $h_s$ |  | $\infty$ | $\infty$ |

If $n_{1.}, n_{2.}, \ldots n_{s.}, \ldots$ be the frequencies of the x variate for the several categories, the values of the ratios of abscissae to standard deviation, or end x's, will be specified as $-\infty, h_1, h_2, \ldots h_s, \ldots$ Here $h_{s-1}$ and $h_s$ are the values on either side of the category $n_{s.}$. Similarly, if the frequencies of the various categories of the y variate be $n_{.1}, n_{.2}, \ldots, n_{.r}$, the value of the ratios of ordinates to standard deviation will be represented by $-\infty, k_1, k_2, \ldots, k_r, \ldots,$ where $k_{r-1}$ and $k_r$ give the end y's on either side of $n_{.r}$.

The means of the categories $n_{.s}$ and $n_{r.}$ will be denoted by $\bar{h}_s$ and $\bar{k}_r$, whereas $\bar{h}_{sr}$ and $\bar{k}_{sr}$ will be the x and y variate means for the sth-rth cell, and $n_{sr}$ the product moment of the frequency in the sth-rth cell.

To find the mean points we make use of the following property[1]: The mean value of that portion of the area under the curve which lies over the interval, $h_{s-1}$ to $h_s$, is found by subtracting the ordinate at $h_s$ from the ordinate at $h_{s-1}$, and dividing the result by the area. Thus the means of the categories $n_{.s}$ and $n_{r.}$ are determined by

$$\overline{h}_s = \frac{H_{s-1} - H_s}{n_{s.}/N} \quad \text{and } \overline{k}_r = \frac{K_{s-1} - K_s}{n_{.r}/N}$$

where

$$H_s = \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}h^2} \quad \text{and} \quad K_r = \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}k^2} .$$

The theoretical cell frequency will be denoted by $\overline{n}_{sr}$, whereas the observed cell frequency is $n_{sr}$. We shall write the real coefficient of correlation of the population as $r$, the coefficient as from a single sth-rth cell as $r_{sr}$, and those from the $n_{s.}$ and $n_{.r}$ array as $r_{s.}$ and $r_{.r}$ respectively. This notation will be used throughout the methods that follow unless otherwise specified.

In the development of the polychoric $r$, Pearson makes use of the tetrachoric functions. The tetrachoric function of the order $t$ is defined

as $$\tau_t = \frac{1}{(t!)^{1/2}} (-d/dx)^{t-1} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}x^2} .$$

Certain properties and notation concerning these functions are useful in the development. We shall write for brevity

$$D_s\tau_t = \tau_t(h_s) - \tau_t(h_{s-1})$$

[1] Burton H. Camp, Elementary Statistics, (Boston: D. C. Heath And Company, 1931), p. 68.

where $\Gamma_t(h)$ denotes the value of the tetrachoric function for $x = h$. The formula for obtaining the successive tetrachoric functions for a given x is

$$\Gamma_t = xp_t\Gamma_{t-1} - q_t\Gamma_{t-2},$$

where the values for $p_t$ and $q_t$ have been tabulated. If

$$z = \frac{N}{2\pi(1 - r^2)^{1/2}} e^{-\frac{1}{2}\frac{(x^2 - 2rxy + y^2)}{1 - r^2}}$$

then

$$z/N = \Gamma_1\Gamma_1' + 2r\Gamma_2\Gamma_2' + 3r^2\Gamma_3\Gamma_3' + \ldots + (t + 1)r^t\Gamma_{t+1}\Gamma_{t+1}' + \ldots$$

where $\Gamma_t = \Gamma_t(x)$ and $\Gamma_t' = \Gamma_t(y)$.

Finally we have

$$\int_{h_{s-1}}^{h_s} \Gamma_t \, dx = -\frac{1}{(t!)^{1/2}} D_s\Gamma_{t-1}$$

and

$$\int_{h_{s-1}}^{h_s} x\Gamma_t \, dx = -\frac{1}{(t)^{1/2}} D_s T_{t-1}$$

where

$$T_{t-1} = (t)^{1/2}\Gamma_t + (t - 1)^{1/2}\Gamma_{t-2}.$$

On the supposition that the surface is normal, has correlation r, and follows the actual marginal frequencies, the following equations are established by Pearson showing how r is related to the known parts of the m by n table:

$$\frac{\bar{n}_{sr}}{N} = \int_{h_{s-1}}^{h_s}\int_{k_{s-1}}^{k_s} \frac{z}{N}\, dx\, dy = D_s\Gamma_o D_r\Gamma_o' + rD_s\Gamma_1 D_r\Gamma_1' + r^2 D_s\Gamma_2 D_r\Gamma_2' + \ldots \quad (2)$$

$$\frac{\overline{n}_{sr}\overline{h}_{sr}}{N} = \int_{h_{s-1}}^{h_s}\int_{k_{s-1}}^{k_s} \frac{xs}{N}\,dx\,dy = D_sT_oD_r\Gamma_o' + rD_sT_1D_r\Gamma_1' + r^2D_sT_2D_r\Gamma_2' + \ldots \quad (3)$$

$$\frac{\overline{n}_{sr}k_{sr}}{N} = \int_{h_{s-1}}^{h_s}\int_{k_{s-1}}^{k_s} \frac{ys}{N}\,dx\,dy = D_s\Gamma_oD_rT_o' + rD_s\Gamma_1D_rT_1' + r^2D_s\Gamma_2D_rT_2' + \ldots \quad (4)$$

$$\frac{\overline{n}_{sr}n_{sr}}{N} = \int_{h_{s-1}}^{h_s}\int_{k_{s-1}}^{k_s} \frac{xys}{N}\,dx\,dy = D_sT_oD_rT_o' + rD_sT_1D_rT_1' + r^2D_sT_2D_rT_2' + \ldots (5)$$

These equations provide us with a large number of ways of determining r. For example:

1. We might find r, that is, $r_{sr}$ from a single cell by writing in (2) $n_{sr}$ for $\overline{n}_{sr}$.

2. We may find $r_{s.}$ from a given column of the table by using the relationship

$$\overline{h}_{s.} = \frac{1}{n_{s.}}\sum_r(n_{sr}\overline{h}_{sr}) = \frac{N}{n_{s.}}\sum_r\left[\frac{n_{sr}}{\overline{n}_{sr}}\left(D_sT_oD_r\Gamma_o' + rD_sT_1D_r\Gamma_1' + \ldots\right)\right] \quad (6)$$

where $\overline{n}_{sr}$ if given by (2), and $\overline{h}_{s.}$ is the known centroid of the $n_{s.}$ marginal total. Hence the above is an equation to find r, that is, $r_{s.}$. If we use this value of $r_{s.}$ in (2) and (4) we obtain the theoretical cell frequency $\overline{n}_{sr}$, and the mean of y for the cell $\overline{k}_{sr}$ as found from a column. Summing $\overline{k}_{sr}$ for every value of r we find $\overline{k}_{s.}$, the y mean of a column, depending on the data as found from the column. Thus

$$\overline{k}_{s.} = \frac{N}{n_{s.}}\sum_r\left[\frac{n_{sr}}{\overline{n}_{sr}}\left(D_s\Gamma_oD_sT_o' + rD_s\Gamma_1D_rT_1' + \ldots\right)\right] \quad (7)$$

This would be an ideal method of determining the mean of a row or column; but it would involve a great deal of hard work, as with the two regression curves we should need to find r for every row and column by an equation of higher order.

3. To find r for the whole table we might assume the product moment components from (5) and sum for all cells. We should have

$$\sum_{s,r} \frac{n_{sr}\bar{n}_{sr}}{N} = r,$$

since the coordinates are measured from the means in terms of the standard deviations as units. Hence substituting in (5) wehhave:

$$r = \sum_{s,r}\left[\frac{n_{sr}}{\bar{n}_{sr}}\left(D_sT_oD_rT'_o + rD_sT_1D_rT'_1 + \cdot \cdot \cdot\right)\right] \qquad (8)$$

Here $n_{sr}$ must be substituted from (2) and we have finally

$$r = \Sigma\left[\frac{n_{sr}}{N}\left(\frac{D_sT_oD_rT'_o + rD_sT_1D_rT'_1 + \cdot \cdot \cdot}{D_s\Gamma_oD_r\Gamma'_o + rD_s\Gamma_1D_r\Gamma'_1 + \cdot \cdot \cdot}\right)\right] \qquad (9)$$

It will be observed that what we are trying to do is to fit a normal correlation surface to a series of cell frequencies. If the observed results are closely normal then $n_{sr}$ would be nearly equal to $\bar{n}_{sr}$. If we might assume the difference $n_{sr}$ and $\bar{n}_{sr}$ so small as to be negligible we should have:

$$r = \sum_{s,r}\left(D_sT_oD_sT'_o + rD_sT_1D_rT'_1 + \cdot \cdot \cdot\right) \qquad (10)$$

4. Let us consider what the most probable value for r might be. We observe $n_{sr}$ as the frequency of the sth-rth cell; we find that with a given correlation r the frequency of this cell would be $\bar{n}_{sr}$, on the assumption that the frequency surface is the normal frequency surface corresponding to the observed marginal totals. Accordingly, the most probable value to give to r would be that which made

$$\chi^2 = \sum_{s,r} \frac{(\overline{n}_{sr} - n_{sr})^2}{\overline{n}_{sr}} \text{ a minimum}$$

or, what is the same thing,

$$\sum_{s,r} \left(\frac{n^2_{sr}}{\overline{n}_{sr}}\right) \text{a minimum.}$$

This leads us, differentiating with regard to r, to

$$\sum_{s,r} \left[\left(\frac{n_{sr}}{\overline{n}_{sr}}\right)^2 \frac{d\overline{n}_{sr}}{dr}\right] = 0$$

or, writing at length, our equation for r is:

$$\sum_{s,r} \left[\left(\frac{n_{sr}}{N}\right)^2 \frac{D_s\Gamma_1 D_r\Gamma_1' + 2r\, D_s\Gamma_2 D_r\Gamma_2' + \ldots}{(D_s\Gamma_o D_r\Gamma_o' + r\, D_s\Gamma_1 D_r\Gamma_1' + \ldots)^2}\right] = 0 \qquad (11)$$

Again if we assume the differences of $n_{sr}$ and $\overline{n}_{sr}$ negligible, we have

$$\sum_{s,r} (D_s\Gamma_1 D_r\Gamma_1' + 2r\, D_s\Gamma_2 D_r\Gamma_2' + \ldots) = 0 \qquad (12)$$

It will be found that the equations (10) and (12) are identically satisfied. Hence our values for r from (8) and (11) depend on $n_{sr}$ differing from $\overline{n}_{sr}$. Without the assumption that $\overline{n}_{sr}$ may be replaced by $n_{sr}$ neither (8) nor (11) are readily solvable. Probably the easiest way will be to obtain an approximate value of r, one well above and one well below this result, so that the real value of r lies between the two. A linear interpolation will probably suffice in most cases to determine r with sufficient accuracy.

The polychoric table as discussed by Ritchie-Scott[1] describes another method of reaching a polychoric coefficient from the weighted means of the

[1] A. Ritchie-Scott, "The Correlation Coefficient of a Polychoric Table", Biometrika, X11, pp. 106-108.

possible tetrachoric values.

The frequency surface divided into p columns and q rows is divided at each point into four quandrants, and for each of these divisions a value for r, that is, $r_{11}$, $r_{12}$, . . . , may be found by the tetrachoric method. These may be regarded as approximations to the true value of r, and their weighted mean found, the weights being determined so that the probable error of the mean r so found shall be a minimum.

$$\text{Let } r = \frac{C_{11}r_{11} + C_{12}r_{12} + \cdots}{C_{11} + C_{12} + \cdots}$$

$$\text{Then } (C_{11} + C_{12} + \cdots)dr = C_{11}dr_{11} + C_{12}dr_{12} + \cdots$$

Squaring, and summing for all possible values and dividing by the number of samples,

$$(\Sigma C_{st})^2 \sigma_r^2 = \Sigma(C_{st}\sigma_{st})^2 + 2\Sigma(C_{st}C_{s't'}\sigma_{st}\sigma_{s't'}R_{sts't'})$$

where $\sigma_{st} = \sigma_{r_{st}}$, and $R_{sts't'} = r_{r_{st}r_{s't'}}$.

If $S = \Sigma(C_{st}\sigma_{st})^2 + 2\Sigma(C_{st}C_{s't'}\sigma_{st}\sigma_{s't'}R_{sts't'})$,

$C = \Sigma(C_{st})$

and

$$\sigma_r^2 = \frac{S}{C^2}.$$

Then for a minimum

$$\frac{\delta(\sigma_r^2)}{\delta C_{st}} = 0$$

or

$$\frac{\delta(\delta r^2)}{\delta C_{11}} = \frac{\delta(S/C^2)}{\delta C_{11}} = \frac{C^2\frac{\delta S}{\delta C_{11}} - 2CS\frac{\delta C}{\delta C_{11}}}{C^4} = 0$$

But $\frac{\delta C}{\delta C_{11}} = 1$ and if we let $S/C = \Gamma$ then $\frac{\delta S}{\delta C_{11}} = \Gamma$.

Similarly $\frac{\delta S}{\delta C_{12}} = \frac{\delta S}{\delta C_{13}} = \dots\dots = \Gamma$,

and one obtains the following equations:

$$C_{11}\sigma_{11}^{2} + C_{12}\sigma_{11}\sigma_{12}R_{11,12} + C_{13}\sigma_{11}\sigma_{13}R_{11,13} + \dots = \Gamma \qquad (13)$$

$$C_{11}\sigma_{11}\sigma_{12}R_{11,12} + C_{12}\sigma_{12}^{2} + C_{13}\sigma_{12}\sigma_{13}R_{12,13} + \dots = \Gamma \qquad (14)$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

The values for the C's are determined by solving these equations simultaneously.

## Approximate Methods

The Ritchie-Scott process is so laborious that it can hardly establish itself in practice, whereas, Pearson's method of evaluating the polychoric coefficient is not too tedious providing one has access to the tables referred to and when a high degree of accuracy is required, it is the best method available. However if one is willing to sacrifice accuracy, there are several simplified methods of approximating r.

A recent publication by Camp[1] provides us with a simplified method of approximating both the tetrachoric coefficient and the polychoric coefficient. These very short methods cannot guarantee to give r accurately to more than one or two decimal places.

Let us first consider the tetrachoric r as found from a 2 by 2 fold classification. Camp replaces the frequencies in Table I by the ratios as

[1] Camp, op. cit., pp. 302-310.

given in Table III.

Table III

| | | | |
|---|---|---|---|
| $A_1$ | $A_2$ | | |
| $B_1$ | $B_2$ | | |
| $F_1$ | $F_2$ | | 1 |

where $f_1 = a + c$, $f_2 = b + d$,

and $A_1 = a/f_1$, $A_2 = b/f_1$, $B_1 = c/f_1$, $B_2 = d/f_2$, $F_1 = f_1/N$, $F_2 = f_2/N$.

The problem again is to find a normal surface which when divided into four cells will present in these cells the relative frequencies observed. The procedure is as follows: Find $x$, $y_1$, $y_2$, from the relations

$$\int_{-\infty}^{x}\phi(x)\,dx = F_1,\quad \int_{-\infty}^{y_1}\phi(x)\,dx = A_1,\quad \int_{-\infty}^{y_2}\phi(x)\,dx = B_2.$$

Then $m = F_1F_2\,\dfrac{y_1 + y_2}{\phi(x)}$, where m is the slope of the regression line y on x.

Since $m = \dfrac{r}{(1 - r^2)^{1/2}}$,

it follows that $r = \dfrac{m}{(1 + m^2)^{1/2}} = \sin\tan^{-1} m.$

This formula rests on three assumptions: first, that for such a division of a normal surface the mean of each column would lie on the regression line; second, that the standard deviation of each column would equal $\sigma_y(1 - r^2)^{1/2}$; and third, that, considered as a one-way distribution in the y direction, the distribution would be normal.

In order to estimate the polychoric r by means of the simplified method given by Camp it will be necessary to subdivide the frequency table in the following manner. Let AB be any of the horizontal divisions of the table which cuts all the frequency columns. In the first column let $a_1$ denote the total frequency above AB, $b_1$ the total frequency below; in the second column use similarly $a_2$ and $b_2$, etc., as indicated in the following Table.

Table IV

| | y \ x | 1 | 2 | 3 | | | | m | |
|---|---|---|---|---|---|---|---|---|---|
| A | Frequency above AB | $a_1$ | $a_2$ | $a_3$ | | | | $a_{m1}$ | B |
| | Frequency below AB | $b_1$ | $b_2$ | $b_3$ | | | | $b_{m1}$ | |
| | Totals | $n_1$ | $n_2$ | $n_3$ | | | | $n_{m1}$ | |

First one finds the mean $\bar{h}_1$, $\bar{h}_2$, etc., which are referred to the mean abscissa of the whole table as the origin, and the units of measurement are $\sigma_x$. Now considering each of the columns individually, let $\sigma_1$, $\sigma_2$, etc., be the standard deviations of the columns. These standard deviations are all approximately equal to $\sigma_y(1 - r^2)^{1/2}$, a fortunate circumstance which is essential to the success of this method. Let $\bar{y}_1$, $\bar{y}_2$, etc., be the distances of the means of the several columns below the line AB in terms of $\sigma_1$, $\sigma_2$, etc., as units. If the columns are normal distributions, these distances can be found from the equations:

$$\int_{-\infty}^{\bar{y}_1}\phi(x)\,dx = b_1/n_1, \qquad \int_{-\infty}^{\bar{y}_2}\phi(x)\,dx = b_2/n_2, \text{ etc.}$$

One may thus obtain, relative to a horizontal axis AB and a vertical axis through the general mean point of the whole table, the following coordinates

of the mean points of the several columns in $\sigma_x$ and $\sigma_y$ units:
$[\bar{h}_1, \bar{y}_1(1 - r^2)^{1/2}]$, $[\bar{h}_2, \bar{y}_2(1 - r)^{1/2}]$, etc.

Now if there is a normal surface satisfying the conditions laid down at the outset, its regression line y on x passes through these mean points; its slope is r in the $\sigma_x$ and $\sigma_y$ units and $\frac{1}{(1 - r^2)^{1/2}}$ in the $\sigma_x$ and $\sigma_1$ units. If these mean points do not lie approximately on a line, there is no normal surface which approximately fits the data and so the method cannot be used.

To find the slope of the line, least squares might be used. If graphical methods are desired it might be necessary to accord greater weight to points representative of greater column frequencies. Also columns in which a or b is very small should be given very little weight.

Since the slope $m = \frac{r}{(1 - r^2)^{1/2}}$,

we can solve for $r = \frac{m}{(1 + m^2)^{1/2}} = \sin \tan^{-1} m$.

Pearson has also contributed simplified methods whereby we may obtain an r which is an approximation to the true correlation. The correlation from marginal centroids, and the mean contingency method will be the two methods of Pearson's that we will discuss.

In discussing the correlation from marginal centroids[1], it will first be necessary to normalize the series. That is, we assign to several groups

[1] Karl Pearson, Biometrika XIV, PP. 128-129.

of an ordered series their proper spacing so that the whole will fit the normal curve. This means of course, that some groups will be squeezed into shorter intervals, others spread over longer ones. Automatically now, the means are at the origin, and the standard deviation are the units.

The means of the categories $n_{s\cdot}$ and $n_{\cdot r}$ are determined by

$$\bar{h}_s = \frac{H_{s-1} - H_s}{n_{s\cdot}/N} \qquad \text{and } \bar{k}_r = \frac{K_{s-1} - K_s}{n_{\cdot r}/N}$$

respectively.[1] The numerical values of $\bar{h}_s$ and $\bar{k}_r$ can now be found. Care must be taken in every case to give the correct sign to $\bar{h}_s$ and $\bar{k}_r$.

Now if there is no correlation, $\bar{h}_s$ and $\bar{k}_r$ combined would give the coordinates of the mean point of the $n_{sr}$ group, and they give a fair approximation to the result if there are numerous categories, that is, if the range of the categories be small. The correlation found from the marginal centroids would then be

$$r = \frac{\Sigma(n_{sr}\bar{h}_s\bar{k}_r)}{N}. \qquad (16)$$

It can be shown that this r is a poorer approximation of the true correlation than the tetrachoric r or the polychoric r. The reason for this is that $\bar{h}_s$ and $\bar{k}_r$ do not give the coordinate of the mean of $n_{sr}$. In fact $n_{sr}\bar{h}_s\bar{k}_r$ is not the contribution of the $n_{sr}$ group to the product moment.

Now let us consider the mean contingency method as developed by Pearson.[2] If $n_{sr}$ be the frequency in the cell of the sth column and the

[1] For details see page 18.

[2] Pearson, "On the Theory of Contingency", Draper's Company Research Memoirs, No. I.

rth row of a correlation table and $n_{s.}$ be the total frequency in the sth column, $n_{.r}$ the total frequency in the rth row, then if two variates are independent, the frequency to be expected in the sth-rth cell will be

$$N\left(\frac{n_{s.}}{N}\right)\left(\frac{n_{.r}}{N}\right) = \frac{n_{s.}n_{.r}}{N}$$

The observed excess over this, i.e., $n_{sr} - \dfrac{n_{s.}n_{.r}}{N}$ is termed the contingency in the cell. The total contingency must of course be zero, that is, the sum of all the cell contingencies.

To find the so-called mean contingency, $\psi$, one sums all the positive excess contingencies and divides by n, obtaining

$$\psi = \frac{1}{N}\sum_{+}\left(n_{sr} - \frac{n_{s.}n_{.r}}{N}\right).$$

Assuming a normal frequency distribution it is possible to deduce the actual correlation from $\psi$, provided that the cells are sufficiently small. Generally a value below that of the true correlation, even if the system be accurately normal, is found. A corrective factor has not as yet been theoretically deduced, but experience seems to show that to add half the correction due to class index correlation[1] gives good results.

If the mean contingency correlation is denoted by $r_{\psi}$, and $r_{xC_x}$ and $r_{yC_y}$ be the class index correlations for x and y, we should take for the true correlation:

[1] Class index correlations denoted by $r_{xC_x}$ and $r_{yC_y}$ give the correlation between variate and its class mark. As we increase the number of classes the index correlations approach unity. Values of class index correlations are tabulated in Biometrika IX pp 121 and 218.

$$r = r_\psi - \frac{1}{2}\left(\frac{r_\psi}{r_{xC_x}\ r_{yC_y}} - r_\psi\right)$$

$$= \frac{1}{2}\left(r_\psi - \frac{r_\psi}{r_{xC_x} r_{yC_y}}\right).$$

# CHAPTER III

## DETERMINATION OF WEIGHTS BY MINIMIZING THE STANDARD ERROR OF ESTIMATE

Let us regard a series of observations of two variables, say x and y, where the independent variable x is classified into groups and the dependent variable y is measured. We shall assume that there exists a linear relationship between the variables. Our problem is to determine the best method of estimating the dependent variable, y, from the classified independent variable, x. In order to do this we will determine weights to be assigned to the ordered classes of x's such that the standard error of estimate for the regression of y on x is minimized. From Chapter I we found that $S_y{}^2 = \sigma_y{}^2(1 - r^2)$, and hence when the standard error of estimate is minimized the correlation coefficient is maximized, and conversely. It follows therefore that our problem may also be considered as a determination of weights for the classes of the x's such that the correlation coefficient, as found by Pearson's product moment formula, shall be a maximum.

Since the data which we use is such that each class of the x's may contain several variates, it will be convenient to determine a value for y from each class that can be used to correspond to each weight assigned to the x's. For this y value we shall use the mean value of the y's within each class. The data which we assume to be classified into nx-classes, can then be represented graphically by n points whose coordinates are the weight assigned to the class and the mean y value for that class. We shall denote the coordinates of the ith point by $(t_i, \bar{y}_i)$. The following table is given to clarify the notation that will be used.

Table V

| x-Class | 1 | 2 | . . . . . | n |
|---|---|---|---|---|
| y's | $y_{11}$ | $y_{21}$ | | $y_{n1}$ |
| | $y_{12}$ | $y_{22}$ | | $y_{n2}$ |
| | | | | |
| | $y_{1f_1}$ | $y_{2f_2}$ | | $y_{nf_n}$ |
| Frequencies | $f_1$ | $f_2$ | | $f_n$ |
| Mean y's | $\bar{y}_1$ | $\bar{y}_2$ | | $\bar{y}_n$ |
| Weights | $t_1$ | $t_2$ | | $t_n$ |

Before going into the methods of determining the weights let us state and prove three theorems related to the problem.

Theorem I. The changing of weights in proportion does not effect the correlation coefficient nor the standard error of estimate.

Consider a series of observations $x_1y_1, x_2y_2, \ldots, x_iy_i, \ldots$ where the y's are numerical and the x's are classified into n classes with the weight $t_i$ assigned to the ith class as indicated in Table V. Since $f_i$ equals the number of variates in the ith class, the total number of variates $N = \sum_{i=1}^{n} f_k$.

We can write the correlation coefficient as

$$r = \frac{\sum_{i=1}^{n} f_i \sum_{i=1}^{n}\sum_{j=1}^{f_i} t_i y_{ij} - \sum_{i=1}^{n} f_i t_i \sum_{i=1}^{n}\sum_{j=1}^{f_i} y_{ij}}{\left[\sum_{i=1}^{n} f_i \sum_{i=1}^{n} f_i t_i^{2} - \left(\sum_{i=1}^{n} f_i t_i\right)^{2}\right]^{1/2} \sum_{i=1}^{n} f_i \sigma_y}$$

Now changing the weights in proportion, i.e. multiplying each $t_i$ by some constant k, the correlation coefficient becomes

$$r' = \frac{\sum_{i=1}^{n} f_i \sum_{i=1}^{n}\sum_{j=1}^{f_i} kt_i y_{ij} - \sum_{i=1}^{n} f_i kt_i \sum_{i=1}^{n}\sum_{j=1}^{f_i} y_{ij}}{\left[\sum_{i=1}^{n} f_i \sum_{i=1}^{n} f_i k^{2} t_i^{2} - \left(\sum_{i=1}^{n} f_i kt_i\right)^{2}\right]^{1/2} \sum_{i=1}^{n} f_i \sigma_y}.$$

Factoring the constant k from the numerator and denominator we have

$$r' = \frac{k\left(\sum_{i=1}^{n} f_i \sum_{i=1}^{n}\sum_{j=1}^{f_i} t_i y_{ij} - \sum_{i=1}^{n} f_i t_i \sum_{i=1}^{n}\sum_{j=1}^{f_i} y_{ij}\right)}{k\left[\sum_{i=1}^{n} f_i \sum_{i=1}^{n} f_i t_i^{2} - \left(\sum_{i=1}^{n} f_i t_i\right)^{2}\right]^{1/2} \sum_{i=1}^{n} f_i \sigma_y}$$

$$\frac{\sum_{i=1}^{n} f_i \sum_{i=1}^{n}\sum_{j=1}^{f_i} t_i y_{ij} - \sum_{i=1}^{n} f_i t_i \sum_{i=1}^{n}\sum_{j=1}^{f_i} y_{ij}}{\left[\sum_{i=1}^{n} f_i \sum_{i=1}^{n} f_i t_i^{2} - \left(\sum_{i=1}^{n} f_i t_i\right)^{2}\right]^{1/2} \sum_{i=1}^{n} f_i \sigma_y}.$$

Therefore $r' = r$.

Theorem II. The correlation coefficient is independent of the weight assigned when the observations are classified into but two classes.

Let $t_1$ and $t_2$ be the weights assigned to the two classes. Now expressing $r$ as a function of $t_1$ and $t_2$ we have

$$r = \frac{(f_1+f_2)(t_1 f_1 \bar{y}_1 + t_2 f_2 \bar{y}_2) - (f_1 t_1 + f_2 t_2)(f_1 \bar{y}_1 + f_2 \bar{y}_2)}{\left[(f_1+f_2)(f_1 t_1^{2} + f_2 t_2^{2}) - (f_1 t_1 + f_2 t_2)^{2}\right]^{1/2} (f_1+f_2)\sigma_y}$$

$$r = \frac{f_1 f_2 (t_2 - t_1)(y_2 - y_1)}{\left[f_1 f_2 (t_2 - t_1)^2\right]^{1/2} (f_1 + f_2)\sigma_y}$$

$$= \frac{(f_1 f_2)^{1/2} (y_2 - y_1)}{(f_1 + f_2)\sigma_y} \; .$$

Since this value of r is independent of $t_i$ we have our conclusion that the correlation coefficient is independent of the weights assigned when the data is classified into two classes.

In our work that follows a proposition taken from the analysis of variance will simplify our considerations.

Theorem III. Within any class the variance of the points from the regression value is equal to the sum of the variance of the points from the column mean and the variance of the column mean from the regression value.

Let us consider any class in Table V, say the ith class. If $y_c$ is the regression value for this class, then we can write

$$y_{ij} - y_c = y_{ij} - \bar{y}_i + \bar{y}_i - y_c$$

then

$$\sum_{j=1}^{f_i} (y_{ij} - y_c)^2 = \sum_{j=1}^{f_i} (y_{ij} - \bar{y}_i)^2 + f_i(\bar{y}_i - y_c)^2$$

$$+ 2 \sum_{j=1}^{f_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - y_c).$$

But

$$\sum_{j=1}^{f_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - y_c) = (\bar{y}_i - y_c) \sum_{j=1}^{f_i} (y_{ij} - \bar{y}_i) = 0.$$

So

$$\sum_{j=1}^{f_i}(y_{ij} - y_c)^2 = \sum_{j=1}^{f_i}(y_{ij} - \bar{y}_i)^2 + f_i(\bar{y}_i - y_c)^2 \qquad (1)$$

From the above theorem we see that since $\sum_{j=1}^{f_i}(y_{ij} - \bar{y}_i)^2$ is independent of the regression line, to minimize the left hand side of equation (1), one needs but to minimize the quantities $(\bar{y}_i - y_c)^2$ for the various classes. This means that the regression line can be determined from the mean points of each column instead of using all the points within each column.

Now in considering the general problem of determining values for the weights $t_1, t_2, \ldots t_n$, to be assigned to the n classes so as to minimize the standard error of estimate, we have from the preceeding development, that the problem can be reduced to that of determining the weights so that the points, $(t_1,\bar{y}_1), (t_2,\bar{y}_2), \ldots (t_n,\bar{y}_n)$, are collinear. This can be done by assigning to any two classes arbitrary values for the weights, say $t_1$ and $t_2$. Then the points, $(t_1,\bar{y}_1)$ and $(t_2,\bar{y}_2)$, determine a straight line and the value of the remaining weights will therefore depend upon the equation of this straight line and the corresponding $\bar{y}$ values. It can readily be justified from analytical geometry that the value of $t_i$ in terms of $t_1$ and $t_2$, to make the point $(t_i,\bar{y}_i)$ fall on the line, is

$$t_i = \frac{t_2\bar{y}_1 - t_1\bar{y}_2 + t_1\bar{y}_i - t_2\bar{y}_i}{\bar{y}_1 - \bar{y}_2}.$$

This value will be appropriate for the ith class as long as the $\bar{y}$'s for the n classes are in the same order as the x-classes.

To verify that these weights will minimize the standard error of estimate obtained in regression, we can approach the problem of determining the weights by means of analysis. That is, we shall consider the problem of determining the weights by using the methods of the calculus so as to maximize the correlation coefficient. Since a more or less inductive approach will be used, we shall first consider the cases where the data are classified into three classes and then four classes before considering the general case where the x's fall into n classes.

Case (i). With the data classified into three classes, let the weights assigned to each class be $t_1$, $t_2$, and $t_3$. Expressing r as a function of $t_1$, $t_2$, and $t_3$, we have

$$r = \frac{(f_1+f_2+f_3)(t_1f_1\bar{y}_1+t_2f_2\bar{y}_2+t_3f_3\bar{y}_3) - (f_1t_1+f_2t_2+f_3t_3)(f_1\bar{y}_1+f_2\bar{y}_2+f_3\bar{y}_3)}{\left[(f_1+f_2+f_3)(f_1t_1^2+f_2t_2^2+f_3t_3^2)-(f_1t_1+f_2t_2+f_3t_3)^2\right]^{1/2}\ (f_1+f_2+f_3)\sigma_y}$$

To maximize r we differentiate r partially with respect to $t_1$, $t_2$, and $t_3$, respectively, and setting each of the derivatives equal to zero, we obtain the following three equations:

(a) $0\cdot t_1^2 + (\bar{y}_1-\bar{y}_3)t_2^2 + (\bar{y}_1-\bar{y}_2)t_3^2 + (\bar{y}_3-\bar{y}_2)t_1t_2 + (-2\bar{y}_1+\bar{y}_2+\bar{y}_3)t_2t_3 + (\bar{y}_2-\bar{y}_3)t_1t_3=0$

(b) $(\bar{y}_2-\bar{y}_3)t_1^2 + 0\cdot t_2^2 + (\bar{y}_2-\bar{y}_1)t_3^2 + (\bar{y}_3-\bar{y}_1)t_1t_2 + (\bar{y}_1-\bar{y}_3)t_2t_3 + (\bar{y}_1-2\bar{y}_2+\bar{y}_3)t_1t_3= 0$

(c) $(\bar{y}_3-\bar{y}_2)t_1^2 + (\bar{y}_3-\bar{y}_1)t_2^2 + 0\cdot t_3^2 + (\bar{y}_1+\bar{y}_2-2\bar{y}_3)t_1t_2 + (\bar{y}_1-\bar{y}_2)t_2t_3 + (\bar{y}_2-\bar{y}_1)t_1t_3=0.$

Since each of the equations are linear in one of the three variables $t_1$, $t_2$, and $t_3$, let us solve for, say $t_3$, from equation (e). We obtain

$$t_3 = \frac{t_2\bar{y}_1 - t_1\bar{y}_2 + t_1\bar{y}_3 - t_2\bar{y}_3}{\bar{y}_1 - \bar{y}_2}$$

Then verifying that this solution satisfies equations (a) and (b) we have a common solution for the above equations, that is $t_3$ in terms of $t_1$ and $t_2$. Hence we may arbitrarily assign values to the weights $t_1$ and $t_2$ from which $t_3$ must be determined to have a maximum value for $r$. This value for $t_3$ is the same as was obtained by fitting the mean points to a straight line.

Case (ii). Similarly, let $t_1$, $t_2$, $t_3$, and $t_4$, be the weights assigned to each class when the data is classified into four classes. Again expressing $r$ as a function of the four weights we have $r$ equal to

$$\frac{(f_1+f_2+f_3+f_4)(t_1f_1\bar{y}_1+t_2f_2\bar{y}_2+t_3f_3\bar{y}_3+t_4f_4\bar{y}_4)-(f_1t_1+f_2t_2+f_3t_3+f_4t_4)(f_1\bar{y}_1+f_2\bar{y}_2+f_3\bar{y}_3+f_4\bar{y}_4}{\left[(f_1+f_2+f_3+f_4)(f_1t_1^2+f_2t_2^2+f_3t_3^2+f_4t_4^2)-(f_1t_1+f_2t_2+f_3t_3+f_4t_4)^2\right]^{1/2}(f_1+f_2+f_3+f_4)\sigma_y}$$

Then differentiating $r$ partially with respect to each of the four weights and setting the derivatives equal to zero we obtain four equations:

$$\text{(d)}\quad 0 \cdot t_1^2 + t_2^2(2\bar{y}_1-\bar{y}_3-\bar{y}_4) + t_3^2(2\bar{y}_1-\bar{y}_2-\bar{y}_4) + t_4^2(2\bar{y}_1-\bar{y}_2-\bar{y}_3) + t_1t_2(-2\bar{y}_2+\bar{y}_3+\bar{y}_4) +$$

$$t_1t_3(\bar{y}_2-2\bar{y}_3+\bar{y}_4) + t_1t_4(\bar{y}_2+\bar{y}_3-2\bar{y}_4) + t_2t_3(-2\bar{y}_1+\bar{y}_2+\bar{y}_3) + t_2t_4(-2\bar{y}_1+\bar{y}_2+\bar{y}_4) +$$

$$t_3t_4(-2\bar{y}_1+\bar{y}_3+\bar{y}_4) = 0$$

$$(e)\ t_1^2(2\bar{y}_2-\bar{y}_3-\bar{y}_4) + 0\cdot t_2 + t_3^2(\bar{y}_1+2\bar{y}_2+\bar{y}_4) + t_4^2(-\bar{y}_1+2\bar{y}_2-\bar{y}_3) + t_1t_2(-2\bar{y}_1+\bar{y}_3+\bar{y}_4) +$$
$$t_1t_3(\bar{y}_1-2\bar{y}_2+ \bar{y}_3) + t_1t_4(\bar{y}_1-2\bar{y}_2+\bar{y}_4) + t_2t_3(\bar{y}_1-2\bar{y}_3+\bar{y}_4) + t_2t_4(\bar{y}_1+\bar{y}_3-2\bar{y}_4) +$$
$$t_3t_4(-2\bar{y}_2+\bar{y}_3+\bar{y}_4) = 0$$

$$(f)\ t_1^2(-\bar{y}_2+2\bar{y}_3-\bar{y}_4) + t_2^2(-\bar{y}_1+2\bar{y}_3-\bar{y}_4) + 0\cdot t_3^2 + t_4^2(-\bar{y}_1-\bar{y}_2+2\bar{y}_3) + t_1t_2(\bar{y}_1+\bar{y}_2-2\bar{y}_3) +$$
$$t_1t_3(-2\bar{y}_1+ \bar{y}_2+\bar{y}_4) + t_1t_4(\bar{y}_1-2\bar{y}_3+\bar{y}_4) + t_2t_3(\bar{y}_1-2\bar{y}_2+\bar{y}_4) + t_2t_4(\bar{y}_2-2\bar{y}_3+ \bar{y}_4) +$$
$$t_3t_4(\bar{y}_1+\bar{y}_2-2\bar{y}_4) = 0$$

$$(g)\ t_1^2(-\bar{y}_2-\bar{y}_3+2\bar{y}_4) + t_2^2(-\bar{y}_1-\bar{y}_3+2\bar{y}_4) + t_3^2(-\bar{y}_1-\bar{y}_2+2\bar{y}_4) + 0\cdot t_4^2 + t_1t_2(\overset{6}{\bar{y}_2}+\bar{y}_3-2\bar{y}_4) +$$
$$t_1t_3(\bar{y}_1+\bar{y}_3-2\bar{y}_4) + t_1t_4(-2\bar{y}_1+\bar{y}_2+\bar{y}_3) + t_2t_3(\bar{y}_2+\bar{y}_3-2\bar{y}_4) + t_2t_4(\bar{y}_1-2\bar{y}_2+\bar{y}_3) +$$
$$t_3t_4(\bar{y}_1+\bar{y}_2-2\bar{y}_3) = 0.$$

In solving these four equations we shall express each of the four equations as the sum of three equations. That is, equation (e) may be expressed as the sum of the following three equations:

$$(1)\ 0\cdot t_2^2 + (\bar{y}_2-\bar{y}_3)t_1^2 + (\bar{y}_2-\bar{y}_1)t_3^2 + (\bar{y}_3-\bar{y}_1)t_1t_2 + (-2\bar{y}_2+\bar{y}_1+\bar{y}_3)t_1t_3+ (\bar{y}_1-\bar{y}_3)t_2t_3= 0$$

$$(2)\ 0\cdot t_2^2 + (\bar{y}_2-\bar{y}_4)t_1^2 + (\bar{y}_2-\bar{y}_1)t_4^2 + (\bar{y}_4-\bar{y}_1)t_1t_2 + (-2\bar{y}_2+\bar{y}_1+\bar{y}_4)t_1t_4+ (\bar{y}_1-\bar{y}_4)t_2t_4= 0$$

$$(3)\ 0\cdot t_2^2 + (\bar{y}_2-\bar{y}_4)t_3^2 + (\bar{y}_2-\bar{y}_3)t_4^2 + (\bar{y}_4-\bar{y}_3)t_2t_3 + (-2\bar{y}_2+\bar{y}_3+\bar{y}_4)t_3t_4 + (\bar{y}_3-\bar{y}_4)t_2t_4= 0.$$

Since equation (1) and equation (b) are the same equation they are satisfied by

$$t_3 = \frac{t_2\bar{y}_1 - t_1\bar{y}_2 + t_1\bar{y}_3 - t_2\bar{y}_3}{\bar{y}_1 - \bar{y}_2}.$$

Equation (2) and equation (b) are similar equations and must be satisfied by similar values so that

$$t_4 = \frac{t_2\bar{y}_1 - t_1\bar{y}_2 + t_1\bar{y}_4 - t_2\bar{y}_4}{\bar{y}_1 - \bar{y}_2} .$$

Equation (3) is satisfied by the solutions from equations (1) and (2), hence equation (e) is satisfied by expressing the values for $t_3$ and $t_4$ in terms of $t_1$ and $t_2$.

Since each of the equations (d), (f), and (g), can be expressed as the sum of three equations which are similar to equations (1), (2), and (3), the values of $t_3$ and $t_4$ which satisfy equation (e) will also satisfy equations (d), (f), and (g). This leads us to the conclusion that to maximize the correlation coefficient in the case of 4 classes, one can arbitrarily choose $t_1$ and $t_2$ and then the values for $t_3$ and $t_4$ are determined.

Now we are ready to discuss data which is classified into n classes. Expressing r as a function of the weights assigned to the n classes we have

$$r = \frac{\sum_{i=1}^{n} f_i \sum_{i=1}^{n} t_i f_i \bar{y}_i - \sum_{i=1}^{n} f_i t_i \sum_{i=1}^{n} f_i \bar{y}_i}{\left[\sum_{i=1}^{n} f_i \sum_{i=1}^{n} f_i t_i^{\,2} - \left(\sum_{i=1}^{n} f_i t_i\right)^2\right]^{1/2} \sum_{i=1}^{n} f_i \sigma_y} .$$

Upon differentiating r partially with respect to each of the n weights and setting the derivatives equal to zero, we obtain n equations. Using summation notation, the ith of these n equations, (i.e. the equation obtained when differentiating with respect to $t_i$), may be expressed as

$$(j) \qquad \sum_{\substack{j,k=1 \\ \text{where } j \neq k \neq i}}^{n} \left[(y_i-y_k)t_j^2 + (y_k-y_j)t_it_j + (-2y_i+y_k+y_j)t_jt_k\right] = 0.$$

The notation $\sum_{\substack{j,k=1 \\ \text{where } j \neq k \neq i}}^{n}$ means that the subscripts j and k take on all values from 1 to n except the ith value but are never equal to each other. The ith equation corresponds to one of the equations (a), (b), and (c), in case (i), and to one of the equations (d), (e), (f), and (g), in case (ii). Now equations(j) may be expressed as the sum of $_{n-1}C_2$ equations of the form

$$(k) \quad 0\cdot t_i^2 + (y_i-y_k)t_j^2 + (y_i-y_j)t_k^2 + (y_k-y_j)t_it_j + (-2y_i+y_j+y_k)t_jt_k + (y_j-y_k)t_it_k=0$$

where $j \neq k \neq i$, and j and k take on values from 1 to n yielding $_{(n-1)}C_2$ different equations. These equations correspond to equations (1), (2), and (3), in case (ii). Furthermore, it can readily be seen that these equations are similar to equations (1), (2), and (3), and therefore similar to equations (a), (b), and (c), also. Since n-2 of these equations contain the variables $t_1$, $t_2$, and one other $t_i$ we can use these equations to solve for n-2 of the t's in terms of $t_1$ and $t_2$. The solutions obtained for n-2 of the t's are similar to the solutions from case (i) and (ii) which would be expected since the equations, as previously stated, are similar. The solutions obtained from the n-2 equations of the type (k) will satisfy the remaining equations of that set. Now the solutions satisfying the ith equations in n variables will satisfy the remaining equations in n variables since the parts they split up into have the same solution. The solution for $t_i$ may be written as

$$\frac{t_2\bar{y}_1 - t_1\bar{y}_2 + t_1\bar{y}_i - t_2\bar{y}_i}{\bar{y}_1 - \bar{y}_2},$$

which checks with the result obtained from the geometric method previously discussed.

We have obtained, therefore, by means of geometry as well as by means of analysis an assignment for weights which depend upon any two of the weights which are given arbitrary values, and the corresponding $\bar{y}$ values. These weights will render a maximum value for the correlation coefficient thus minimizing the standard error of estimate.

## CHAPTER IV

## A NUMERICAL EXAMPLE

We shall now regard a numerical illustration employing the method of assignment of weights that has been discussed in Chapter III. Using these weights, the regression line, the standard error, and the correlation coefficient will be computed. This numerical example for comparison will also include the computation of the regression line, standard error, and the correlation coefficient from an assignment of evenly spaced weights[1] together with weights derived from marginal centroids.[2]

The data used in the following illustration is the mathematical placement test scores and the first term mathematics grades compiled from 207 freshman students entering the University of Oregon in the fall of 1939. The score received in the math placement test will be denoted by y, the measurable series, while the grade received in the respective math courses taken will be considered as the ordered series x. There are twenty classes for x since there were four different freshman courses classified as I, II, III, and IV,[3] and within each course are five classifications according to grades received. The order of the classes is such

[1] Carl F. Kossack, "Mathematics Placement at the University of Oregon," American Mathematical Monthly XLIX, No. 4 (April 1942).

[2] See pages 27 and 28.

[3] Courses classified as I, II, III, and IV, denote Introduction to Algebra, Intermediate Algebra, College Algebra, and Introduction to Analysis, respectively.

that the first class in the x series consists of students that received F(failure) in Course I, the second class, those who received D in Course I, etc., and the twentieth class would therefore consist of students that received an A in Course IV. The raw data is given in the appendix in Table VIII.

We shall arbitrarily let the weights range from 0 to 100, remembering that in Theorem I it was established that the correlation coefficient is not effected when the weights are changed in proportion. In order to determine the minimizing weights it will be necessary to arbitrarily assign values to any two of the weights. It is convenient to let $t_1$ equal 0 and $t_{20}$ equal 100. The remaining weights are then determined from the formula derived in Chapter III

$$t_i = \frac{t_{20}\bar{y}_1 - t_1\bar{y}_{20} + t_1\bar{y}_i - t_{20}\bar{y}_i}{y_1 - \bar{y}_{20}}$$

These data are somewhat irregular in that the means of the y series are not in exactly the same order as the x-classes. There are three mean y values that are out of order and consequently the values for the corresponding weights will not be in the desired order. A second set of weights are determined by making an arbitrary adjustment to the above weights such that the weights will be in order. Still another set of weights are determined by choosing a set of weights that are evenly spaced. These are obtained by dividing the range from 0 to 100 so that with the first weight equal to zero the value of the remaining 19 weights will differ from each successive weight by 100/19 units. Finally a set of weights referred to as normalized centroid weights are determined in the same manner as in the technique discussed in Chapter II that involved marginal centroids. The centroid values thus found are multiplied by the proper constants so that the value of the centroids range from 0 to 100. A table of the four assignments of weights is given below.

Table VI

| | Minimising Weights | Adjusted Weights | Even Spaced Weights | Centroid Weights |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | .894 |
| 2 | 2.69 | 2.69 | 5.263 | 14.507 |
| 3 | 5.38 | 5.38 | 10.526 | 23.633 |
| 4 | 13.08 | 13.08 | 17.789 | 29.406 |
| 5 | 11.64 | 17.00 | 21.052 | 33.566 |
| 6 | 21.54 | 21.54 | 26.316 | 35.573 |
| 7 | 39.49 | 39.49 | 31.579 | 38.677 |
| 8 | 39.71 | 39.71 | 36.842 | 44.679 |
| 9 | 48.31 | 48.31 | 42.105 | 51.466 |
| 10 | 51.54 | 51.54 | 49.368 | 56.019 |
| 11 | - - - | - - - | 52.631 | 56.019 |
| 12 | 62.82 | 62.82 | 57.894 | 58.916 |
| 13 | 72.58 | 63.00 | 63.157 | 61.192 |
| 14 | 63.95 | 63.95 | 68.420 | 63.862 |
| 15 | 80.51 | 80.51 | 73.683 | 65.807 |
| 16 | 63.69 | 81.50 | 78.946 | 67.670 |
| 17 | 83.25 | 83.25 | 84.210 | 70.277 |
| 18 | 84.00 | 84.00 | 89.473 | 75.492 |
| 19 | 89.35 | 89.35 | 94.736 | 84.721 |
| 20 | 100.00 | 100.00 | 100.000 | 100.000 |

The regression line, the standard error, and the correlation coefficient, are summarised in the following table according to the weights used.

Table VII

| | Regression Line | Standard Error | Correlation Coefficient |
|---|---|---|---|
| Minimizing Weights | y = .65x + 7.0 | 9.70 | .885 |
| Adjusted Weights | y = .643x + 7.108 | 9.97 | .879 |
| Even Spaced Weights | y = .636x + 7.907 | 10.31 | .866 |
| Centroid Weights | y = .865x + 5.588 | 10.56 | .859 |

It should be noted from the above table that the minimizing weights are such that even when adjusted to keep them in natural order they yield a regression with a significantly smaller standard error than in either of the other cases.

## CHAPTER V

## SOME UNANSWERED PROBLEMS

In the development of the weights above we have assumed that the "best" weights are those which minimize the standard error of estimate. However, this definition of "best" might very well be questioned. In reression problems one usually thinks of the data as being a sample of some population and the predicting equation one obtains from the data is used on new variates which were not present in the original problem. This procedure gives rise to the question, "Will the standard error of estimate, or the errors made in future estimations, be also minimized if the weights determined by minimizing the standard error for the known sample population is used?" Perhaps one should approach the whole problem from the point of view of bivariate distribution function rather than from the finite sample approach used in this paper. It appears that such an investigation will be needed to answer the above question.

In Chapter II several elaborate techniques of computing r as an estimation of the correlation coefficient from a non-quantitative bivariate sample were discussed. However in the development of a system of weights to be used in regression very little use was made of the information contained in this chapter. In fact all that was done was to make a comparison between the minimizing weights developed in Chapter III and the weights that have been used in the past. It seems that a more thorough study of the connection between these two problems could be made.

How does the r found from using the minimizing weights compare with those developed in estimating the correlation coefficient? Can the methods of successive approximations developed in Chapter II for adjusting r so as to make it a closer estimate of the correlation coefficient be applied in adjusting the weights for regression?

The problem considered in this paper could readily be enlarged. We have considered only the determination of weights for the independent variable, which we assume to be ordered but unmeasured. Using these weights we may obtain by regression the best estimate, in the sense of minimizing the standard error, for the dependent variable y which is measurable. A new problem arises if one considers the dependent variable as the non-measurable series while the independent variable is the measured series. Here the problem is not as straight-forward as the one considered in this paper, for one can make r equal to one, by simply giving the same weight to every class of the dependent variable. This, of course, does not help in the regression problem, as there would be no way of discriminating between the classes. What is apparently needed is a set of unequal weights, but how to vary these weights so as to obtain the best regression could not be determined by the technique used in this paper. One could further consider the problem of having both variables unmeasured and then the next step would be to consider more than two variables. Thus we see that this paper merely scratches the surface of a whole series of problems associated with nonmeasured variables in regression.

## CONCLUSION

In this paper we have shown that if one defines the best weights to be assigned to classes of a non-measured dependent variable as those which make the standard error of estimate a minimum, then these weights are dependent upon arbitrary weights assigned to two of the classes. The weight for the ith class is determined to be

$$t_i = \frac{t_2\bar{y}_1 - t_1\bar{y}_2 + t_1\bar{y}_i - t_2\bar{y}_i}{\bar{y}_1 - \bar{y}_2},$$

where $t_1$ and $t_2$ are the arbitrary weights, and $\bar{y}_i$ is the mean of the y's of the variates of the ith class. We have shown by means of an example that the weights so determined are different than those determined by methods that have been used in the past. In fact the differences between these sets of weights gives rise to the question of whether there is some better way to define "best" for determining the weights in this problem, and finally, how the methods used in this paper might be used to approach the many additional problems associated with using non-measured variables in regression.

# BIBLIOGRAPHY

Camp, Burton H., Elementary Statistics, Boston: D. C. Heath and Company, 1931.

Elderton, W. Palin, Frequency Curves and Correlation, 3rd edition, London: Cambridge University Press, 1938.

Ezekial, Mordecai, Methods of Correlation Analysis, New York: John Wiley and Sons, 1941.

Kenney, John F., Mathematics of Statistics, New York: D. Van Nostrand Company, 1939.

Kossack, Carl F., "Mathematics Placement at the University of Oregon", American Mathematical Monthly, XLIX, No. 4, April, 1942.

Pearson, Frank A., and Kenneth R. Bennet, Statistical Methods, New York: John Wiley and Sons, 1942.

Pearson, Karl, "Mathematical Contributions to the Theory of Evolution," Philosophical Transactions, Vol. 195, A, Feb., 1900, pp. 1-6.

_______, "On the Theory of Contingency," Draper's Company Research Memoirs, No. 1.

_______, Tables for Statisticians and Biometricians, London: Cambridge University Press, 1914, pp. 42-57.

_______, and Egon Pearson, "On Polychoric Coefficients of Correlation", Biometrika, XIV, pp. 127-156.

Ritchie-Scott, "The Correlation Coefficient of a Polychoric Table", Biometrika, XII, pp. 106-108.

# APPENDIX

## Table VIII

Table VIII

| | I | II | | III | IV | |
|---|---|---|---|---|---|---|
| | Elements of Algebra | Intermediate Algebra | | College Algebra | Introduction to Analysis | |
| F | 0 | 2 | 24 | | 24 | |
| | 4 | 10 | 26 | | 48 | |
| | 12 | 20 | 42 | | 48 | |
| | 12 | 20 | | | 56 | |
| | | 24 | | | 66 | |
| D | 2 | 14 | 42 | 25 | 62 | 68 |
| | 6 | 14 | 44 | 46 | 52 | |
| | 6 | 22 | 48 | 48 | 60 | |
| | 8 | 24 | 52 | 60 | 60 | |
| | 8 | 28 | | 58 | 62 | |
| | 12 | 32 | | 60 | 62 | |
| | 12 | 36 | | | 66 | |
| | 16 | 36 | | | 68 | |
| C | 2 | 14 | 32 | 42 | 34 | |
| | 2 | 16 | 34 | 42 | 50 | |
| | 2 | 18 | 34 | 44 | 56 | |
| | 3 | 18 | 36 | 46 | 58 | |
| | 4 | 20 | 38 | 48 | 60 | |
| | 8 | 22 | 38 | 50 | 60 | |
| | 8 | 24 | 38 | 56 | 62 | |
| | 10 | 24 | 38 | 60 | 62 | |
| | 12 | 26 | 40 | 64 | 62 | |
| | 14 | 26 | 42 | 68 | 64 | |
| | 14 | 28 | 44 | 76 | 64 | |
| | 16 | 28 | 44 | | 66 | |
| | 18 | 30 | 48 | | 72 | |
| | 34 | 30 | 50 | | 76 | |
| | | 32 | 52 | | 78 | |
| | | 32 | 54 | | | |
| B | 7 | 20 | 38 | 40 | 44 | 72 |
| | 8 | 26 | 40 | 42 | 60 | 74 |
| | 10 | 32 | 42 | 48 | 60 | 74 |
| | 10 | 32 | 42 | 48 | 60 | |
| | 12 | 34 | 42 | 50 | 60 | |
| | 16 | 34 | 44 | 54 | 62 | |
| | 16 | 34 | 44 | 58 | 68 | |
| | 20 | 36 | 46 | | 70 | |
| | 26 | 36 | 48 | | 70 | |
| | 30 | 36 | 62 | | 72 | |
| A | 6 | 12 | 44 | 50 | 50 | |
| | 9 | 24 | 46 | 52 | 70 | |
| | 12 | 24 | 46 | 58 | 70 | |
| | 16 | 32 | 48 | 58 | 84 | |
| | 20 | 32 | 48 | 66 | 86 | |
| | 24 | 44 | 50 | 72 | | |
| | | 44 | 56 | | | |
| | | 44 | 58 | | | |

typed by

BESSIE KAMARAD