UNITED STATES DEPARTMENT OF AGRICULTURE

Bureau of Agricultural Economics

CORRELATION THEORY AND METHOD APPLIED TO AGRICULTURAL RESEARCH

By

Bradford B. Smith

Washington, D. C.

August 1926

Table of Contents

Note: The use of a typewriter with standard keyboard for this manuscript has made necessary in a few cases the substitution of other characters than those commonly used. The summation sign Σ (Greek letter Sigma) is usually expressed by (S). The symbol for the correlation ratio η (Greek letter Eta) is expressed by (N). The addition sign is in many cases, expressed by the symbol (+). The sign for the standard deviation is expressed by the symbol (σ). No confusion should result from the use of these symbols, since they are used in the manuscript for no other purpose.

In the description of multiple correlation, the symbol representing a net regression coefficient is underlined ($\underline{b}$) to distinguish it from the symbol representing one of the independent varables (b).

CORRELATION THEORY AND METHOD APPLIED TO AGRICULTURAL RESEARCH

Collected and prepared for the use of
Statisticians of the Bureau of Agricultural Economics

by
Bradford B. Smith, Economic Analyst,
Division of Statistical and Historical Research.

Introduction

In no one volume are the theory and methods of correlation as applied to agricultural research in the Bureau of Agricultural Economics brought together in a form readily adapted to reference purposes. On the other hand not a few new statistical methods have been developed by members of the staff. This publication is an attempt to bring together and coordinate such methods. In order to be complete it presents usual correlation theory, but from a point of view easily adaptable to include the more recent theory and method. The approach is essentially that which has been given in the Graduate School of the Department of Agriculture for the past two years and is very similar in its treatment of simple correlation to that found in Prof. Frederick C. Mill's excellent text, "Statistical Methods Applied to Economics and Business". The new subject matter on multiple linear and curvilinear correlation, joint relationships, application to time series, and apportionment of importance to contributing variables is founded partly on articles published by members of the staff and partly on material as yet unpublished. Statisticians are especially indebted to H. R. Tolley and Mordecai Ezekiel of this Bureau for the notable contributions they have made to correlation methods, designated later in this work. Appreciation is also extended to E. M. Daggit of this Bureau for assistance in preparing the manuscript.

I THE FIELD OF CORRELATION.

The biologist studying effects of changing environmental factors on inheritance finds that one experiment does not always precisely verify the results of another; for influences other than the specific one under observation are at work, and change the apparent effect from time to time.

The entomologist, studying the effectiveness of practical methods of checking the inroads of pests, frequently has the true effect obscured by the interference of unmeasured factors. This prevents his establishing the true quantitative relations.

The economist---or student in the social sciences--- is especially at a disadvantage in either detecting or demonstrating the social laws for it is practically impossible to secure experimental conditions when dealing with mass human reactions: conditions where the influence of all factors but the ones under consideration are held constant throughout the recording of the data. There may for example be a perfectly simple theoretical relation between the price, p, and the supply, s, of a commodity, such as $p = \frac{1}{a+b.s}$. But in practice this is difficult to demonstrate, for other factors influencing the price, such as changing demand, quality of the product and the value of gold, through their changes obscure the relation. Thus that which appears to be the quantitative relation in one case, is modified in the second, the third and so on. What

then is the true relation? Correlation method is but a tool for learning from such variable data, taken as a whole, what the most probable relation is. This method has its foundation in the theory of probability. In addition to giving a means of ascertaining what the most probable relation is, it gives an idea as to how closely this most probable relation is fulfilled.

Correlation methods are evidently not needed for experiments in the natural sciences such as chemistry and physics where by exact standardization under experimental conditions the effect of a given cause can be demonstrated precisely time after time. In those undertakings, however, where we are unable to ascertain the precise relation existing between one factor and another since other factors exert a variable influence, it becomes necessary to draw out from the mass of often conflicting data some conclusion which we may say represents on the whole the most probable relation. The correlation methods are the most effective tools yet devised for dealing with such situations. They owe their origination largely to the English School of Biometricians, pre-eminently Karl Pearson. Recently there have been a number of texts applying these methods to the problems of the social sciences as distinguished from the biological. Each writer has his own particular approach to the problem. The approach presented here is based on the belief that simple correlation--the correlation of two variables--is but

the beginning and the least effective of all the correlation methods for handling economic agricultural problems. Therefore any treatment of simple correlation should be such as is easily and logically expansible to include multiple and partial correlation, as well as curvilinear correlation. Nearly every important economic research study dealing with quantitative aspects is appreciably more satisfactory when such methods are used. This is because nearly every such study is inevitably confronted with the necessity for taking into account the influence of several related factors in detecting, demonstrating or applying economic law. The only known method of doing this is the multiple correlation method. This should not be interpreted to mean that the correlation methods are by any means sufficient. They have definite and serious limitations; so much so, that it is almost safe to say that they are misused more often than not. However, the methods are improving continually with each contribution.

II GROSS CORRELATION

1. Regression.

Suppose that a series X, is associated with a series Y, so that each value of X has a corresponding value of Y. These may for example, represent the price and the quantity of melons sold on different days in a central market; n may represent the number of pairs of observations.

The problem of studying the relation between the two series may be divided into two parts: (1) to measure how great the divergence in X (price) from its average is associated with a unit divergence in Y (quantity); and (2) to obtain some idea as to how closely the relation is fulfilled.

Deviations from the averages are taken because frequently a direct proportion may be discovered between such deviations which is not apparent between the original values. The two following series (X and Y) serve to illustrate:

X	X-av	: Y	Y-av.
10	-4	: 8	+4
12	-2	: 6	+2
14	0	: 4	0
16	+2	: 2	-2
18	+4	: 0	-4

Evidently the relation between deviations in x and in y may be expressed by the factor, -1.0 We may say:

$$X-M_x = -1 \ (Y-M_y) \ . \ . \ . \ . \ . \ . \ (1)$$

where M stands for the respective arithmetic averages. The

relation between the original values cannot be otherwise expressed On the other hand if there should be a direct relation between X and Y as shown in the following table, where X = 2Y, the deviations from average will show the relation quite as successfully as if the ratio of the original items was taken:

X	X-mx	:	Y	Y-my
2	-4	:	4	-8
4	-2	:	8	-4
6	0	:	12	0
8	+2	:	16	+4
10	+4	:	20	+8

Since there is something to be gained and nothing to be lost, therefore, the attempt should be to establish the proportional relation between deviations rather than between original items. The first problem can now be stated more specifically: we wish to discover the value of b in expressions of the type:

$$X-M_x = b(Y-M_y) \qquad \ldots (2)$$

or, letting x and y represent the deviations from average of X and Y respectively:

$$x = by \qquad \ldots (3)$$

which may also be written, of course,

$$X = M_x - bM_y + bY \qquad \ldots (4)$$

or, since b M_y will be a constant, more briefly

$$X = K + bY \qquad \ldots (5)$$

which is the familiar expression of the formula for a straight line on graph paper of coordinates, X and Y.

Since X is determined by Y, in this set-up X may be

termed the dependent variable and Y the independent. If

$$Y = K+bX \qquad \ldots (6)$$

were written, Y would be the dependent and X the independent.

Since the first task cited above is to find how great a divergence from average in X is associated with a unit divergence in Y, a value of b in formula (3) should be found so that the equation would be true for every pair of observations. The capital letters X and Y are taken to represent original values, while the small letters x and y represent deviations.

Taking numerical data as shown in table 1, and considering the first case, it is possible to solve for a value of b, for if $x = b.y$, then since $x = 0$ and $y = -1$, b must be equal to 0.

In like manner b is found to equal 0 in the second case; but in the third observation b is found to equal -.5; in the fourth, infinity.

Evidently there is no one value of b which will satisfy perfectly for all the observations the formula

$$x = by$$

and it is only in the most unusual circumstances that series will be found where a single value of b will satisfy perfectly for all observations the requirement, $x = b.y$.

Logically, therefore, the value of b which will come the nearest to satisfying all the observations,- -the "most probable value" of b, should be found, and described as the probable relation existing between the deviations in the two series.

In the application of the laws of probability a general procedure for finding such values has been evolved. This is the method of least squares. 1/ It assumes that the criterion of "best value" or "best fit" is that the sum of the squared residuals shall be a minimum. The meaning of this can be readily understood if a dot chart or Galton graph such as shown in Figure 1 is made, where the ordinates are the values of x and the abscissae the values of y. Each dot, when properly located, represents an x and a y value by the distances from the x and y axes.

Because x = b.y (see formula 3), the value of x will be zero when the value of y is zero; hence any line drawn to represent the relation of x to y must pass through the origin - where x=0 and y=0.

A line drawn diagonally from the upper left corner to the lower right will obviously come nearer passing through most of the dots representing data from Table 1 than will a horizontal line. Should such a line be drawn on the graph and the positive and negative vertical distances between each dot and the line measured, the measurements found would represent the "residuals." The value of b defines the slope of the line, for b, according to the formula, x = b.y, is nothing but the distance covered in the x direction when there is a unit change in the y direction. This is true no matter what portion of the line is considered, so the line must necessarily be a straight line. Relationships producing such lines are therefore described as a linear, to distinguish them from relationships producing a curved line, which are described as curvilinear. In the above example, b must be a negative value, since whenever y in-

1/ Merriman, Mansfield. Elements of the Method of Least Squares. London, 1877.

creases in value, x decreases. The criterion of "best fit"--that is, that the sum of the residuals squared must be a minimum--means merely that the value, b, or the slope of the line on the chart must be such that the sum of the squared vertical distances between the dots and the line will be a minimum. This then is the "most probable value" of b. The finding of b will be the accomplishment of the first portion of the problem.

In order to secure the values of the differences, or residuals, one must know the ordinate values of the points on the line to substract from the known values of x. But for any given y value this value will be b.y, which may be termed x', since it will be the value of x secured if it were estimated from its described relation to y as shown by the line. Letting z represent any residual, therefore, $z = x-x'$ or

$$z = x-by \qquad \ldots (7)$$

And the best value of b in the relation, $x = b.y$, will be such that Sz^2 will be a minimum. 2/ The value, b, in correlation terminology is the "regression coefficient" or the "regression of X on Y." It is the amount of change in X associated with a unit change in Y. It may be distinguished from other regression coefficients by subscripts, the initial subscript indicating the dependent. The subscripts will change in order according to whether the regression of X on Y, or the regression of Y on X, is being described. Thus

---

2/ The designation "S" before a term may be taken to mean the sum of the terms like that before which it stands. It is frequently expressed by the Greek letter Sigma, "Σ".

$$x = b_{xy}y \qquad \text{and}$$

$$y = b_{yx}x \qquad \ldots (8)$$

To find the value of b_{xy} by methods of least squares, multiply each observation equation through by the coefficient of b_{xy} in that equation: (subscripts to x and y designate the observation number)

Observation:	Coeff. of b_{xy} :	Extension
$x_1 = b_{xy}y_1$:	y_1 :	$x_1y_1 = b_{xy}y_1^2$
$x_2 = b_{xy}y_2$:	y_2 :	$x_2y_2 = b_{xy}y_2^2$
$x_n = b_{xy}y_n$:	y_n :	$x_ny_n = b_{xy}y_n^2$
:	:	$Sxy = b_{xy}Sy^2$

Summing the product equations so secured, gives

$$Sxy = b_{xy}\, Sy^2 \qquad \ldots (9)\ a$$

The values Sxy and Sy^2 may be conveniently secured by some such arrangement as shown in table 1. Equation (9 a) or, for the example, $-244 = b_{xy}$ 278, is called a "normal equation" and its solution gives the value of b_{xy} :

$$b_{xy} = \frac{Sxy}{Sy^2} = \frac{-244}{278} = -.878$$

The value, Sxy/n or p_{xy}, is termed the product moment. Thus $b_{xy} = p_{xy}/\sigma_y^2$ which is perhaps a more convenient way of defining b_{xy}.

Having found the value of b_{xy} it can immediately be inserted in the equation and the equation written,

$$x' = b_{xy}y \qquad \ldots (9)$$

or $\qquad x' = -.878y$

Equation (9) is termed the "regression equation" in correlation terminology. The values of x' and y may easily be expressed in terms of original items instead of deviations from average by writing in the equivalents of x and y.

$$X' - M_{x'} = b_{xy}(Y - M_y) \qquad \ldots (10)$$

or

$$X' = M_{x'} - b_{xy}M_y + b_{xy}Y \qquad \ldots (11)$$

M_x may be substituted for $M_{x'}$ since $Sx' = Sx = 0$, thus leaving the means of the X' and X series identical.

The first object has now been accomplished: to measure how great a divergence in the dependent is probably associated with a unit divergence in the independent. A value of b_{xy} has been secured which, if applied to each value of y, will give products, x', that will most nearly equal the associated values of x; the sum of the values, $z=x-x'$, squared, will be less than if any other value of b_{xy} should be used.

As stated previously, the line on the graph, $x' = b_{xy}Y$, must pass through the origin, where both x and y are zero; and letting y equal any other value, x' may be determined; the plotting of this point then determines the line. In correlation terminology this line is called the regression line.

2. Correlation.

The second problem as stated in the beginning, is to secure some idea as to how truly the described relation is fulfilled, or as can be said now, how closely the x' values coincide with the x values, or in another way, how nearly the x values come to being on the regression line. An inspection of the graph, Figure 1, gives some idea, and prompts the thought that a numerical measure of the agreement might be obtained by averaging the residuals. Accordingly:

$$\text{Each} \qquad z = x - x'$$

$$= x - b_{xy}y$$

$$\text{Therefore} \qquad Sz = Sx - Sy \cdot b_{xy} \qquad \ldots \quad (12)$$

But both Sx and Sy are zero, hence Sz is zero, which means that only by disregarding the signs of z could such a numerical measure be secured. This, however, would prohibit any further rigid mathematical treatment of the measure. So, just as the signs are eliminated in securing standard deviation by squaring the differences, so the signs of the residuals may all be made positive by squaring them. In other words, the standard deviation of the residuals may be taken as an inverse measure of the closeness of fit of the line. This standard deviation, σ_z, is evidently $\sqrt{Sz^2/n}$ since the mean of the series is zero. In correlation terminology it is called the "standard error of estimate," and is sometimes symbolized, s_{xy}.

But taking the standard deviation of the residuals is not enough, for this value can change according to the value of the scale or unit used. For comparative purposes a measurement more on the order of a coefficient that will not change with changes in units or scales is needed. To do this σ_z should be expressed in terms of some form of the series, X, and since the standard deviation of z is being taken, what would be more logical than to take the standard deviation of the original series? Let the mathematical relation existing between them accordingly be determined.

$$\text{Each } z = x - x'$$

$$" \quad z^2 = x^2 - 2xx' + x'^2$$

$$\therefore \frac{Sz^2}{n} = \frac{Sx^2}{n} - 2\frac{Sxx'}{n} + \frac{Sx'^2}{n} \qquad \ldots (13)$$

But Sxx'/n, or the product moment, $p_{xx'}$, of x and the estimates of x from the regression equation, may be shown to equal Sx'^2/n:

$$\text{Each } xx' = x \cdot b\, y$$

$$= x\left(\frac{Sxy}{Sy^2}\right)y$$

$$\therefore Sxx' = Sxy\,\frac{Sxy}{Sy^2} \qquad \ldots (14)$$

$$\text{Also, each } x'^2 = b^2y^2$$

$$= \frac{Sxy\ Sxy}{Sy^2Sy^2}\ y^2$$

$$\therefore Sx'^2 = Sxy\,\frac{Sxy}{Sy^2} \qquad \ldots (15)$$

Cancelling in equation 13

$$Sz^2/n = Sx^2/n - Sx'^2/n \qquad \ldots (16)$$

The mean of the x' values in zero, so Sx'^2/n is the standard deviation squared of the x' values, the estimates of x secured by

Multiplying each y by b_{xy}, all of which values lie on the regression line. Accordingly,

$$\sigma_z^2 = \sigma_x^2 - \sigma_{x'}^2 \quad \ldots (17)$$

or $$\sigma_{x'}^2 = \sigma_x^2 - \sigma_z^2 \quad \ldots (18)$$

These are very important formulae. The meaning of the relations shown may be stated thus: The proportion of the total squared variability in the dependent that is due to y as expressed by x' values may be expressed by the fraction $\sigma_{x'}^2 / \sigma_x^2$ and the proportion of the total squared variability in x not explained by y may be expressed by the fraction σ_z^2/σ_x^2. The expression $\sigma_{x'}^2 / \sigma_x^2$ may be termed a coefficient of determination, d_{xy}, and σ_z^2 / σ_x^2 is the proportion of unaccounted-for squared variability. The sum $d_{xy}+d_{xz}$ must always equal unity, and since the two represent derivations from sums of squares they are always positive in sign.

If we take the square roots of these measures, getting them back to coefficients of the first order, we have $\sigma_{x'}/\sigma_x$ and σ_z/σ_x. The expression σ_z/σ_x is the ratio of the standard deviation of the residuals to the standard deviation of the original x series. Whenever this value becomes large it means that there is very little relation between the x series and the y series, and when small it means that the standard deviation of residuals compared to the original standard deviation of x is small, and hence that the values of $b_{xy}y$ approximate closely to the original values of x. Since it varies inversely with the closeness of the relation between x and y it is termed the coefficient of alienation. The coefficient $\sigma_{x'}/\sigma_x$ or r_{xy}, since it varies directly with the closeness of the relation

between x and y, is termed the Pearsonian Coefficient of Correlation, after its originator.

If the value of b_{xy} is negative the coefficient of correlation is said to be negative and is always preceded by a negative sign. Its meaning in that event is that the dependent becomes smaller in value as the independent becomes larger, as is the case in the example, Table I, and Figure 1.

From the formula, derived from 18,

$$\frac{\sigma_z^2}{\sigma_x^2} \div \frac{\sigma_{x'}^2}{\sigma_x^2} = 1 \quad \ldots\ldots (19)$$

It is apparent that neither coefficient can ever be greater than plus or minus 1.00. If $\frac{\sigma_{x'}}{\sigma_x}$ is either plus or minus 1.00 (and $\frac{\sigma_z}{\sigma_x}$ therefore zero), it means that a perfect interpretation of x can be made from y, i.e., that all observed values of x, when plotted, lie on the regression line, $x' = b_{xy}y$; conversely, if $\frac{\sigma_{x'}}{\sigma_x}$ is zero it means that the standard deviation of the estimates is zero, therefore that the regression line is horizontal, and furthermore that the value of b_{xy} is zero. It means, thus, either (1) that there is no relation between x and y, or else (2) if there is any relation, it can not be expressed by a straight line. Many students are prone to explain low values of r by the first reason, whereas there may be a very high degree of relation shown when the proper curve is substituted for the straight line. It is customary, however, to speak of two variables as uncorrelated when r is very small; but it must not be forgotten that the second meaning given above may also be the explanation of low values of r.

The formula for the usual method of computing (and defining) r may be derived as follows:

From (15), $\overline{\sigma x'}^2 = \frac{pxy^2}{\overline{\sigma y}^2}$

And $\overline{\sigma x'} = \sqrt{\frac{pxy^2}{\overline{\sigma y}^2}} = \frac{pxy}{\overline{\sigma y}}$ .... (22)

Dividing by $\overline{\sigma x}$ to secure r_{xy}

$$r_{xy} = \frac{\overline{\sigma x'}}{\overline{\sigma_x}} = \frac{pxy}{\overline{\sigma x}\overline{\sigma y}} \quad \text{or} \quad \frac{\Sigma_{xy}}{n\overline{\sigma x}\overline{\sigma y}} \qquad \ldots\ldots (23)$$

which is the usual definition of the Pearsonian Coefficient of Correlation.

If y is made the dependent and the value of b_{yx} (note change in order of subscripts) solved for in

$$y = byx^x \qquad \text{then}$$

$$r_{yx} = \frac{\overline{\sigma y'}}{\overline{\sigma y}} \qquad \ldots\ldots (24)$$

This value, r_{yx} may also be shown to equal $pxy/\overline{\sigma x}\ \overline{\sigma y}$ and hence is equal to r_{xy}. The order of writing the subscripts is therefore unimportant. This is not the case with the regression coefficients, however, since they represent two different lines on the graph. b_{yx} is graphed on Figure 1 as a broken line.

Since $b_{xy} = \frac{p}{\overline{\sigma y}^2}$ it takes its sign from the sign of p as will the coefficient of correlation when computed from formula 23. Both regressions are evidently of the same algebraic sign with the coefficient of correlation.

The coefficient b_{xy} may be expressed in terms of r, σ_x and σ_y as follows:

$$b_{xy} = \frac{p}{\sigma_y^2} = \frac{p}{(\sigma_x \sigma_y)} \cdot \frac{\sigma_x}{\sigma_y} = r \frac{\sigma_x}{\sigma_y} \quad \ldots.(25)$$

from which the regression may be written in the form:

$$x = r \frac{\sigma_x}{\sigma_y} y \quad \ldots.(26)$$

or

$$\frac{x}{\sigma_x} = r \frac{y}{\sigma_y} \quad \ldots.(27)$$

Similarly the other regression equation (8) may be written

$$\frac{y}{\sigma_y} = r \frac{x}{\sigma_x} \quad \ldots.(28)$$

The inference to be drawn from these formulae gives an additional meaning to the conception of the coefficient of correlation, i.e., when the variations in the series are reduced to comparable denominators, their standard deviation, r expresses completely the relation between them--both the amount of change in each associated with unit changes in the other and also the degree to which this relation may be expected to be maintained.

An interesting relation useful as a check on computations may be shown to exist between r, b_{xy} and b_{yx} :

As shown above (25)

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \text{ , and } b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

Multiplying the two equations together gives

$$b_{xy} . b_{yx} = r \frac{\sigma_x}{\sigma_y} . r \frac{\sigma_y}{\sigma_x}$$

or $$r = \sqrt{b_{xy} . b_{yx}} \quad \ldots (28)\ a$$

3. Summary of formulae

I $\sigma_x^2 = \sigma_{x'}^2 + \sigma_z^2$

II $\sigma_{x'}^2/\sigma_x^2 + \sigma_z^2/\sigma_x^2 = 1$

III $d_{xz} = \sigma_z^2/\sigma_x^2$

IV $d_{xy} = \sigma_{x'}^2/\sigma_x^2$

V $d_{xz} + d_{xy} = 1$

VI $r_{xy} = r_{yx}$

VII $= \sqrt{d_{xy}}$

VIII $= \sqrt{d_{yx}}$

IX $= \sqrt{b_{xy}\, b_{yx}}$

X $= \sigma_{x'}/\sigma_x$

XI $= \sigma_{y'}/\sigma_y$

XII $= \dfrac{n}{\sigma_x \sigma_y}$

XIII $b_{xy} = \dfrac{n}{\sigma_y^2}$

XIV $= r\, \dfrac{\sigma_x}{\sigma_y}$

XV $\dfrac{x}{\sigma_x} = r\, \dfrac{y}{\sigma_y}$

XVI $\dfrac{y}{\sigma_x} = r\, \dfrac{x}{\sigma_y}$

XVII $x' = b_{xy} y$

XVIII $y' = b_{yx} x$

XIX $X' = Mx - My\, b_{xy} + b_{xy}\, Y$

XX $\dfrac{Sxx'}{n} = p_{xx'}$

XXI $= \sigma_{x'}^2$

XXII $\dfrac{Syy'}{n} = p_{yy'}$

XXIII $= \sigma_{y'}^2$

XXIV $s_{xy} = \sigma_z = \sigma_x \sqrt{1 - r_{xy}^2}$

XXV $r_{xy} = \sqrt{1 - \dfrac{s_{xy}^2}{\sigma_x^2}}$

Summary--Cont'd

XXVI $= \sqrt{1 - d_{xz}}$

XXVII $\frac{S_{zy}}{n} = p_{zy}$

XXVIII $= 0$

XXIX $r_{zy} = 0$

XXX $r_{zx} = \sqrt{d_{x}}$

XXXI $= \sigma_{z}/\sigma_{x}$

XXXII $\sigma_{x} = \sqrt{\frac{S_{x}^2}{n}}$

XXXIII $p_{xy} = \frac{S_{xy}}{n}$

XXXIV $r_{x'y} = 1.00$

Note: It is useful to observe that adding a constant to each item in a series in no wise affects the correlation, for since every item is changed similarly the deviations remain unchanged. Multiplying by a constant does not change the correlation, for numerator and denominator in $r = p/\sigma_x\sigma_y$ are changed proportionally. The regressions change, however.

Coefficients of regression, alienation, correlation and determination have been secured, which provide fairly complete tools with which to meet the two problems outlined at the beginning of this discussion. This, however, is all based on the assumption that relationships are linear. In the event that such be not the case; it leads to the need

for modification of these measures. The correlation ratio, and the correlation index, are such modifications. Their formulae and meaning will subsequently be developed along substantially the same lines as were those of the measures described in this section.

4 Arithmetic Methods.

Some formulae useful in reducing labor of calculating the various coefficients discussed in the previous sections may be derived.

An inspection of the formulae for all of these coefficients shows that they may be computed from three values, $O_{\bar{x}}$, $O_{\bar{y}}$ and p_{xy}. The computations involved, once these values are secured, are but slight. The tedious part of computing the coefficients is in securing these three values. In treating the original series the first step involved was to secure the deviations, x and y, of each X and Y value from their respective means; and since the differences so secured are more apt than not to be fractions, the arithmetic of squaring and multiplying them together is extended. Formulae for securing the values of Sxy/n, $O_{\bar{x}}$ and $O_{\bar{y}}$ may be easily developed which eliminate the necessity of using these deviations. They represent a marked saving of labor, especially for those who have access to any of the standard computing machines.

The standard deviation of X may be expressed as follows:

$$\sigma_x^2 = \frac{1}{n} (Sx^2)$$

$$= \frac{1}{n} S (X-M_x)^2$$

$$= \frac{1}{n} (S X^2 - 2 M_x S X + n M_x^2)$$

$$= \left(\frac{SX^2}{n} - 2 \frac{SX}{n} \frac{SX}{n} + \frac{SX}{n} \frac{SX}{n}\right)$$

$$= \frac{SX^2}{n} - M_x^2$$

$$\sigma x = \sqrt{\frac{SX^2}{n} - M_x^2} \qquad \ldots\ldots (29)$$

The use of this formula obviates the need of taking deviations in securing the standard deviation of X. Substituting Y for X in the formula, of course, gives an expression for the standard deviation of Y.

Any whole number approximating the mean may be used as an assumed mean from which to secure deviations and a subsequent correction will give the true standard deviation. This method means handling smaller values than used in formula 29 and yet eliminates the fractional difference.

Let $M_x + e = \bar{M}_x$, the assumed mean, differing from the true, M_x, by the amount, e.

Then each deviation, $\bar{x}$, secured from the assumed mean is described

$$\bar{x} = X-\bar{M}_x$$

$$= X-M_x-e$$

and hence, each $\bar{x} = x-e$

where x is the true deviation, $X-M_x$

Then each $\bar{x}^2 = x^2 - 2xe+e^2$

and $S\bar{x}^2 = Sx^2 - 2eSx + ne^2$

But Sx is by definition equal to zero.

Therefore $S\bar{x}^2 = Sx^2 + ne^2$

and $\frac{S\bar{x}^2}{n} = \sigma_{\bar{x}}^2 + e^2$

or $\sigma_x^2 = \frac{S\bar{x}^2}{n} - e^2$ .....(30)

That is, the true standard deviation squared is secured by subtracting from the mean square of the differences from the assumed average the square of the difference between the true and assumed averages. It might be noted that the formula, (29) represents a special case of formula (30), for in the former the assumed mean is zero and hence the correction to be made is the full value of the mean squared. It is helpful to note, also, that the correction is always a subtraction whether the value of e is plus or minus, since the squaring takes out the negative signs. A convenient assumed mean is one which is a multiple of ten and of approximately the value of the smallest item in the series. It is very easy to take the differences then, and there are practically no negative quantities to confuse the computation of the product moment.

Just as the standard deviations may be defined in terms of the original values, so also may the product moment:

Each $xy = (X - M_x)(Y - M_y)$

$= XY - M_y X - M_x Y + M_x M_y$

$Sxy/n = SXY/n - M\ SX/n - M\ SY/n + \frac{nM_x M_y}{n}$

$= SXY/n - M_x M_y$ .....(31)

Or if it is desired to use assumed means,

$\bar{M}_x = M_x + c_x$

$\bar{M}_y = M_y + e_y$

Then each $\bar{x} = x - e_x$

and each $\bar{y} = y - e_y$

Then each $\bar{x}\,\bar{y} = xy - xe_y - ye_x + c_x c_y$

and $S\bar{x}\,\bar{y}/n = \frac{Sxy}{n} - e_y \frac{Sx}{n} - e_x \frac{Sy}{n} + e_x e_y$

or $p_{xy} = Sxy/n = S\bar{x}\,\bar{y}/n - e_x e_y$ .....(32)

It should be noted that the correction, $e_x e_y$, may be either positive or negative; care should accordingly be given to its sign in determining p_{xy} by this method. The most convenient method of listing the arithmetic involved in computation of p, σ_x and σ_y by the above formulae is shown in the tabular form given below.

Observation Number	X or x	Y or y	X^2 or x^2	Y^2 or y^2	XY or xy
1	x_1	y_1	x_1^2	y_1^2	x_1y_1
2	x_2	y_2	x_2^2	y_2^2	x_2y_2
3	'	'	'	'	' '
'	'	'	'	'	' '
'	'	'	'	'	' '
'	'	'	'	'	' '
n	x_n	y_n	x_n^2	y_n^2	x_ny_n
Sums	Sx	Sy	Sx^2	Sy^2	Sxy
Means	Mx	My	Mx^2	My^2	Mxy
Corrections			$-M_x^2$	$-M_y^2$	$-M_xM_y$
Sum to secure . . .	. . . .		σ_x^2	σ_y^2	Pxy

In the event that computing machines such as the Monroe are used in which it is possible to cumulate the squares and products, the values, $S\,X^2$, $S\,Y^2$, and $S\,XY$ may be secured without listing them. Thus, put the value of X_1 into the keyboard and extend the product, X_1^2 into the register. Without clearing the register extend X_2^2 and so on. $S\,X^2$ may be read from the register after the last extension has been made. If, when using the Monroe, the digit, "1", in the left row of the key-board is locked down throughout the securing of $S\,X^2$, the value, S X, will appear in the left of the register. This either saves the adding of X as a separate operation, or partially check the extensions in securing $S\,X^2$.

Table 1

Record No.	Independent, Y			Dependent, X			y^2	x^2	xy
	Y	M_y	$y=Y-M_y$	X	M_x	$x=X-M_x$			
1	18	19	-1	10	10	0	1	0	0
2	18	19	-1	10	10	0	1	0	0
3	17	19	-2	11	10	+1	4	1	-2
4	19	19	0	11	10	+1	0	1	0
5	21	19	+2	9	10	-1	4	1	-2
6	20	19	+1	8	10	-2	1	4	-2
7	16	19	-3	12	10	+2	9	4	-6
8	20	19	+1	8	10	-2	1	4	-2
9	22	19	+3	8	10	-2	9	4	-6
10	20	19	+1	10	10	0	1	0	0
11	19	19	0	9	10	-1	0	1	0
12	23	19	+4	7	10	-3	16	9	-12
13	5	19	-14	23	10	+13	196	169	-182
14	21	19	+2	7	10	-3	4	9	-6
15	18	19	-1	12	10	+2	1	4	-2
16	24	19	+5	6	10	-4	25	16	-20
17	20	19	+1	10	10	0	1	0	0
18	19	19	0	9	10	-1	0	1	0
19	19	19	0	11	10	+1	0	1	0
20	21	19	+2	9	10	-1	4	1	-2
n=20	380	-	0	200	-	0	278	230	-244
Means	19	-	0	10	-	0	13.9	11.5	-12.2
Standard deviations							3.73	3.39	
Product moment									-12.2

Double frequency table

Table 2

X		Y										Frequencies summed
Classes	Assumed Values	300-304	305-309	310-314	315-319	320-324	325-329	330-334	335-339	340-344	345-	
		Assumed values										
		-5	-4	-3	-2	-1	0	1	2	3	4	
35-	+4							1				1
33-34	+3			2	1	2	1					6
31-32	+2	2	1	3	17	19	17		3			62
29-30	+1		1	8	28	56	42	13		2		150
27-28	0		1		12	15	70	25	14			137
25-26	-1			4	5	10	25	56	28	7		135
23-24	-2				1	3		20	20	8	1	53
21-22	-3						3	15	11	5	2	36
19-20	-4						1	2		1		4
17-18	-5											0
Frequencies summed		2	3	17	64	105	159	132	76	23	3	584

TABLE 4. COMPUTATION OF $\sigma_{\bar{y}}$

Y Values (assumed) $\bar{y}$	$\bar{y}^2$	Frequency f	$f\bar{y}$	$f\bar{y}^2$
- 5	25	2	-10	50
- 4	16	3	-12	48
- 3	9	17	-51	153
- 2	4	64	-128	256
- 1	1	105	-105	105
0	0	159	0	0
÷ 1	1	132	÷132	132
÷ 2	4	76	÷152	304
÷ 3	9	23	÷ 69	207
÷ 4	16	3	÷ 12	48
Sums		584	÷ 59	1303
Means			÷.101	2.23
$M_{\bar{y}}^2$				.01
$\sigma_{\bar{y}}^2$				2.22
Square root = $\sigma_{\bar{y}}$ =				1.490
$\sigma_{\bar{y}} \cdot C_y = \sigma_y$; C_y =				5
σ_y =				7.450

TABLE 3. COMPUTATIONS OF $\sigma_{\bar{x}}$

X values (assumed) $\bar{x}$	$\bar{x}^2$	Frequency f	$f\bar{x}$	$f\bar{x}^2$
+ 4	16	1	+ 4	16
+ 3	9	6	+ 18	54
+ 2	4	62	+124	248
+ 1	1	150	+150	150
0	0	137	0	0
- 1	1	135	-135	135
- 2	4	53	-106	212
- 3	9	36	-108	324
- 4	16	4	- 16	64
- 5	25	0	- 0	0
Sums	-	584	- 69	1203
Means			-.1181	2.06
$M_{\bar{x}}$ squared				.01
$\sigma_{\bar{x}}^2$				2.05
Square root = $\sigma_{\bar{x}}$				1.432
$\sigma_{\bar{x}}$ x C_x = $\sigma_{\bar{x}}$;			C_x =	2
Product = $\sigma_{\bar{x}}$ =				2.864

TABLE 5. COMPUTATION OF P_{xy}

Assumed values; + Products					Assumed values; - Products				
$\bar{x}$	$\bar{y}$	$\bar{x}\bar{y}$	f	$f\bar{x}\bar{y}$	$\bar{x}$	$\bar{y}$	$\bar{x}\bar{y}$	f	$f\bar{x}\bar{y}$
+ 4	+ 1	+ 4	1	+ 4	+ 3	- 3	- 9	2	-18
+ 2	+ 2	+ 4	3	+ 12	+ 3	- 2	- 6	1	- 6
+ 1	+ 1	+ 1	13	+ 13	+ 3	- 1	- 3	2	- 6
+ 1	+ 3	+ 3	2	+ 6	+ 2	- 5	- 10	2	-20
- 1	- 3	+ 3	4	+ 12	+ 2	- 4	- 8	1	- 8
- 1	- 2	+ 2	5	+ 10	+ 2	- 2	- 4	17	-68
- 1	- 1	+ 1	10	+ 10	+ 2	- 1	- 2	19	-38
- 2	- 2	+ 4	1	+ 4	+ 1	- 4	- 4	1	- 4
- 2	- 1	+ 2	3	+ 6	+ 1	- 3	- 3	8	-24
					+ 1	- 2	- 2	28	-56
					+	- 1	- 1	56	-56
					- 1	+ 1	- 1	56	-56
					- 1	+ 2	- 2	28	-56
					- 1	+ 3	- 3	7	-21
					- 2	+ 1	- 2	20	-40
					- 2	+ 2	- 4	20	-80
					- 2	+ 3	- 6	8	-48
					- 2	+ 4	- 8	1	- 8
					- 3	+ 1	- 3	15	-45
					- 3	+ 2	- 6	11	-66
					- 3	+ 3	- 9	5	-45
					- 3	+ 4	- 12	2	-24
					- 4	+ 1	- 4	2	- 8
					- 4	+ 3	- 12	1	-12
Sums				+ 77					-813
									77
									-716
	Divide by n (584) gives mean								- 1.226
	Subtract $m_{\bar{x}}$ $m_{\bar{y}}$ (-.118) (.101)								+ .012
									- 1.214

$P_{xy} = C_x C_y P_{\bar{x}\bar{y}} = -1.214 \times (5.2)$

-1.214

$\times \quad 10$

$P_{xy} = -12.14$

TABLE 6. COMPUTATION OF MEAN OF COLUMNS

Column: (Type)	Σf	$\Sigma f \bar{x}$	$\frac{\Sigma f \bar{x}}{\Sigma f} = M\bar{x}$
$\bar{y} = -5$	2	+ 4	+ 2.00
$\bar{y} = -4$	3	+ 3	+ 1.00
$\bar{y} = -3$	17	+ 16	+ .94
$\bar{y} = -2$	64	+ 58	+ .91
$\bar{y} = -1$	105	+ 84	+ .80
$\bar{y} = 0$	159	+ 41	+ .26
$\bar{y} = +1$	132	- 132	- 1.00
$\bar{y} = +2$	76	- 95	- 1.25
$\bar{y} = +3$	23	- 40	- 1.74
$\bar{y} = +4$	3	- 8	- 2.67

Check against $\Sigma f \bar{x}$ in $O\bar{x}$:-69

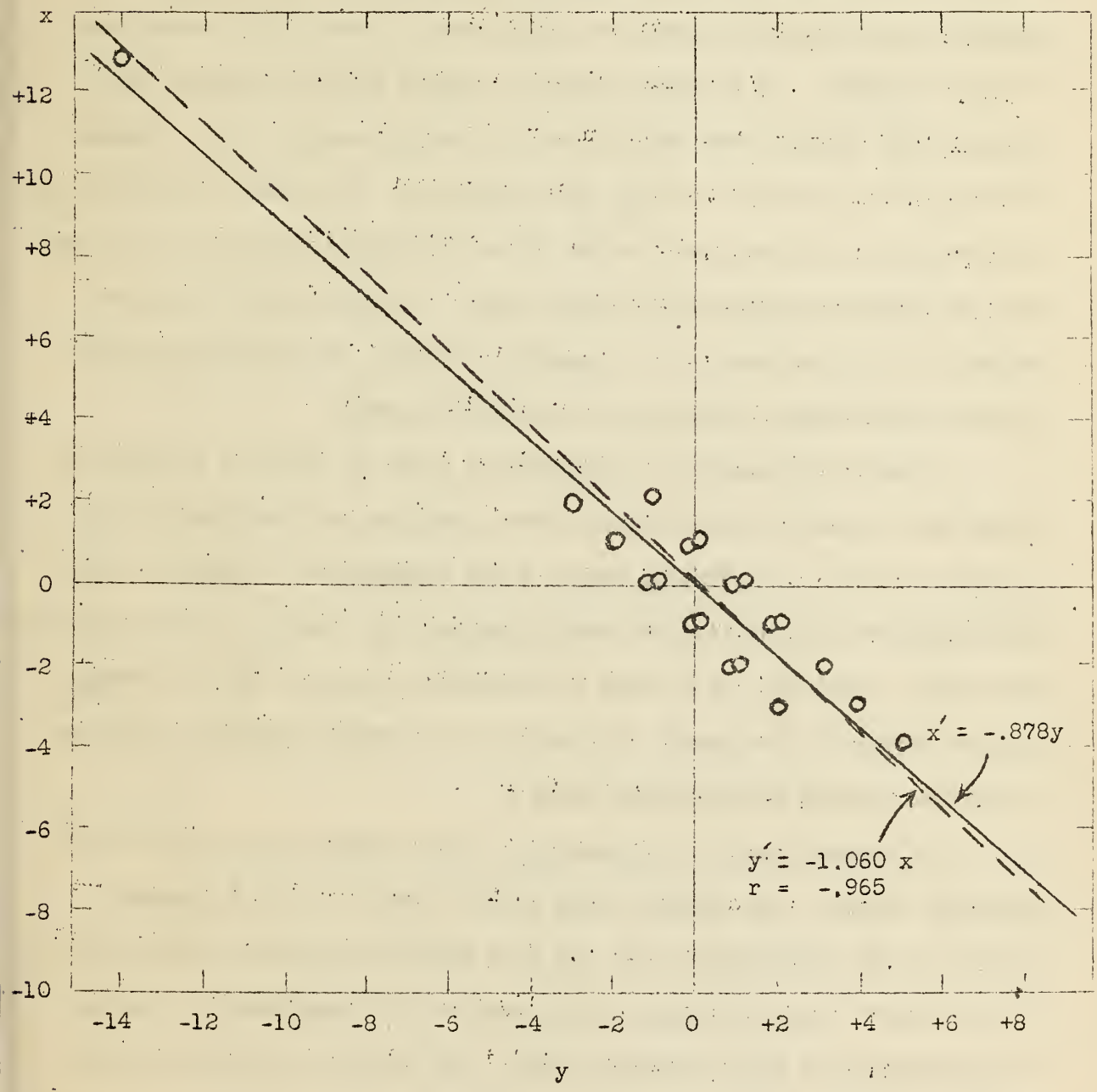
$M\bar{x} = \Sigma f \bar{x} / \Sigma f$:- .1181

TABLE 7. COMPUTATION OF $O\bar{m}$ and N_{xy}

$n = \Sigma f$	From table 5: m	m^2	$n\,m^2$
2	+ 2.00	4.00	8.00
3	+ 1.00	1.00	3.00
17	+ .94	.86	14.62
64	+ .91	.83	53.12
105	+ .80	.64	67.20
159	+ .26	.68	108.12
132	- 1.00	1.00	132.00
76	- 1.25	1.56	118.56
23	- 1.74	3.02	69.46
3	- 2.67	7.13	21.39
584 = n	-	-	595.47
Mean			1.02
Subtract $\frac{(\Sigma n\,m)^2}{n}$	[Table 5]	$.1181^2$ =	.01
$O\bar{m}^2$			1.01
$O\bar{m}$			1.005

$$N_{xy} = \sqrt{1 - \frac{O\bar{z}^2}{O\bar{x}^2}} = \frac{O\bar{m}}{O\bar{x}} = \frac{1.005}{1.432} = .70$$

Figure 1. Graphing of x and y from table 1.



Double Frequency Table.

When there are a considerable number of observations, there are probably a large number of repetitions of given values. The process of squaring and multiplying could be facilitated if these like values were grouped together. It is much simpler to square 347, for example, and multiply the square times the number of times it occurs, than to square 347 each time it appears and add the resultants. For those who have access to punch card machines, the process of securing this grouping is very simple and needs no explanation at this point. For those who do not have access to such machines, or to computing machines, the double frequency table is practically invaluable in its labor saving.

A double frequency or "correlation" table is simply a grouping of items into classes by one variable with each group reclassified by the second variable. In planning such a table care should be taken to insure that the class interval be the same throughout the classes of each variable. The value of any item in a class is customarily taken to be the average of the limits of that class. The nature of a double frequency table can be easily grasped by inspecting table 2.

In constructing a frequency table, the columns and rows are first properly titled. The operator then notes in which vertical grouping, X class, each observation lies, and then moves horizontally across the line to the proper Y class making a tally mark in the compartment. When all observations have been properly tallied, the number of tallies in each compartment is written in. The numbers appearing in the body of a finished

double frequency table therefore represent the number of observations of a given X value which occured in conjunction with the given Y value.

Note (in Table 2) that the class values for X are listed in descending magnitude. This, though not usual, is done for the purpose of bringing the tabular scales into conformity with the scales used on graphs such as shown in figure 1. Indeed, a rough form of graph can be superimposed on the table when the scales are so arranged.

The vertical distributions are called columns, the horizontal, rows. Since it makes no difference statistically which variable runs vertically or horizontally, either row or column may be called an "array". An array is adequately designated by the value of the class, either X or Y as the case may be, in which its entire distribution lies. This class value is termed the "type" of the array.

Also note that space is provided adjacent to the class designations for "assumed values." These assumed values assist very materially in the labor of computation.

Using the assumed values, the standard deviations can be very quickly determined:

(1) Multiply each of the summed frequencies found in the right hand column (table 2) times their corresponding assumed values of X and sum the products. Dividing by the total munber of cases evidently gives the mean of the X values, $S(\overline{X})/n = M_{\overline{X}}$, in terms of assumed values.

(2) Repeat, except instead of using the assumed values of X, use the assumed values squared. Secure the mean of the summed products, $S(\overline{X^2})/n$.

(3) From the mean of the squares so secured it is only necessary to subtract the square of the mean, $M_{\bar{x}}$, found in "(1)" above, to secure the squared standard deviation of the X series in terms of the assumed values, i.e.

$$\sigma_{\bar{x}}^2 = S\bar{x}^2/n - M_{\bar{x}}^2$$

(4) Multiplying $\sigma_{\bar{x}}$ by the class interval yields the true standard deviation of X:

$$\sigma_x = c_x\sigma_{\bar{x}}$$

An analagous procedure secures the standard deviation of the Y values. These computations are illustrated in table 3 and table 4 for the data given in table 2.

To secure the product moment:

(1) A series of products, $f\bar{x}\bar{y}$, are secured, wherein $\bar{x}\bar{y}$ is the product of the corresponding X and Y assumed values for any compartment of the table, and f is the number of observations listed in the compartment. This is simply securing the product of each associated X and Y value (in terms of assumed values) and weighting each different product by the number of times that particular combination occurs.

(2) Subtract from the mean of such products, $S(f\bar{x}\bar{y})/n$, the product of the two means, $M_{\bar{x}}M_{\bar{y}}$, fouud when securing the standard deviations; the difference is the product moment, $p_{\bar{x}\bar{y}}$, in terms of the assumed values, i.e.

$$p_{xy} = S(f\bar{x}\bar{y})/n - M_{\bar{x}}M_{\bar{y}}$$

(3) Multiplying $p_{\bar{x}\bar{y}}$ by the product of the class intervals of X and Y, c_xc_y, gives the product moment in terms of original values, i.e.

$$r_{xy} = c_xc_yp_{\bar{x}\bar{y}}$$

Having the values of the product moment and the two standard devia-

tions, it is possible to secure immediately the coefficient of correlation, $r_{xy} = p_{xy}/\sigma x\, \sigma y$. Table 6 shows the computations.

An algebraic demonstration of the authenticity of using assumed values is given below:

Each assumed value of X, $\bar{X}$, is equal to the original value of X, less the value of X corresponding to the zero on the assumed scale, V_x, divided by the class interval, i.e.,

$$\bar{X} = (X - V_x)/c_x$$

$$\text{Hence} \quad X = c_x \bar{X} + V_x$$

The mean of X, M_x, is therefore equal to the class interval times the mean of the assumed values, plus the value of X in the assumed zero class:

$$M_x = \frac{SX}{n} = \frac{c_x S\bar{X} + nV_x}{n} = c_x M_{\bar{x}} + V_x$$

Substituting the above values of each X and the mean of X below in the proper place, the product moment may be derived as follows:

Each deviation in X, x, $= X - M_x$

$$= c_x\bar{X} + V_x - c_x M_{\bar{x}} - V_x$$

$$= c_x(\bar{X} - M_{\bar{x}})$$

Similarly $y = c_y(\bar{Y} - M_{\bar{y}})$

Therefore, each $xy = c_x c_y(\bar{X}\bar{Y} - \bar{X}M_{\bar{y}} - \bar{Y}M_{\bar{x}} + M_{\bar{x}}M_{\bar{y}})$

Summing and dividing by n

$$\frac{Sxy}{n} = p_{xy} = c_x c_y\left(\frac{S\bar{X}\bar{Y}}{n} - M_{\bar{x}}M_{\bar{y}}\right)$$

$$= c_x c_y p_{\bar{x}\bar{y}}$$

Standard deviations as follows:

$$\text{Each } x^2 = c^2(\bar{X} - M_{\bar{x}})^2$$

Summing and dividing

$$\sigma_{\bar{x}}^2 = \frac{S_{\bar{x}}^2}{n} = c_x^2\left(\frac{S_{\bar{X}}^2}{n} - 2 M_{\bar{x}} \frac{S_{\bar{X}}}{n} + M_{\bar{x}}^2\right)$$

$$= c_x^2 \sigma_{\bar{x}}^2$$

Similarly

$$\sigma_{\bar{y}}^2 = c_y^2 \sigma_{\bar{y}}^2$$

III THE CORRELATION RATIO

Using the assumed values throughout--as though they were the only known values of X and Y, find the value of $b_{\bar{x}\bar{y}}$ By formula (9),

$$b_{\bar{x}\bar{y}} = p_{\bar{x}\bar{y}} / \sigma_{\bar{x}}^2 = -1.214/3.05 = -.592$$

The regression line must, on a graph, pass through the intersection of the means of $\bar{x}$ and $\bar{y}$. Using the correlation table and its scales as a rough form of graph, the line may be plotted. See table 2.

The correlation coefficient was defined as the relation between the standard deviations of the X values if they had all been on the line and their actual standard deviation, $\sigma_{\bar{x}}'/\sigma_{\bar{x}}$. Put in another way, the correlation between $\bar{X}$ and $\bar{Y}$ is a function of the scatter of the $\bar{X}$ values around the regression line compared to their scatter around their mean line. The less the scatter around the regression line, the better the correlation, the more dependable the line as a method of estimating the dependent from its described relation to the independent. But perhaps a curved line could be put on the graph which would come more closely to fitting all the $\bar{X}$ values than does the straight line. As a method of ascertaining if this is so, let the mean value of $\bar{X}$ for each column be secured and indicated by a small circle. (See table 6 for computation.)

The averages are connected with a broken line)

Inspection of the curve of these averages indicates that a closer approximation to the true values of $\bar{X}$ would be secured if this curve were used instead of the $b_{\bar{x}y}$ line.

Bearing in mind the significance of the scatter as a measure of the relation, just as r_{xy} may be defined as $\sqrt{1 - \bar{\sigma}_z^2 / \bar{\sigma}_x^2}$ where the residuals are measured as deviations from the regression line, we may in analagous fashion define N_{xy}(eta) as $\sqrt{1 - \bar{\sigma}_z^2/\bar{\sigma}_x^2}$ when the residuals are measured as deviations from the averages of the arrays, thus obtaining the measure of relation from the scatter around the average line instead of around the regression line.

$$N_{xy} = \sqrt{1 - \bar{\sigma}_z^2/\bar{\sigma}_x^2} \qquad \ldots\ldots (33)$$

N_{xy} is called the "Correlation Ratio" as distinguished from the correlation coefficient. It may be calculated very simply from the relation that may be shown to exist between $\bar{\sigma}_z$, $\bar{\sigma}_x$ and $\bar{\sigma}_m$, the standard deviation of the means of the columns of $\bar{X}$ values weighted upon the number of observations in each group:

The total squared deviation from m_1 in the first group (column) of n_1 observations, $S(d_1^2)$, where m_1 is the value of the mean of the X values for the group may be written

$$\begin{aligned} Sd_1^2 &= S(\bar{x}_1 - m_1)^2 \\ &= S\bar{x}_1^2 - 2m_1 S\bar{x}_1 + n_1 m_1^2 \\ &= S\bar{x}_1^2 - 2m_1^2 n_1 + n_1 m_1^2 \\ &= S\bar{x}_1^2 - m_1^2 n_1 \qquad \ldots\ldots (34) \end{aligned}$$

For all columns $S(d^2)$ becomes $S(z^2)$ and we may write:

$$S(z^2) = \begin{pmatrix} Sx_1^2 - m_1^2 n_1 \div \\ Sx_2^2 - m_2^2 n_2 \div \\ \vdots \\ Sx_n^2 - m_n^2 n_n \end{pmatrix} \quad \ldots\ldots\ldots (35)$$

Letting $S(m^2n)$ represent the sum of terms like $m_1^2 n_1$, $S(z^2) = S(\bar{x}^2) - S(m^2n)$

$$\text{or } \sigma_z^2 = \sigma_{\bar{x}}^2 - \sigma_m^2 \quad \ldots\ldots\ldots (36)$$

Hence,

$$1 - \frac{\sigma_z^2}{\sigma_{\bar{x}}^2} = \frac{\sigma_m^2}{\sigma_{\bar{x}}^2} \quad \ldots\ldots\ldots (37)$$

and $$Nxy = \bar{\sigma}m/\bar{\sigma}x \quad \ldots\ldots\ldots (38)$$

To secure the correlation ratio, Nxy, then, it is only necessary to find the standard deviation of the means of the $\bar{X}$ values for each $\bar{Y}$ type, weighting each mean by the number of items in that array.

Form for computation.
(For computations see table 6)

$S\bar{x}$	n	m	$nm^2 = S\bar{x}.m$
$S\bar{x}_1$	n_1	m_1	$n_1 m_1^2$
$S\bar{x}_2$	n_2	m_2	$n_2 m_2^2$
Sx_n	n_n	m_n	$n_n m_n^2$
Sx	Sn		Snm^2

$$\bar{\sigma}m = \sqrt{\frac{Snm^2}{Sn} - \left(\frac{S\bar{x}}{Sn}\right)^2}$$

No matter which of the two variables is considered the dependent, the correlation coefficient is the same; this is not necessarily true for the correlation ratio. The ratio N xy does not necessarily equal N yx.

The correlation ratio is always as large as the correlation coefficient, usually larger. If the means of the arrays lie along a straight line, the ratio then obviously becomes the equivalent of the correlation coefficient, since the scatter around the regression line and the line of the means would be identical.

A marked difference between the ratio and the coefficient indicates often that a straight line does not satisfactorily describe the relation between X and Y. Since the ratio is derived from the formula, $N_{xy} = \frac{\sigma_a}{\sigma_{\bar{x}}} = \sqrt{1 - \frac{\sigma_z^2}{\sigma_{\bar{x}}^2}}$ it may be either positive or negative in sign. It is customary to consider the ratio as positive.

It is not possible to get a regression equation from the correlation ratio that will fit all parts of the line of averages, because there is no consistent relation between any two points of the curve. The best that can be done is to describe the curve itself--the graph of the line of averages--as the functional relation existing between dependent and independent.

IV. THE CORRELATION INDEX.

If one should now, either by mathematical process, or free hand, smooth the curve of averages derived by methods shown in the previous section, on the basic and usually justifiable assumption that the effect of gradual changes in the independent is a gradual change in the dependent-

a continuous change; and if one should further compute the value of the "root-mean-square" deviation residual from this curve and use this value instead of the residuals from the means line in the formula for the correlation ratio, $\sqrt{1 - \sigma_z^2 / \sigma_x^2}$; a value would be secured which is called the "Correlation Index" 3/ designated by "ρ", rho. This value has

3/ The correlation index is the most recent of the correlation measures. Its formula was devised and the name "index" given it, independently by Mordecai Ezekiel and F. C. Mills. See Ezekiel, Mordecai. A Method of Handling Curvilinear Correlation for Any Number of Variables. Amer. Statis, Assoc. Jour 19; 431 - 453. 1924. Mills, F. C. The Measurement of Correlation and the Problem of Estimation. Amer. Statis. Assoc. Jour. 19; 273 - 300. 1924.

properties characterizing both the ratio and the coefficient, but superior to both. A regression curve is obtained, comparable to the straight regression line for the coefficient. Like the ratio, there are always two index figures for each pair of variables; these however tend to approach each other in value like the coefficient, since changes in the magnitude of the indexes attributable to extreme items is less probable than with the ratio, since the smoothing permits the values in adjoining arrays to prevent the curve from following extreme items in any one array. Often the theoretical considerations underlying a problem predicate that curvilinear relations exist; the index affords a method of describing quantitatively such relations. Like the correlation ratio, the index is considered to be positive in sign. Similar to both ratio and coefficient its value can never be greater than 1.0. Like the coefficient the curve permits the estimating of values of the dependent from the described relation to the independent.

The only satisfactory way to compute the index of correlation is actually to measure the residuals and secure the standard error of estimate in that fashion. However, in the event that the curve used is some mathematical curve fitted by least squares, there are short-cuts which may be used and will be described when Multiple Correlation is discussed.

It is also possible to secure σ_z if the differences between the curve (smoothed) value and the array mean is known for each array. Compute the root-mean-square of these differences, weighting each difference by the number of items in the array--similar to the manner in which σ_m, was computed--and denote it by σ_d. Then $\sigma_z^2 = \sigma_x^2 - \sigma_m^2 \div \sigma_d^2$

TABLE 8 - ASSEMBLY OF SEVERAL MEASURES.

Measure	Formula (not all formulas given)		Value in example (Table 2)	
	In terms of original values	Original in terms of assumed values	For assumed values	For original values
r_{xy}	$\frac{\sigma_{x'}}{\sigma_x}$; $p_{xy}/\sigma_x\sigma_y$	$p_{\bar{x}\bar{y}}/\sigma_{\bar{x}}\sigma_{\bar{y}}$; $\frac{\sigma_{\bar{x}'}}{\sigma_{\bar{x}}}$	-.569	-.569
d_{xy}	r^2_{xy}	$r^2_{\bar{x}\bar{y}}$	.324	.324
σ_z	$\sigma_x\sqrt{1 - r^2_{xy}}$	$C_x\sigma_{\bar{x}}\sqrt{1 - r^2_{\bar{x}\bar{y}}}$	1.18	2.36
b_{xy}	$r\frac{\sigma_x}{\sigma_y}$; $\frac{p_{xy}}{\sigma_y^2}$	$r\frac{\sigma_{\bar{x}}C_x}{\sigma_{\bar{y}}C_y}$; $\frac{p_{\bar{x}\bar{y}}C_x}{\sigma_{\bar{y}}^2C_y}$	-.547	-.219
σ_x	$\sqrt{\frac{Sx^2}{n}}$; $\sqrt{\frac{SX^2}{n} - M^2_x}$	$C_x\sqrt{\frac{S\bar{x}^2}{n}}$; $C_x\sqrt{\frac{S\bar{X}^2}{n} - M^2_{\bar{x}}}$	1.432	2.864
$\sigma_{x'}$	$(b_{xy}\sigma_y)$	$(b_{\bar{x}\bar{y}}\sigma_{\bar{y}})C_x$	.815	1.630
p_{xy}	$\frac{Sxy}{n}$; $\frac{SXY}{n} - M_xM_y$	$\frac{S\bar{x}\bar{y}}{n}C_xC_y$;		
		$C_xC_y(\frac{S\bar{X}\bar{Y}}{n} - M_{\bar{x}}M_{\bar{y}})$	-1.214	-12.14
N_{xy}	$\frac{\sqrt{1 - \sigma_z^2}}{\sigma_x^2}$; $\frac{\sigma_{mx}}{\sigma_x}$	$\frac{\sqrt{1 - \sigma_{\bar{z}}^2}}{\sigma_{\bar{x}}^2}$; $\frac{\sigma_{m\bar{x}}}{\sigma_{\bar{x}}}$	.70	.70
ρ_{xy}	Same as N_{xy} except residuals measured from smoothed curve		-	-
σ_y	Similar to σ_x	-	7.490	7.450

V. Multiple Linear Correlation [4]

(1) Regression

The preceding sections have been devoted to methods of measuring the relation of one (independent) variable to the dependent variable, and the reliability or constancy of that relation. The following sections will be devoted to methods of measuring the relation of several independent variables to the dependent variable and the reliability or constancy of the several relations. Since there are several independents instead of but one, a further problem is automatically introduced, that of determining the relative significance of the several relationships discovered.

In most analyses of problems it will develop that several factors influence the given, dependent factor. For example, there are numerous factors which influence the price of a commodity. For convenience it may be said, then, that the dependent variable (price) is some function of the other variables (supply, demand, price level, etc.). This may be symbolized

$$X = F\ (A, B, C \ldots)$$

wherein X denotes the dependent and the initial letters in the alphabet the independents, and F means "function of". The problem is to determine the nature of the function. The gross correlation methods enabled us to determine the best values of the necessary constants when it was assumed that the functional relationship was essentially linear. In like manner, if we assume that the nature of the relationships of the

---

[4] For the original presentation of the least square approach to multiple correlation see Tolley, H. R., and Ezekiel, Mordecai. A Method of Handling Multiple Correlation Problems. Amer. Statis. Assoc. Journ. 18: 993 - 1003. 1923.

dependent to the several independents is linear, multiple correlation methods, by a simple extension of gross correlation methods, permit us to determine the best values of the necessary constants. This assumption of linearity may be given algebraic expression by writing

$$x = b_1 a + b_2 b + b_3 c + \ldots \quad (39)$$

wherein x, a, b, c represent deviations from average in X, A, B, C. Deviations are employed, for, as in simple correlation, there are advantages in comparing deviations from average rather than original values. Formula (39), comparable with formula (3), implies not only a linear relation between the dependent and each independent, but also that the components of the independents are added together before being equated to x. This may be a quite inappropriate assumption in some cases. Products rather than sums might be a more valid type of functional relationship to assume in certain cases. Nevertheless, linear multiple correlation method is incapable of comprehending any other type of relationship than that shown in formula (39). It is true that by using logarithms or reciprocals or some other functions of the different variables included, a certain amount of elasticity in this last assumption may be obtained. Nevertheless, once these logarithms or reciprocals or other functions are determined upon, linear correlation method can do no more than provide the best values in formula (39), no matter how appropriate or inappropriate the type of relationship therein assumed may be. Evidently multiple correlation method provides means for testing within but a narrow range that which the analysis of the problem may evolve, i. e. that the dependent is some function of several other factors. These limitations

must be borne in mind in any interpretation of multiple correlation results.

Suppose a dependent X, with deviations from mean represented by x; and similarly deviations in A, B, C ..., etc. of a, b, c, ..., etc. Then the constants $\underline{b}$ in formula (39) may be obtained by an extension of the process wherein $\underline{b}$ was found for formula (3).

In the case of gross regression, formula (3), the value of $\underline{b}$ was found by forming a normal equation and solving. In the present case the several values of $\underline{b}$ may be found by forming several normal equations in analagous fashion, there being as many normal equations as unknown values, $\underline{b}$, and then solving these normal equations by any simulataneous method the investigator cares to use.

The first normal equation is obtained by multiplying each observation equation, formula (39), thru by the coefficient of the first unknown constant and summing the product equations so secured, i.e.

$$a_1 (x_1 = \underline{b}_1 a_1 + \underline{b}_2 b_1 + \underline{b}_3 c_1) = (a_1 x_1 = \underline{b}_1 a_1^2 + \underline{b}_2 a_1 b_1 + \underline{b}_3 a_1 c_1)$$

$$a_2 (x_2 = \underline{b}_1 a_2 + \underline{b}_2 b_2 + \underline{b}_3 c_2 + \ldots) = (a_2 x_2 = \underline{b}_1 a_2^2 + \underline{b}_2 a_2 b_2 + \underline{b}_3 a_2 c_2)$$

- -

$$a_n (x_n = \underline{b}_1 a_n + \underline{b}_2 b_n + \underline{b}_3 c_n) = (a_n x_n = \underline{b}_1 a_n^2 + \underline{b}_2 a_n b_n + \underline{b}_3 a_n c_n)$$

Summing gives

$$\text{Normal Equation 1} \ldots\ S(ax) = \underline{b}_1 . S(a^2) + \underline{b}_2 . S(ab) + \underline{b}_3 . S(ac)$$

Multiplying each observation equation through by the coefficient of the second unknown, $\underline{b}_2$, and summing the product equations gives the second normal equation, i.e.

$$b_1 \quad (x_1 = \underline{b}_1 a_1 + \underline{b}_2 b_1 + \underline{b}_3 c_1) = (b_1 x_1 = \underline{b}_1 a_1 b_1 + \underline{b}_2 b_1^2 + \underline{b}_3 b_1 c_1)$$

- -

$$b_n \quad (x_n = \underline{b}_1 a_n + \underline{b}_2 b_n + \underline{b}_3 c_n) = (b_n x_n = \underline{b}_1 a_n b_n + \underline{b}_2 b_n^2 + \underline{b}_3 b_n c_n)$$

Normal Equation II. ... $S(bx) = \underline{b}_1 . S(ab) + \underline{b}_2 . S(b^2) + \underline{b}_3 . S(bc)$

In a similar manner normal equations are constructed for each of the unknowns. In the case of three independent and three unknowns, $\underline{b}_1$, $\underline{b}_2$, and $\underline{b}_3$, the normal equations are:

$$\left.\begin{aligned} \text{I.} \quad & \underline{b}_1 . S(a^2) + \underline{b}_2 . S(ab) + \underline{b}_3 . S(ac) = S(ax) \\ \text{II.} \quad & \underline{b}_1 . S(ab) + \underline{b}_2 . S(b^2) + \underline{b}_3 . S(bc) = S(bx) \\ \text{III.} \quad & \underline{b}_1 . S(ac) + \underline{b}_2 . S(bc) + \underline{b}_3 . S(c^2) = S(cx) \end{aligned}\right\} \qquad \text{.....(40-a)}$$

The absolute terms are given on the right of the equality signs since arithmetically this arrangement represents a somewhat simpler solution.

If all terms are divided by n the coefficients of the unknowns in the normal equations become familiar product moments and standard deviations, Thus:

$$\left.\begin{aligned} \text{I.} \quad & \underline{b}_1 \sigma_a^2 + \underline{b}_2 p_{ab} + \underline{b}_3 p_{ac} = p_{ax} \\ \text{II.} \quad & \underline{b}_1 p_{ab} + \underline{b}_2 \sigma_b^2 + \underline{b}_3 p_{bc} = p_{bx} \\ \text{III.} \quad & \underline{b}_1 p_{ac} + \underline{b}_2 p_{bc} + \underline{b}_3 \sigma_c^2 = p_{cx} \end{aligned}\right\} \qquad \text{. .....(40-b)}$$

An easy method of remembering these coefficients is to imagine a box-like figure with columns and rows designated by the variables, as

given below:

	a	b	c	x
a	aa	ab	ac	ax
b	ba	bb	bc	bx
c	ca	cb	cc	cx
x	ax	bx	cx	xx

In each compartment the appropriate column and row letter designations are listed which then give the subscripts to the product moments. Subscripts "aa", "bb", and "cc" evidently designate standard deviations. These latter are styled "the diagonal terms". An inspection shows a symmetrical distribution of coefficients around the diagonals.

Thus where the coefficient of b_2 is p_{ab} in the first row, second column, the same value occurs in the second row, first column, as a coefficient of b_1. The symmetrical nature of these normal equations adapts them to certain time saving methods of solution; and it is eminently worth while for the investigator to learn these special methods if he anticipates the necessity of solving any number of such simultaneous equations. They will be discussed briefly in the section dealing with arithmetic methods. Any method of solving simultaneous equations, however, is perfectly valid for the determining of the values of b.

The significance of the values of b when determined, is, as in the case of gross correlation, that by the use of these values the sum of the squared residuals will be a minimum; and by this criterion the values of b will be the "best possible values". The residuals are found, as in the case of simple correlation, by finding estimates of x, x', from the regression equation and subtracting these estimates from the associated values of x, i.e.,

$$z = x - x'$$

In the case of gross correlation, x' was determined by multiplying the associated values of y by the determined value of $\underline{b}$, that is $x' = \underline{b}y$, in which y was the independent. In the case of multiple correlation, an analogous procedure is followed. The value, x', is determined by multiplying each associated value of the independents by the appropriate determined values of $\underline{b}$ and adding, i.e.,

$$x' = \underline{b}_1 a + \underline{b}_2 b + \underline{b}_3 c \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (41)$$

The residual, z, may then be expressed algebraically,

$$z = x - x'$$

$$= x - \underline{b}_1 a - \underline{b}_2 b - \underline{b}_3 c \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (42)$$

And by the theory of least squares the sum of the squared residuals, $S(z^2)$, will be a minimum,

$$S (x - \underline{b}_1 a - \underline{b}_2 b - \underline{b}_3 c)^2 = \text{minimum.}$$

In gross correlation the value of $\underline{b}$ was termed "the regression coefficient". In multiple correlation the values of $\underline{b}$ are termed "net regression coefficients", the word, "net", implying that more than one independent variable is used, and that the effects of one or more other variables are removed. And, just as in the case of gross correlation, the values of $\underline{b}$ may be interpreted as indicating the amount of change in x which is associated, on the average, with a unit change in the given independent variable.

Since there are several independent variables, the representation of $\underline{b}$ as the slope of a regression line on a single dot chart is slightly more complicated than in the case of simple correlation. Nevertheless, this representation can be made and is quite helpful. It is essential to a comprehension of multiple curvilinear correlation, which will be described later.

Before a dot chart and a regression line representing the effect of a variable, a, on the dependent, x, can be constructed, the amount of influence of the other variables, b and c, must be eliminated from x, or else the true, or net, relation of a to x will be obscured. This means that we cannot plot the values of x against the values of a in making a dot chart, but that we must plot the values of x corrected for the influence of the other independents - with the effect on x of these other independents removed - against the values of a.

Since the amount of influence of b and c on x is given by the net regressions of x on these two variables, the removal of this influence may be accomplished for any instance of x by merely substracting from x the quantities b_2 b and b_3 c for associated values of b and c, i.e. x corrected for the influence of b and c is given by

$$x - b_2 b - b_3 c = j \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(43)$$

If j, then, is plotted against the values of a, a dot chart will be secured which will show graphically the net relation of a to x. And, if on this dot chart a straight line be drawn to pass through as many of the dots as possible - or rather, be drawn so that the summed squared deviations from the line (in the "j" direction) will be a minimum - this line will have a slope b_1. It would be identical with the slope of a line obtained from determining the regression of j on a by the method of gross correlation.

The identity of the two may be easily demonstrated by showing that any gross regression coefficient is such that the correlation between the independent and the residulas is zero, and by then showing that the correlation between $(j - b_1 a = z)$ and a is also zero.

Thus when $\underline{b}$ is determined in $x = \underline{b}y$

$$b = \frac{S(xy)}{S(y^2)} \quad \text{and} \quad x' = y\frac{S(xy)}{S(y^2)}$$

and

$$z = x - x'$$

$$= x - y\frac{S(xy)}{S(y^2)}$$

Then

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

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

$$= 0$$

No other value of $\underline{b}$ than $\frac{S(xy)}{S(y^2)}$ will give $p_{yz} = 0$. The correlation between $(j - \underline{b}_1 a = z)$ and a may also be shown to equal zero:

Thus

$$z = j - \underline{b}_1 a$$

$$= x - \underline{b}_1 a - \underline{b}_2 b - \underline{b}_3 c$$

Then

$$p_{az} = \frac{1}{n} . S[a(x - \underline{b}_1 a - \underline{b}_2 b - \underline{b}_3 c)]$$

$$= \frac{1}{n}[S(ax) - \underline{b}_1 .S(a^2) - \underline{b}_2 .S(ab) - \underline{b}_3 .S(ac)]$$

$$= p_{ax} - \underline{b}_1 \sigma_a^2 - \underline{b}_2 p_{ab} - \underline{b}_3 p_{ac}$$

But by the solution of the normal equations [I.] ,

$$\underline{b}_1 \sigma_a^2 + \underline{b}_2 p_{ab} + \underline{b}_3 p_{ac} = p_{ax}$$

hence

$$p_{az} = p_{ax} - p_{ax}$$

$$= 0 \qquad \text{.......................................} \qquad (44)$$

This is an important theorem to remember: Independent variables are uncorrelated with the residuals; and only the least-square values of $\underline{b}$ will produce this result. This proves the identity previously mentioned.

The dot chart with ordinates of j and abscissae, a, having been constructed, this chart then shows the net relation between the variable, a, and x. In an exactly analagous manner, charts may be made to show the net relation between x and the other independents, b and c, the ordinates in one case being $(x - \underline{b}_1 a - \underline{b}_3 c)$ and in the other $(x - \underline{b}_1 a - \underline{b}_2 b)$.

The sum of the squared residuals is, of course, the same for all the charts, since a residual is in each case given by formula (42).

We have now developed methods for finding and representing graphically the relationship of several independent variables to the dependent upon assumptions implicit in formula (39)

(2) Correlation

The next proposition, as in the case of gross correlation, is to develop some measure of the consistency or reliability of the relationships discovered. As in that case, a logical procedure is to compare the standard deviation of the residuals with the standard deviation of the original x values. And, as in that case, some relation between the σ_z^-, σ_x^-, and $\sigma_{x'}^-$ may be found to enable us to compute the measures easily. As in the case of gross correlation,

$$\sigma_x^{-2} = \sigma_{x'}^{-2} + \sigma_z^{-2}$$

is found to express the relationship between the three values, as follows:

$$\text{Each } z = x - x'$$

$$= x - \underline{b}_1 a - \underline{b}_2 b - \underline{b}_3 c$$

$$\text{Each } z^2 = (x - \underline{b}_1 a - \underline{b}_2 b - \underline{b}_3 c)^2$$

Expanding, summing and dividing by n

$$\left.\begin{aligned} \sigma_z^2 = {} & \sigma_x^2 - \underline{b}_1 p_{ax} - \underline{b}_2 p_{bx} - \underline{b}_3 p_{cx} \\ & - \underline{b}_1 p_{ax} + \underline{b}_1^2 \sigma_a^2 + \underline{b}_1 \underline{b}_2 p_{ab} + \underline{b}_1 \underline{b}_3 p_{ac} \\ & - \underline{b}_2 p_{bx} + \underline{b}_1 \underline{b}_2 p_{ab} + \underline{b}_2^2 \sigma_b^2 + \underline{b}_2 \underline{b}_3 p_{bc} \\ & - \underline{b}_3 p_{cx} + \underline{b}_1 \underline{b}_3 p_{ac} + \underline{b}_2 \underline{b}_3 p_{bc} + \underline{b}_3^2 \sigma_c^2 \end{aligned}\right\} \quad \ldots\ldots(45)$$

But taking the first normal equation and multiplying thruout by $\underline{b}$, we have

$$\underline{b}_1^2 \sigma_a^2 + \underline{b}_1 \underline{b}_2 p_{ab} + \underline{b}_1 \underline{b}_3 p_{ac} = \underline{b}_1 p_{ax}$$

$\underline{b}_1 p_{ax}$ may therefore be substituted for the equivalent three terms in the second row of the expanded square above. Substituting in like manner for the equivalents of $\underline{b}_2 p_{bx}$ and $\underline{b}_3 p_{cx}$ from the second and third normal equations, and collecting terms, (45) becomes

$$\sigma_z^2 = \sigma_x^2 - (\underline{b}_1 p_{ax} + \underline{b}_2 p_{bx} + \underline{b}_3 p_{cx}) \quad \ldots\ldots(46)$$

This is a perfectly general development applicable to any number of variables. It now remains to be shown that the terms enclosed in the parentheses, formula (46), are equal to $\sigma_{x'}^2$.

$$\text{Each } x' = \underline{b}_1 a + \underline{b}_2 b + \underline{b}_3 c$$

$$\text{and} \quad \frac{S(x'^2)}{n} = \left.\begin{aligned} &\underline{b}_1^2 \sigma_a^2 + \underline{b}_1 \underline{b}_2 p_{ab} + \underline{b}_1 \underline{b}_3 p_{ac} \\ &+ \underline{b}_1 \underline{b}_2 p_{ab} + \underline{b}_2^2 \sigma_b^2 + \underline{b}_2 \underline{b}_3 p_{bc} \\ &+ \underline{b}_1 \underline{b}_3 p_{ac} + \underline{b}_2 \underline{b}_3 p_{bc} + \underline{b}_3^2 \sigma_c^2 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots(47)$$

$$= \left.\begin{aligned} &\underline{b}_1(\underline{b}_1 \sigma_a^2 + \underline{b}_2 p_{ab} + \underline{b}_3 p_{ac}) \\ &+ \underline{b}_2(\underline{b}_1 p_{ab} + \underline{b}_2 \sigma_b^2 + \underline{b}_3 p_{bc}) \\ &+ \underline{b}_3(\underline{b}_1 p_{ac} + \underline{b}_2 p_{bc} + \underline{b}_3 \sigma_c^2) \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots(48)$$

But the terms enclosed in the three parentheses in (48) are from the normal equations respectively equal to p_{ax}, p_{bx} and p_{cx}.

Hence $$\sigma_{x'}^2 = \underline{b}_1 p_{ax} + \underline{b}_2 p_{bx} + \underline{b}_3 p_{cx} \quad \ldots\ldots\ldots\ldots\ldots\ldots(49)$$

Formula (49) is a very useful formula and should be remembered.

Substituting (49) in (46) we have

$$\sigma_z^2 = \sigma_x^2 - \sigma_{x'}^2 \quad \text{or}$$

$$\sigma_x^2 = \sigma_{x'}^2 + \sigma_z^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(50)$$

which is identical with formula (18).

Just as in the case of simple, or gross, correlation, a measure of the closeness and reliability of the relationships discovered may bd had by expressing $\sigma_{x'}^2$ as a percentage or decimal fraction of σ_x^2; and a coefficient of "alienation" by expressing σ_z^2 as a percentage, or decimal fraction of σ_x^2. The relationship between the two measures is, of course,

$$\frac{\sigma_{x'}^2}{\sigma_x^2} + \frac{\sigma_z^2}{\sigma_x^2} = 1$$

If $\frac{\sigma_{x'}^2}{\sigma_x^2}$ is reduced to the first order it becomes $\frac{\sigma_{x'}}{\sigma_x}$ and may be designated, $R_{x.abc}$ wherein the subscripts to the right of the point designate the independent variables employed. $R_{x.abc}$ is classically known as "the coefficient of multiple correlation", and literally represents the decimal fraction that the standard deviation of estimates is of the standard deviation of the original X values. $R_{x.abc}$ is also numerically equivalent to the ordinary gross coefficient of correlation between the original X values and the estimates, X'; i.e.,

$$R_{x.abc} = r_{xx'} \qquad \text{.....(51)}$$

This is a useful concept in interpreting $R_{x.abc}$ and the equality between the two may be proved easily, as follows:

$$R_{x.abc} = \frac{\sigma_{x'}}{\sigma_x}$$

and $$r_{xx'} = \frac{p_{xx'}}{\sigma_x \sigma_{x'}} \quad \text{by definition,}$$

but $$p_{xx'} = \frac{1}{n} S\left[x(\underline{b}_1 a + \underline{b}_2 b + \underline{b}_3 c)\right]$$

$$= \frac{1}{n}\left[\underline{b}_1 . S(xa) + \underline{b}_2 . S(xb) + \underline{b}_3 . S(xc)\right]$$

$$= \underline{b}_1 p_{xa} + \underline{b}_2 p_{xb} + \underline{b}_3 p_{xc} \qquad \text{.....(52)}$$

But by comparing (52) with (49) it becomes apparent that

$$p_{xx'} = \sigma_{x'}^2 \qquad \text{.....(53)}$$

Hence substituting for $p_{xx'}$,

$$r_{xx'} = \frac{\sigma_{x'}^2}{\sigma_{x'} \sigma_x}$$

$$= \frac{\sigma_{x'}}{\sigma_x}$$

$$= R_{x.abc} \qquad \text{.....(54)}$$

Formula (49) provides a simple and easy way of computing $R_{x.abc}$ once the constants in the multiple regression equation are determined. Thus

$$R^2_{x.abc} = \frac{\sigma_{x'}^2}{\sigma_x^2}$$

$$= \frac{b_1 p_{ax} + b_2 p_{bx} + b_3 p_{cx}}{\sigma_x^2} \quad \ldots\ldots\ldots (55)$$

The evaluation of formula (55) requires no other values than those already provided by the solution of the normal equations. If σ_z^2 is known, a simple way of computing $R_{x.abc}$ is from the relationship,

$$R^2_{x.abc} = 1 - \frac{\sigma_z^2}{\sigma_x^2} \quad \ldots\ldots\ldots\ldots\ldots (56)$$

This formula becomes of great importance in handling curvilinear multiple correlation problems.

$R_{x.abc}$ is a coefficient of correlation which gives a measure of the reliability of the various regression relationships discovered, when taken as a group, or in another sense it is a measure of the reliability of estimating the dependent from its discovered relationship to the several independents. It does not, however, provide any means of apportioning importance to the various independent variables. This is the next problem before us.

(3) Determination, part and partial correlation.

Recalling the coefficients of determination developed in connection with gross correlation, it was found that $(\sigma_{x'}^2/\sigma_x^2 = d_{xy})$ represented the proportion of total squared variability in the dependent that was attributable to the independent, y. An analagous procedure may be employed in the case of multiple correlation.

Thus from (50) the proportion of the total squared variability in the dependent attributable to the independents is given by $\sigma_{x'}^2/\sigma_x^2$. It becomes necessary to apportion the variability of $\sigma_{x'}^2$ to the different variables which make up x', and thus attain our object. This may be accomplished from a consideration of formula (49). Here $\sigma_{x'}^2$ is defined as the sum of three terms, one each for each independent. The proportion that each independent contributes to the squared variability of x' may be determined then by taking the decimal fraction that each of the three terms is of $\sigma_{x'}^2$. To relate these fractions to the independent it is necessary to multiply by the proportions

$$\sigma_{x'}^2/\sigma_x^2$$

Thus

$$d_{xa.bc} = \frac{b_1 p_{ax}}{\sigma_{x'}^2} \cdot \frac{\sigma_{x'}^2}{\sigma_x^2}$$

$$= \frac{b_1 p_{ax}}{\sigma_x^2}$$

$$d_{xb.ac} = \frac{b_2 p_{bx}}{\sigma_x^2} \qquad \text{.....(57)}$$

$$d_{xc.ba} = \frac{b_3 p_{cx}}{\sigma_x^2}$$

$d_{xa.bc}$ symbolizes the "net determination of x by a". Subscripts to the right of the point designate the other independent variables included in the study. The order in which these latter subscripts are listed is immaterial. The first of the subscripts to the left of the point should always designate the dependent, the second, the independent under consideration. 5/

5/ For original treatment of coefficients of determination see the following: Wright, S. Correlation and Causation. Jour. Agr. Research 20: 557 - 585. 1921. Smith, B. B. Forecasting the Acreage of Cotton. Amer. Statis. Assoc. Jour. 20: 31 - 47. 1925.

Inspection of the formulae for coefficients of determination shows that whenever the net regression of x on the given independent is of opposite sign from the gross regression (which takes the sign of the product moment) of x on the independent, the coefficient of net determination is negative. This makes the interpretation of these coefficients difficult. When such a condition arises it is usually due to relatively high inter-correlation among independents, as compared with the correlation between independents and dependent. It may be interpreted that the influence of one variable upon the dependent is greater through a second variable than directly. It is thus traceable back to the inappropriateness of the implicit assumptions of the multiple net regression equation. There are, however, certain characteristics of determination coefficients which give them definite meaning. The sum of them equals R^2, or the total determination of the dependent by the several independents. Furthermore, if two or more terms such as b_1 a and b_2 b be added together prior to the computing of x', then the determination by the joined series may be found by adding together the coefficients from the separate series: i.e.,

$$d_{x(a+b).c} = d_{xa.bc} + d_{xb.ac} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (57)$$

This is a useful theorem, for in the event that negative determinations appear, they may be added to the determinations by other closely related independents; and the determination of that particular group of independents, as a whole, may thus be ascertained.

Another method of measuring the relative importance of a given independent to the dependent may be developed by recalling to mind the dot graphs described as a means of showing graphically the net relation of the various independents to the dependent. Here the ordinates were in each case x minus

the functional contributions (such as $\underline{b}_2$ b and $\underline{b}_3$ c) of all other independents except the given one. Thus in one case the ordinates were ($j = x - \underline{b}_2 b - \underline{b}_3 c$) and the abscissae were a. The regression line had a slope of $\underline{b}_1$. Taking this individual dot chart, we have essentially a gross correlation problem, j being the dependent and a the independent. Ordinary gross correlation methods may therefore be used to measure the importance of a to j. Thus

$$r_{aj} = \frac{\sigma_{j'}}{\sigma_j} \qquad \text{.....(58)}$$

This is but the relation of the standard deviation of estimates of j represented by points on the regression line, compared to the standard deviation of j itself.

But since each $j' = \underline{b}_1 a$ .....(59)

Then $\sigma_{j'} = \underline{b}_1 \sigma_a$ .....(60)

Hence $r_{aj} = \frac{\underline{b}_1 \sigma_a}{\sigma_j}$ .....(61)

The value σ_j may be determined as follows:

The basic formula, $\sigma_x^2 = \sigma_{x'}^2 + \sigma_z^2$ may be paralleled and

$$\sigma_j^2 = \underline{b}_1^2 \sigma_a^2 + \sigma_z^2 \qquad \text{.....(62)}$$

σ_z^2 is, of course, given by the formula

$$\sigma_z^2 = \sigma_x^2 (1 - R^2_{x.abc}) \qquad \text{.....(63)}$$

Substituting (63) in (62)

$$\sigma_j^2 = \underline{b}_1^2 \sigma_a^2 + \sigma_x^2 (1 - R^2_{x.abc}) \qquad \text{.....(64)}$$

Substituting (64) in (61)

$$r^2_{aj} = \frac{\underline{b}_1^2 \sigma_a^2}{\underline{b}_1^2 \sigma_a^2 + \sigma_x^2 (1 - R^2_{x.abc})}$$

$$= \frac{1}{1 + \dfrac{\sigma_x^2 (1 - R^2_{x.abc})}{\underline{b}_1^2 \sigma_a^2}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (65)$$

This is the definition of an, as yet, little used measure, b/ and may be called the "coefficient of part correlation" as contrasted with "the coefficient of partial correlation", which has a different meaning. It is always positive in sign, varies between the limits of 0.0 and + 1.0, is quite easily computed, and perfectly general for systems of any number of inscribed independents. It is only necessary to insert the proper $\underline{b}^2$ and σ^2 values in (65) to determine it, the other values in (65) remaining constant for any given system.

A generalized notation may be developed, thus

$$a r^2_{x - bcde \cdots n} = \frac{1}{1 + \dfrac{\sigma_x^2 (1 - R^2_{x.abc\ldots n})}{\underline{b}^2_{xa.bcd\ldots n} \sigma_a^2}} \qquad \ldots\ldots\ldots (66)$$

The first subscript to the right of r designates the dependent, the subscripts following the minus sign represent the independent whose functions have been subtracted from x and the subscript to the left of r the remaining independent which is correlated with the dependent when so treated.

Subscripts to $\underline{b}$ are similar in meaning to subscripts for $\underline{d}$ already explained - the first one designating the dependent, the second the given independent and subscripts to the right of the point the remaining

b/ This measure was worked out together by Mordecai Ezekiel and B. B. Smith and is here published for the first time.

independents included.

$a^{r}x - bcde \ldots n$ is literally the plus or minus coefficient of correlation between

$$(x - \underline{b}_{xb.acd\ldots n}b - \underline{b}_{xc.abc\ldots n}c \ldots - \underline{b}_{xn.abc\ldots(n-1)}n)$$

and a.

Measures of "determination" and of "part correlation" have been developed in the above as means of measuring the importance of the various independent variables to the dependent. A third measure which is classical in its use, but much more laborious to compute will now be described. This is the coefficient of "partial" or "net" correlation 7/

If, as before, we take the case of four variables (deviation from means) x, a, b, c and determine the partial correlation of x and a, expressed by $r_{xa.bc}$, this could be done as follows:

(1) Find the values of $\underline{b}$ in the following:

$$x = \underline{b}_{xb.c}b + \underline{b}_{xc.b}c$$

and in $$a = \underline{b}_{ab.c}b + \underline{b}_{ac.b}c$$

(2) determine the two sets of residuals

$$x - (\underline{b}_{xb.c}b + \underline{b}_{xc.b}c) = z_1$$

$$a - (\underline{b}_{ab.c}b + \underline{b}_{ac.b}c) = z_2$$

(3) and find $r_{z_1z_2} = r_{xa.bc}$

The value of $r_{z_1z_2} = r_{xa.bc}$ may be determined from the following considerations:

z_2 is uncorrelated with b or c, i.e.,

$$r_{z_2b} = 0 \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (67)$$

$$r_{z_2c} = 0$$

7/ Yule, G. U. An Introduction to the Theory of Statistics. 6th ed., rnl. London, 1922, chap. XII.

This is because, by a previously demonstrated theorem, residuals are uncorrelated with independents, the product moments being zero. (See formula (44) et seq.)

In a similar manner

$$r_{z_1 b} = 0 \qquad \text{.....} \quad (68)$$

$$r_{z_1 c} = 0$$

If now z_1 and z_2 be correlated and z_1 be estimated from z_2 by the resulting regression equation, these estimates may be subtracted from z_1 giving z_3 or the error of estimating z_1 from z_2, i.e.,

$$z_3 = z_1 - \underline{b} z_2 \qquad \text{.....} \quad (69)$$

wherein $\underline{b}$ is the gross regression of z_1 on z_2.

Then by application of previously developed theory,

$$r^2_{xa.bc} = r^2_{z_1 z_2} = 1 - \frac{\sigma^2_{z_3}}{\sigma^2_{z_1}} \qquad \text{.....} \quad (70)$$

It remains to determine the values of σ_{z_3} and σ_{z_1}

Now it may be shown that both b and c are uncorrelated with z_3 for, taking the variable, c, as a case,

p_{cz_3} may be shown to equal zero:

Thus

$$p_{cz_3} = \frac{1}{n} .S(c.z_3)$$

$$= \frac{1}{n} .S[c(z_1 - \underline{b} z_2)]$$

$$= p_{cz_1} - \underline{b} p_{cz_2} \qquad \text{.....} \quad (71)$$

But by (67) and (68) both p_{cz_1} and p_{cz_2} are zero and hence

$$p_{cz_3} = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (72)$$

Similarly, b is also uncorrelated with z_3.

Not only is z_3 uncorrelated with b and c but it is also obviously uncorrelated with z_2. It is therefore also uncorrelated with a, for

$$p_{z_3z_2} = 0$$

$$= \frac{1}{n}.S\left[z_3(a - \underline{b}_{ab.c}\, b - \underline{b}_{ac.b}c)\right]$$

$$= p_{az_3} - \underline{b}_{ab.c}p_{bz_3} - \underline{b}_{ac.b}\, p_{cz_3}$$

But since p_{bz_3} and p_{cz_3} are zero, it follows that

$$p_{az_3} = p_{z_3z_2}$$

$$= 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (73)$$

In short, z_3 or the errors of estimating z_1 from z_2, are uncorrelated with all the variables in the system, with the exception of x. It follows, therefore, by the theory developed in connection with formula (44), that the values of z_3 are those that would be secured if estimates of x were secured from the multiple regression equation,

$$x' = \underline{b}_{xa.bc}a + \underline{b}_{xb.ac}b + \underline{b}_{xc.ab}c$$

and subtracted from x, i.e.,

$$z_3 = x - \underline{b}_{xa.bc}a - \underline{b}_{xb.ac}b - \underline{b}_{xc.ab}c$$

and thus $$\sigma^2_{z_3} = \sigma^2_x (1 - R^2_{x.abc}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (74)$$

The value $\sigma^2_{z_1}$ may be written, of course,

$$\sigma^2_{z_1} = \sigma^2_x (1 - R^2_{x.bc}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (75)$$

The equivalents given in (74) and (75) may be substituted in (70) to give, then, a formula which may be utilized in the computation of partial coefficients of correlation:

$$r^2_{xa.bc} = 1 - \frac{\sigma_x^2 (1 - R^2_{x.abc})}{\sigma_x^2 (1 - R^2_{x.bc})} \quad \ldots\ldots\ldots\ldots\ldots\ldots (76)$$

In practice, of course, the σ_x^2 values cancel. This formula represents a new concept of the coefficient of partial correlation - it is a function of the ratio of the errors of estimating x from all the independents to the errors of estimating x from all the independents except the given one. Or, again, it is a function of the amount that the error of estimating x is reduced by including the given independent in the estimating. This last is a convenient concept of the partial correlation coefficient. But its literal meaning should not be forgotten - it is the correlation between any two variables in a system after the effect of all other variables in the system have been eliminated from both of them, by least square correlation methods.

The formulae for the partial correlations of the other variables, b and c, similar to (76) are given:

$$r^2_{xb.ac} = 1 - \frac{\sigma_x^2 (1 - R^2_{x.abc})}{\sigma_x^2 (1 - R^2_{x.ac})}$$

$$r^2_{xc.ab} = 1 - \frac{\sigma_x^2 (1 - R^2_{x.abc})}{\sigma_x^2 (1 - R^2_{x.ab})} \quad \ldots\ldots\ldots (76\text{-a})$$

Note that in these equations the numerator of the fraction, representing the absolute standard error ($\sigma_{z_3}^2$) of estimating the x term, (z_1), from the other, (z_2) does not change for the various partials. The difference in the partial correlation for the different independent variables is thus attributable to changing values in the denominator of

the fraction which represents the variability remaining in x, (z_1), after elimination of the influence of variables indicated by subscripts to R to the right of the point.

The greatest value of a partial will occur when the denominator term becomes largest. The limit is evidently σ_x^2, when $P^2_{x.bc}$ is zero. In this event the partial coefficient becomes equal to the multiple coefficient. This is a useful theorem. Partial coefficients can never be greater than the multiple, or conversely, the multiple coefficient is always as large and usually larger than the largest partial coefficient.

Partial coefficients of correlation as computed by formula (76) may be taken as either the plus or the minus root of the squared value. It is customary to give the coefficient the sign of the net regression coefficient which has subscripts identical to it.

Before leaving the subject of partial correlation the generalized formula may be given.

$$r^2_{xn.abcd\cdots(n-1)} = \left(\frac{(1 - R^2_{x.abc\ldots n})}{(1 - R^2_{x.abc\ldots(n-1)})}\right) \cdots\cdots\cdots\cdots (77)$$

Although the development of the theory of partial correlation coefficients has been given with particular reference to four variables only, it will be recognized that this development is so presented as to be perfectly general in its application to any number of variables. A limited number of variables were employed in order to avoid confusion resulting from too profuse subscript notation otherwise necessitated.

(4) Arithmetic methods.

A complete and very detailed description of the arithmetic methods of working out multiple correlation solutions, whether they be large scale or small scale in scope, has already been prepared [8], and it does not seem advisable to burden this discussion with the bulk of that description. Only the more general points will here be discussed.

There are two laborious tasks in determining net regression coefficients and multiple correlation coefficients. The first is to compute the necessary product moments and standard deviations to make up the normal equations, the second to solve these equations.

Since the preparation of the various product moments and standard deviations necessary to the forming of the normal equations involves the inter-multiplying of all possible pairs of variables and the squaring of all variables, it is eminently practical to reduce these variables to as simple arithmetic values as possible prior to their multiplication. This involves only the coding of these variables as previously explained in connection with gross correlation. A systematic notation of all coding processes should be made, however, in order to avoid confusion at a later time when it comes to decoding the results. Thus if $\bar{C}$ represents the code of any value, C; the value of $\bar{C}$ in terms of C should be noted. This relationship can always be expressed in the form of a linear equation, i. e., $\bar{C} = k_1 + k_2 C$. When it comes time to write the multiple regression equation in terms of original values, it is only necessary to substitute for $\bar{C}$ its equivalent, $k_1 + k_2 C$ and simplify the equation, involving

---

[8] Smith, B. B. Use of Punched Card Tabulating Equipment in Multiple Correlation Problems. U. S. Dept. Agr., Bur. Agr. Econ., 1923. 24 pp. Mimeographed.

only ordinary algebraic processes.

The process of multiplying together and squaring the variables, or "making the extensions" is simplified if the coded values are used directly, rather than the deviations from means. This, requires, of course, the use of formulae such as (29) and (31) in the computation of the product moments and standard deviations. But each investigator will quickly work out for himself methods of systematizing these processes, or he may find them already prepared for him in the aforementioned publication.

It should be noted, that once the variables are coded, the fact that they have been coded may be forgotten in all phases of the interpretation of results, save only in the case of regression coefficients; correlation coefficients of all kinds, and coefficients of determination will be identical with those that would have been secured were original, uncoded, values employed. Only the regression values are changed by the use of the codes.

After the variables have been coded, and prior to any extension of them, it is advisable to introduce a "check-sum". This check-sum is merely the sum of all associated (coded) values of independents and dependent. It serves as a means of checking all extensions and also carries on through the solution of the normals as an almost complete check on all arithmetical work. The operation of this check-sum in checking extentions may be explained as follows:

Suppose variables, A, B, C, and X and a check-sum, $U = A + B + C + X$; then it obviously follows that

$$S(A^2) + S(AB) + S(AC) + S(AX) = S[a(a + b + c + x)]$$
$$= S(AU)$$

The computation of S(AU) serves to check the computation of $S(A^2)$, S(AB), S(AC) and S(AX).

After the normal equations have been prepared the method of solution advocated is the Doolittle Method 9/. The arithmetic processes of this method as applied specifically to multiple correlation problems may be found in detail in an aforementioned publication of the Bureau of Agricultural Economics.

It may be remembered, however, that any standard method of solving simultaneous equations is valid for the determination of the values of b.

(5) Use of Multiple Correlation Methods for the Fitting of Parabolae.

If it be assumed that the relation of Y to X, X being the dependent, is of the nature of a parabola rather than of a straight line, instead of using approximation methods discussed in considering the correlation index to find this curvilinear relationship, multiple correlation methods may be used.

The assumption is, of course,

$$X = K + b_1 Y + b_2 Y^2 + \dots + b_n Y^n \quad \dots\dots\dots\dots\dots\dots (78)$$

All that is necessary to do is to substitute the appropriate powers of Y for the independent variables in the usual multiple regression equation:

$$x = b_1 a + b_2 b + \dots + b_n n$$

and solve for the values of b. Writing the multiple regression equation in terms of original values, rather than deviations fromaverage, supplies the value of K in (78).

9/ The theory and method of this solution may be found in the following:
Adams, O. S. Geodesy - Application of the Theory of Least Squares to the Adjustment of Triangulation. 1915. U.S. Coast and Geodetic Survey. Spec. Pub. 28.
Wright, T. W. and Hayford, J. F. Adjustment of Observations by Method of Least Squareswith Applications to Geodetic Work. 2d ed. N.Y., 1906.

Since there is a constant, mathematical relationship between Y and its powers, only one curve on a coordinate graph is necessary to describe the relationship, rather than one curve for each variable as in the usual multiple correlation. This curve can, of course, be determined by assuming values of Y in (78) and evaluating for X, which then gives as many points on the curve as one cares to compute. The curve may then be drawn to pass through the points.

If in the process of the analysis of the relation of X to several independent variables, it is assumed that the net relation to one of them is best described by a parabola, rather than a straight line, the necessary powers of the given independent variable may be introduced as new independents and the multiple correlation proceed as usual. The coefficient of multiple correlation may be computed by the usual process. The computation of partial correlation coefficients, wherein the several powers of the given independent are treated jointly, is too complicated to be practical. The coefficient of determination, however, may be secured by simply adding together the several coefficients computed for the various powers of the given independent. The coefficient of part correlation may be computed by substituting for the term,

$$b^2_{xa.bcd...n}\ \sigma^2_{a'}$$

in formula (66) a term which is equivalent to the squared standard deviation of contributions from the various powers of the given variable added together. Supposing that the variable b represented variable

a, squared. The term would then be

$$\frac{1}{n}.S(\underline{b}_1 a + \underline{b}_2 b)^2$$

$$= \frac{1}{n}[\underline{b}_1^2.S(a^2) + 2\underline{b}_1 \underline{b}_2.S(ab) + \underline{b}_2^2.S(b^2)]$$

$$= \underline{b}_1^2\sigma_a^2 + 2\underline{b}_1 \underline{b}_2 p_{ab} + \underline{b}_2^2 \sigma_b^2$$

This value would then be substituted for $\underline{b}_1^2 \sigma_a^2$ in formula (65). Its computation involves only the net regression coefficients and standard deviations and product moment already computed.

The multiple correlation method is of course, adaptable to the fitting of other types of curves susceptible to determination by methods of least squares.

VI. Multiple Curvilinear Correlation 10/

Linear multiple correlation assumed that the functions in the following equation were linear and thus provided weights or regression coefficients defining the slope of the straight lines,

$$x = F_1(a) + F_2(b) + F_3(c) + \ldots + F_n(n) \quad \ldots\ldots\ldots\ldots \quad (79)$$

Curvilinear multiple correlation makes no assumptions as to the nature of the functions save that they may be represented by a smooth curve. It permits the data, of themselves, to reveal the nature of the functions. The functions are found by first assuming certain curves to be descriptive of the functions and then by methods of simultaneous approximation these assumed curves are adjusted and modified so as to give minimum squared residuals when applied to the independents in estimating the dependent.

It is apparent that formula (79) represents a much broader case

---

10/ For original presentation of curvilinear correlation see Ezekiel, Mordecai. A Method of Handling Curvilinear Correlation for Any Number of Variables. Amer. Statis. Assoc. Jour. 19: 431 - 453. 1924.

of the qualitative assumption,

$$x = f(a, b, c \ldots n),$$

than does the multiple linear regression equation.
It is limited, however, in that it assumes that an adding together of the several functions gives a best representation of x, and in this respect may suffer from the same inappropriateness of assumption as in the case of multiple linear correlation.

To visualize the distinction between linear and curvilinear regression curves it is only necessary to recall the graphic method of representing net regression lines discussed in the initial description of multiple net regression. Here, the values of x, corrected for contributions of all variables except the given independent, a, were the ordinates on a coordinate dot chart. The abscissae were the values of the given independent, a. Suppose, now, that the distribution of the dots on this chart was such that it were apparent that a better fit to the dots could be had by constructing a curve, rather than a straight line. A free-hand curve may accordingly be substituted for the net regression line. And in similar manner curves may be substituted for the net regression lines in the dot charts representing the net relation of x to the other independent variables. These curves then describe the functional relation of each independent to the dependent.

Since the first step in the process of determining curvilinear net regression is the construction of these net dot charts, an easy method of preparing these dot charts may be described. Using for illustrative purposes the four variables, x, a, b, c which are deviations

from averages of items in the four series, X, A, B, and C, which are shown in Table 9, the first step is to determine the values of $\underline{b}$ by ordinary multiple correlation methods in the following:

$$x = \underline{b}_1 a + \underline{b}_2 b + \underline{b}_3 c \qquad \text{.....(80)}$$

x, of course, being the dependent.

For the sake of illustration an arithmetic example is given.

Table 9.- Data for illustration[11/]

Item number	A	B	C	X	Sum
1	11	10	9	14	44
2	20	19	15	24	78
3	6	6	0	4	16
4	6	12	6	8	32
5	8	8	26	16	58
6	9	8	8	12	37
7	11	8	8	13	40
8	14	16	16	18	64
9	12	10	0	9	31
10	8	8	8	11	35
11	4	5	10	11	30
12	23	26	26	28	103
13	14	12	10	17	53
14	10	16	14	14	54
15	10	10	15	15	50
16	20	13	20	26	79
17	12	12	12	16	52
18	10	2	8	21	41
19	16	6	5	19	46
20	20	20	30	27	97
Sums	244	227	246	323	1,040
Means	12.20	11.35	12.30	16.15	52.00

Extensions - Preparing Normal Equations[12/]

a - 1	3,504.0	3,196.0	3,463.0	4,515.0	14,673.0
a - 2	2,976.8	2,769.4	3,001.2	3,940.6	12,688.0
a - 3	527.2	426.6	461.8	574.4	1,990.0
b - 1		3,207.0	3,373.0	4,097.0	13,873.0
b - 2		2,576.5	2,792.1	3,666.0	11,804.0
b - 3		630.5	580.9	431.0	2,069.0
c - 1			4,296.0	4,740.0	15,872.0
c - 2			3,025.8	3,972.9	12,792.0
c - 3			1,270.2	767.1	3,080.0
x - 1				6,035.0	19,377.0
x - 2				5,216.5	16,796.0
x - 3				808.5	2,581.0

11/ Taken from table in Mordecai Ezekiel's article cited previously. (See footnote 3).

12/ After the manner shown in Use of Punched Card Equipment cited previously. (See footnote 8).

Table 10.- Normal Equations and Solution 13/

Equation II.	Terms in b_1	Terms in b_2	Terms in b_3	Absolute term	Check sum	b_p	d
I.	527.2	426.6	461.8	574.4	1,990.0		
II.		630.5	580.9	431.0	2,069.0		
III.			1,270.2	767.1	3,080.0		
	527.2	426.6	461.8	574.4	1,990.0		
	-1.0000	-.8092	-8759	-1.0895	-3.7746		
		630.5	580.9	431.0	2,069.0		
		-345.2	-373.7	-464.8	-1610.3		
		285.3	207.2	- 33.8	458.7		
		-1.0000	-.7263	+.1185	-1.6078		
			1,270.2	767.1	3080.0		
			-404.5	-503.1	-1743.0		
			-150.5	24.6	-333.2		
			715.2	288.6	1003.8		
		b_3 =		.4035	767.1	309.5	.3828
	b_2	-.1185	-.2931 =	-.4116	431.0	-177.4	-.2194
b_1	1.0895	+.3331	-.3534 =	1.0692	574.4	614.1	.7596
					$R^2 = S.(d)$ =		.9230
					R =		.96

Proof of b

III. $461.8b_1 + 580.9b_2 + 1270.2b_3 = 767.1$

Coeff xb 493.7 - 239.1 + 512.5 = 767.1

13/ After the manner described in Use of Punched Card Equipment cited previously. (See footnote 8).

Table 11.-Tabulation of residuals, with A, B, and C

	A	B	C	Z_1	Z_2 1/
1	11	10	9	-0.1	+0.1
2	20	19	15	+1.7	+0.1
3	6	6	0	-2.7	+0.1
4	6	12	6	+1.6	+2.5
5	8	8	26	-2.8	-1.9
6	9	8	8	-0.4	+0.1
7	11	8	8	-1.6	-1.4
8	14	16	16	+0.5	+0.1
9	12	10	0	-2.4	+0.1
10	8	8	8	-0.3	+0.1
11	4	5	10	+1.9	+2.1
12	23	26	26	+1.1	-0.4
13	14	12	10	+0.1	+0.1
14	10	16	14	+1.8	+1.1
15	10	10	15	-0.5	-0.9
16	20	13	20	-1.2	-2.4
17	12	12	12	+0.5	+0.1
18	10	2	8	+4.7	+2.6
19	16	6	5	-0.8	-1.4
20	20	20	30	-1.1	-1.9
2	244	227	246	0	-1.1
Means	12.2	11.35	12.3	0	
Regres - sions	1.1054	-.4720	.4179		
$s(Z_1^2)$				63.36	
$s(Z_1^2)/n\sigma_x^2$	=			.0758	
$R_{X.ABC}$	=	$\sqrt{1-.0758}$	=	0.96	

1/ Brought from table 12 for use in obtaining second approximation curves.

Table 12.-Readings of functional relations of X to independents from first approximation curves (in Figure 2)

Observation number	F(A)	F(B)	F(C)	S(F)	S(F) + K[1] = X'	X	Z_2 = X-X'	Z_2^2
1	-1.5	+0.5	-1.0	-2.0	13.9	14	+0.1	.01
2	+9.5	-3.5	+2.0	+8.0	23.9	24	+0.1	.01
3	-7.5	+3.0	-7.5	-12.0	3.9	4	+0.1	.01
4	-7.5	-0.5	-2.5	-10.5	5.5	8	+2.5	6.25
5	-5.0	+1.5	+5.5	-2.0	17.9	16	-1.9	3.61
6	-4.0	+1.5	-1.5	-4.0	11.9	12	+0.1	.01
7	-1.5	+1.5	-1.5	-1.5	14.4	13	-1.4	1.96
8	+2.0	-2.0	+2.0	+2.0	17.9	18	+0.1	.01
9	0.0	+0.5	-7.5	-7.0	8.9	9	+0.1	.01
10	-5.0	+1.5	-1.5	-5.0	10.9	11	+0.1	.01
11	-10.0	+3.5	-0.5	-7.0	8.9	11	+2.1	4.41
12	+13.0	-6.0	+5.5	+12.5	28.4	28	-0.4	.16
13	+2.0	-0.5	-0.5	+1.0	16.9	17	+0.1	.01
14	-2.5	-3.0	+1.5	-5.0	12.9	14	+1.1	1.21
15	-2.5	+0.5	+2.0	0.0	15.9	15	-0.9	.81
16	+9.5	-1.0	+4.0	+12.5	28.4	26	-2.4	5.76
17	0.0	-0.5	+0.5	0.0	15.9	16	+0.1	.01
18	-2.5	+6.5	-1.5	+2.5	18.4	21	+2.6	6.76
19	+4.5	+3.0	-3.0	+4.5	20.4	19	-1.4	1.96
20	+9.5	-3.5	+7.0	+13.0	28.9	27	-1.9	3.61
Sums	+0.5	+4.0	+1.5	+6.0	324.1	323	-1.1	36.59

$$P_{X.ABC} = \sqrt{1 - \frac{S(Z_2^{\,2})}{S(x^2)}} = \sqrt{1 - \frac{36.59}{835.5}}$$

$$= \sqrt{1 - .043746} = \sqrt{.956253}$$

$$= .977$$

---

[1] $K = M_x - \frac{1}{n} S[S(F)]$

$= 16.3 - \frac{6}{30}$

$= 15.9$

Figure 2.- Net relation of x to A,B,& C, for first approximation

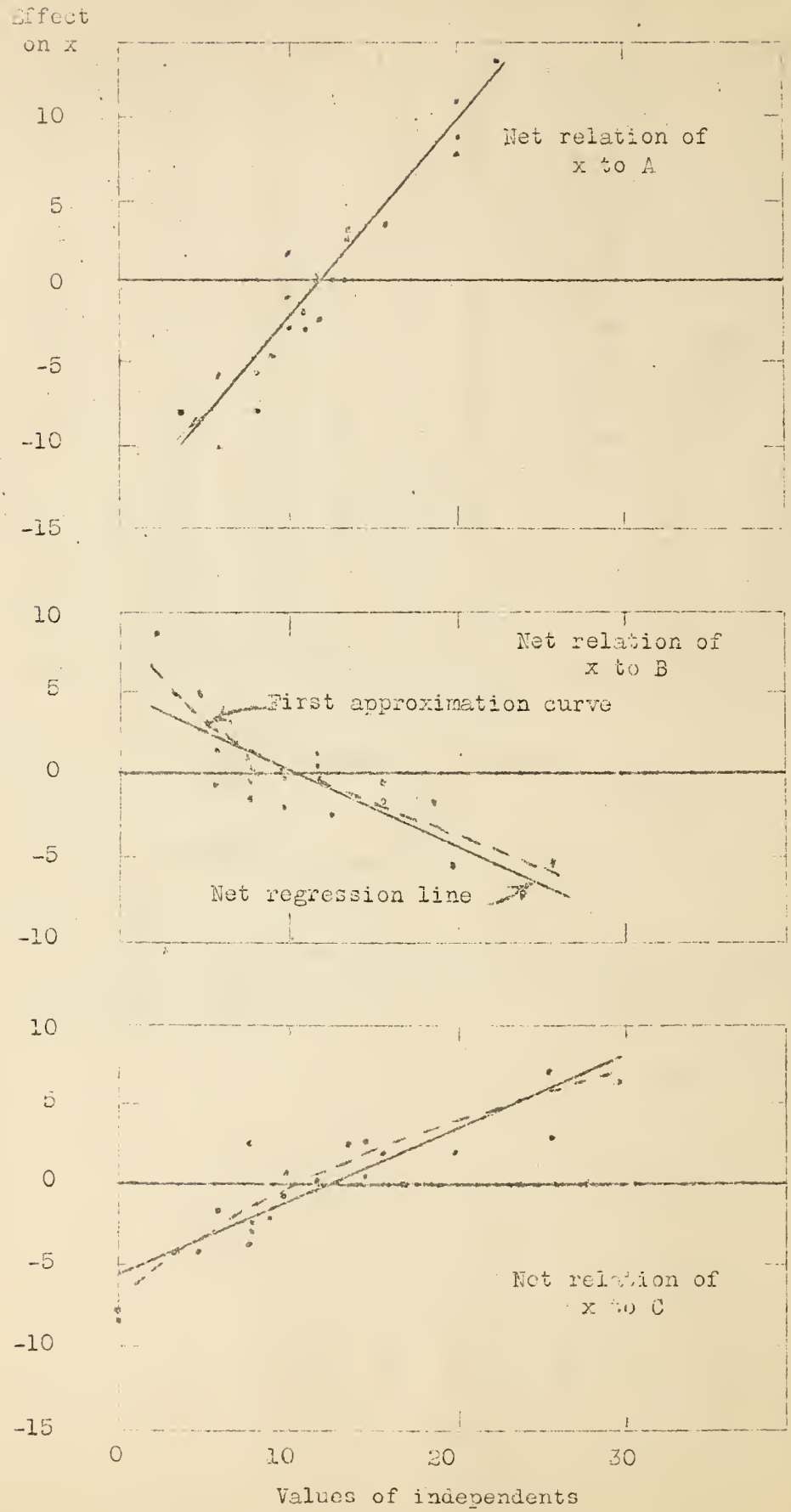
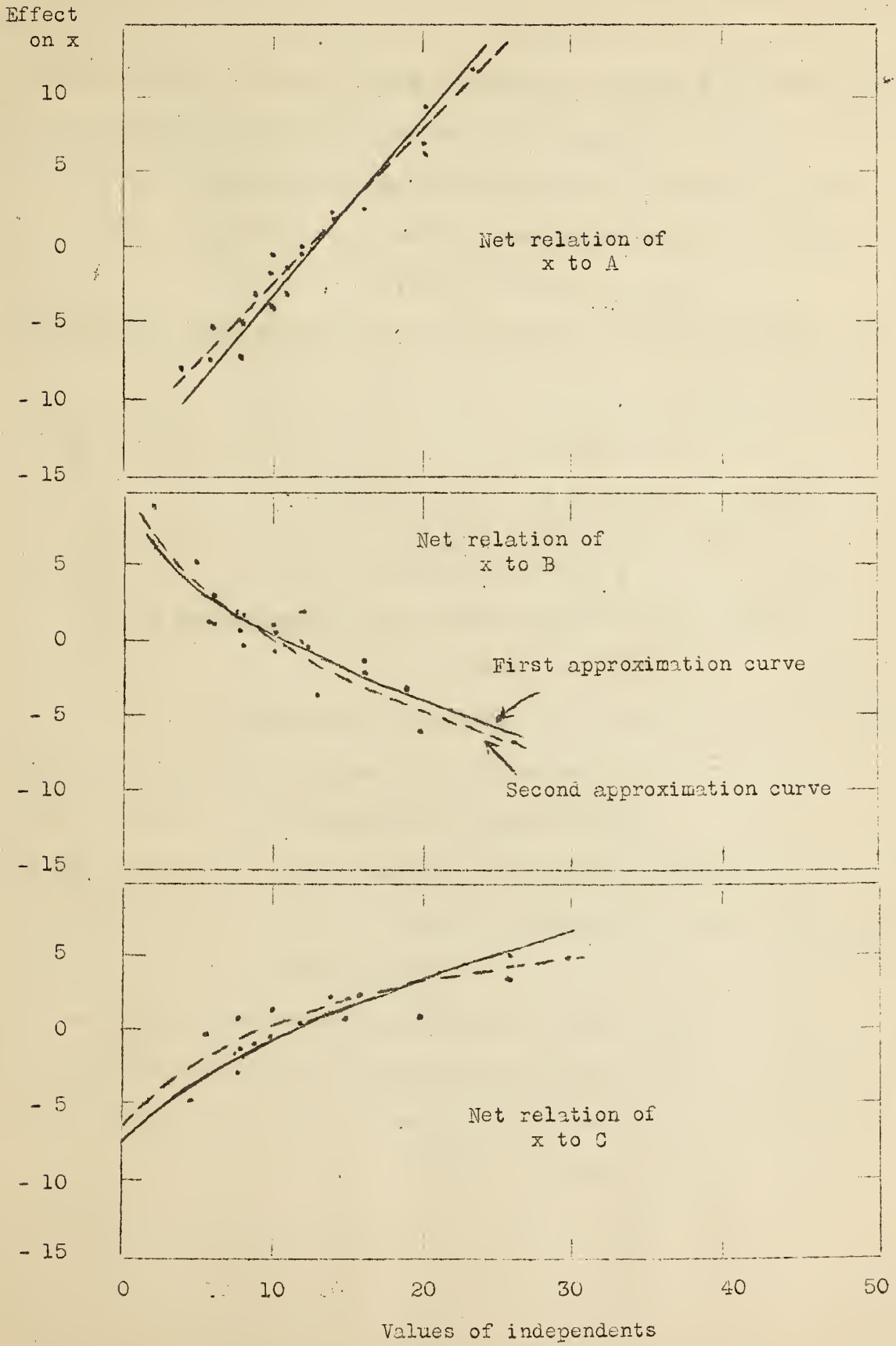


Figure 3. Net relation of x to A, B, and C, for second approximation.



The values of $\underline{b}$ having been found by multiple linear correlation methods, the dot charts may be constructed on the basis of the following considerations. Since the ordinates for the chart showing the relation of a to x are the values of x less the functional terms $\underline{b}_2$ b and $\underline{b}_3$ c, i.e.,

$$(j_1 = x - \underline{b}_2\, b - \underline{b}_3\, c),$$

the ordinates may also be defined as the residuals plus the contributions of a, $\underline{b}_1$ a_1 i. e.,

$$j = \underline{b}_1\, a + z \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (81)$$

For $z = x - \underline{b}_1\, a - \underline{b}_2\, b - \underline{b}_3\, c$

$z + \underline{b}a = x - \underline{b}_2\, b - \underline{b}_3\, c$

$= j_1$, by definition.

Thus, to construct the graph it is only necessary to

(1) Find the values of z

(2) Graph the regression of x on a$(= \underline{b}_1)$

(3) Plot the residuals as ordinate deviations from the regression line with abscissae the associated values of a.

The ordinate value of the regression line takes care of the term $\underline{b}_1$ a in (81) for any given value of a, and it is therefore only necessary to add the value of z to locate the point.

In a similar manner, to make the dot chart showing the net relation of b to x it is only necessary to plot the residuals as ordinate deviations from the regression of x on b with abscissae the associated values of b. And likewise for c.

The advantage of this method of constructing the dot charts is that it saves labor. If j were determined repeatedly for each case, the three following series would have to be secured,

$$j_1 = x - \underline{b}_2 b - \underline{b}_3 c$$

$$j_2 = x - \underline{b}_1 a - \underline{b}_3 c$$

$$j_3 = x - \underline{b}_1 a - \underline{b}_2 b$$

which repeats three times the processes involved in securing

$$z = x - \underline{b}_1 a - \underline{b}_2 b - \underline{b}_3 c$$

The residuals are shown in table II. These residuals were found by simplifying the regression equation interms of original values,

$$X - M_X = \underline{b}_1 (A - M_a) + \underline{b}_2 (B - M_b) + \underline{b}_3 (C - M_c)$$

to read

$$X = K + \underline{b}_1 A + \underline{b}_2 B + \underline{b}_3 C \quad \text{.....(82)}$$

by merely collecting the constant terms to give K,

$$K = M_X - \underline{b}_1 M_a - \underline{b}_2 M_b - \underline{b}_3 M_c$$

and then subtracting the evaluation of this equation (82) from associated values of X. Since this is a comparatively simple process the tables showing the arithmetic are omitted.

As a check on the computation of $R_{x.abc}$ and of z, the standard deviation of the residuals may be determined and used in the formula

$$R = \sqrt{1 - \frac{\sigma_z^2}{\sigma_X^2}}$$

The arithmetic is shown at the bottom of table II.

With the residuals once determined, the charts are laid out, with regression lines of slope, b, and the dots plotted in the manner described. These charts, for the example, are shown in figure 2.

Notice that although the X scale is in terms of deviation from average, the abscissae scales are in terms of original values. This is merely a convenience. In constructing the graphs the abscissae scales might also have been constructed as deviations from average, but this would involve ascertaining the deviation of any given A value before plotting its associated z as a deviation from the regression line. The dots are located in precisely the same spots as if this process had been gone through with, provided only that the regression line is drawn to pass through the point representing the intersection of the mean of the independent and the zero value of X (the latter scale only being constructed to show deviations from average.) This point is the "mean of the distribution."

It is advisable in this connection to point out that once the regression lines have been graphed, these graphs may be used as a means of computing the estimates of X. Thus, for any given value of A, to determine the contribution of A to the estimate of x it is only necessary to read the ordinate of the regression line at the abscissa corresponding to the given value of A. This shows the net deviation from average in X accompanying the given value of A. In like manner the contributions of B and C to the estimate of X may be secured from the appropriate graphs. Summing these three readings together gives the aggregate deviation from average in X that may be attributed to the independent variables. To transform the total or aggregate

deviation from average to an absolute value it is obviously only necessary to add the average of X. In passing, it might be noted that with this last addition an estimate of X, X', is secured. The value Z is secured by subtracting X' from X.

This represents a somewhat clumsy manner of securing estimates of X, and residuals, when we are dealing with linear, multiple correlation. Nevertheless, if, instead of straight lines, the net regression lines were free-hand curves, it would be practically the only way of securing estimates of X. And, indeed, this is the method used as soon as we depart from the representation of the net relation of an independent variable to the dependent by other than a straight line.

This departure is made forthwith by examining the dot charts (figure 2) and observing that in two of the cases, a curved line would give a better approximation to the dots, showing the net relations of X to B and C than do the straight lines. Calling upon his judgment, and remembering that the functional relation is to be expressed by a smoothed curve the operator next draws in these smoothed curves, shown by the broken lines. The straight lines are then superseded by the curves (broken lines) as representations of the relation subsisting between x and B and C. For the present the relation of X to A remains represented by the straight line. The new curves are called the "first approximations".

It next becomes desirable to obtain some measure both of how well these first approximations have been drawn, and of how completely they explain the variation in X. This is to be accomplished by securing estimates of X on the basis of these curves, rather than on the basis

of the first straight lines, and correlating the estimates with the actual X values. The estimates may be secured as explained in the fourth preceding paragraph. It is only necessary to read the functions of the independents from the curves for values of A, B, and C associated with any given X value, and sum them.

The necessary readings and sums are given in table 12. In order to use these sums of functions as estimates of X such a constant should be added to them as would make the average of estimates, $M_{x'}$, equivalent to the average of the actuals, M_x. This constant of course, can be ascertained by taking the difference between the average of the sums of functions and the average of X; thus

$$K = M_x - (1/n)\,S\,[S(F)] \qquad \ldots\ldots\ldots\ldots (83)$$

and the new equation for estimating X may be written

$$X' = K + F_1'(A) + F_2'(B) + F_3'(C) \qquad \ldots\ldots\ldots\ldots (84)$$

Correlating X' with X gives the multiple correlation index, $P_{X.ABC}$. If the curves are better representations of the relationship than the straight lines, the correlation index should be higher than the correlation coefficient. The correlation index is conveniently computed from new residuals, because: as we shall observe later, it is desirable to compute these new residuals. The new residuals may, of course, be computed by subtracting the estimates from the "actuals", i.e.

$$Z_2 = X - X'.$$

The computation of these residuals is shown in table 12. The multiple correlation index is then found by the formula

$$P^2_{X.ABC} = 1 - \frac{S(Z_2^2)}{S.(x^2)} \qquad \ldots\ldots\ldots\ldots (85)$$

which is recognizable as precisely analogous to the familiar formula (56).

In considering the function curves it comes to mind that if there is a high degree of correlation between any two of the independent variables, the process of drawing smoothed curves to pass through as many of the dots as possible may be overdone, for if there is a grouping of positive residuals for one independent causing the operator to draw a curve so as to pass through that group, there is apt to be the same grouping of positive residuals for the closely related second independent causing the operator to pass a curve through that group again. Thus, the same deviation may be doubly accounted for by being attributed to the two independents. In order to test for this case, second approximations may be secured in a manner identical to the securing of the first approximations. Thus, new dot charts are to be constructed in which the ordinates are the values of X corrected for the influence of all independents (as defined by the constructed curves) except the given one, and the abscissae the associated values of the given independent. If the first approximation curve is entirely satisfactory the dots will group themselves closely around the superimposed first approximation curve showing the relation of the dependent to the given independent.

Just as there was simplification in the process of making the first dot charts by plotting residuals as ordinate deviations from the regression lines for associated values of the given independent, it is likewise desirable to construct the charts in this case by plotting the new, or "second" residuals, Z_2, as ordinate deviations from the first approximation regression curves. To accomplish this it is only necessary to reconstruct the first approximation curves on new graphs and proceed to plot the residuals as given originally in Table 12, and secondarily in Table 11, since they are in

the latter table conjoined with the values of the independents necessary to have in plotting the dots. The new dot charts described are shown in Figure 3.

These show that further improvement in the curves representing the relationship of X to the various independents may be had by modifying the expressed relationships somewhat, as illustrated by the broken curves. The new curves (broken lines) may then be taken as the second approximations.

If it is felt that there still remains room for improvement, third, fourth, fifth and more approximations may be made. In actual practice it is not unusual for eight or ten approximations to be made before the investigator is satisfied that his final curves represent the best possible expression of relationships between the dependent and the independents. As long as it is possible by further approximations to raise the value of rho (P), continued approximations are justified. When rho can no longer be increased by changing the curves it indicates the futility of further approximations.

The method of multiple curvilinear correlation may be summarized as follows: By the usual methods of multiple correlation the net regression lines showing the net effect of each factor upon the dependent are plotted. The values of the dependent as obtained from the regression equation are determined, as are the residuals. These residuals are in turn plotted against each independent factor as deviations from each of the regression lines, the lines then being curved to pass through the plotted points in so far as consistent with the hypothesis of a "smooth curve" function. From these curves the dependents are again estimated by reading

from the curve the dependent values associated with each independent. New residuals are obtained and plotted as deviations from the regression curves, and the process is continued until the residuals can not be reduced further.

It is seen that this method is one of approximation, and as such is not susceptible to the mathematical demonstration of validity and probability as are many other statistical methods. On the other hand, as measured by the closeness with which the dependent values may be estimated from the independent factors and by empirical tests, the method is considerably superior to ordinary linear multiple correlation.

In studying the method of multiple curvilinear correlation, the investigator might observe that the first determination of the linear regression lines is not required of necessity. The process of approximation might be commenced with the dot charts showing the gross relation of the dependent to the various independents. But if this were done the investigator would have to contend with the errors introduced by intercorrelation amongst the independents, cited in a previous paragraph, to an even greater degree than in the method as presented. For if the weighted average slope of curves are first determined by methods of linear multiple correlation, the effect of intercorrelation amongst the independents is eliminated except in so far as deviations from linearity are correlated. It is, therefore, always advisable to let the initial approximation to the functional relations be the net regression lines.

The apportionment of variability in the dependent to the various independent factors may be accomplished in much the same manner as that described for multiple linear correlation. This may be accomplished by taking the functions of independents read from the final curves as

the independent variables rather than the actual variables themselves and correlating with the dependent. Then if product moments and standard deviations be computed, and net regression coefficients of X on the functions of independents, by the usual methods of multiple correlation, all the figures necessary to the computation of coefficients of determination and part correlation will be available. Coefficients of partial correlation, however, would be very difficult to interpret if computed from these series.

It should be noted that the net regressions of the dependent on the functions of the independents should all be 1.0 since the function curves were so constructed that by adding together readings from them (with weights of 1.0) the dependent could be estimated. If a given net regression coefficient should prove to have a value other than 1.0, the given function curve should be modified or "tilted" so that the value of b would come out 1.0. Thus if the value of b should prove to be .8 the curve would have to be changed so that, for any given abscissae the ordinate value would be .8 of what it would formerly. This, then, would cause the net regression, b, to be 1.0.

The curvilinear correlation methods which have been described have enabled us to determine the functional relations in a formula of the type

$$X = K + F_1 (A) + F_2 (B) + \ldots + F_n (n)$$

Methods have also been designed [14] to enable us to determine the functions in relations of the following type

$$F(X) = K + F_1 (A) + F_2 (B) + \ldots + F_n (n) \quad \ldots\ldots\ldots\ldots\ldots\ldots (86)$$

This, of course, is equivalent to

$$X = F_0[K + F_1 (A) + F_2 (B) + \ldots + F_n (n)] \quad \ldots\ldots\ldots\ldots\ldots (87)$$

and it is in this latter form that F_0 is determined.

14/ Bruce, Donald. "On Possible Modifications in the Ezekiel Method for Handling Curvilinear Multiple Correlation" MS filed in Library of the U. S. Dept. of Agriculture.

Formula (86) enables us to reduce somewhat the rigidity of the implicit assumption inherent in formula (79), and thus represents an even further step towards analytical methods free from such assumptions. Thus where (79) propounds that X is a sum of functions of the independents, (86) propounds that some function of X is a sum of functions of independents.

The method of determining F_o in formula (87) is quite simple. In addition to graphing the residuals against each of the independents as ordinate deviations from the net regression lines, the residuals are also graphed against the sum of the functions read from the curves plus the constant term, i.e. against X', as an ordinate deviation from a forty-five degree line on a coordinate chart in which the abscissae and ordinate scales are identical. The X' values are taken as abscissae. The ordinates then represent $F_o(X')$, and, of course, for the purpose of first approximating F_o, the relation between X' and X, or X' and $F_o(X')$ is taken as a "one-to-one" relation; hence the forty-five degree slope. But in later approximations, just as the residuals are plotted as deviations from curves determined by preceding processes, so also they are plotted from the curve determined in the preceding process for F_o. New residuals are of course found by substracting $F_o(X')$ from X, for in reality, $F_o(X')$ is the true estimate of X, rather than X' alone.

VII Joint Relationships. 15/

We have seen how by methods of correlation certain specific cases of the general theorem, that the dependent variable is some function of the independents, may be tested for. These methods have taken care of linear and non linear addative functions. It is now proposed to show how with a limited number of variables and a large number of observations it is possible to completely eliminate the limitations of the above cited methods. In short it is proposed to define the function in the following equation, based solely upon the data themselves and with no suppositions as to addative or other types of relationships and provided only that whatever the function be it changes systematically with changing values of the independents, i.e. it is a "smoothed" function.

$$X = F(A,B) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (88)$$

As in the case of curvilinear multiple correlation, the methods employed are approximation methods and are thus not subject to the same rigid mathematical demonstration of validity that other methods are. Nevertheless it is possible to adapt certain of the measures of relationship to this case and thus obtain measures of agreement. The nature of the relationships discovered can best be represented graphically.

The appropriateness of methods for determining F in (88) as opposed to the other types of relationships discussed may first be considered. Suppose that X represented the yield per acre of a crop, A the rainfall during the growing season, and B the temperature. Then a given deviation in A from its normal would have some effect upon X. But this effect

---

15/ For original treatment of joint relationships in multiple correlations see Ezekiel, Mordecai. Determination of Correlation "Surfaces" in the Presence of Other Variables. MS. submitted to Amer. Statis. Assoc.

on X would be different according to whether or not the temperature were high or low. Thus with warm weather a given increase in rainfall might easily have a more pronounced effect on yield than the same increase in rainfall might have with cold weather. It thus becomes apparent that the discovery of the relationships as represented by F in the following,

$$F(X) = F(A) + F(B)$$

is inadequate, for F(A) changes with values of B. We have here not an additive relationship but a joint relationship. This type of relationship occurs in many cases.

The method of determining F in (88) is to construct a three dimensional diagram, in which the depths are the values of A, the widths the values of B and the heights the values of X. A smoothed surface is then constructed to be as representative of the plotted points as consistent with the hypothesis of a smoothed surface. To do this smoothing graphically requires considerable skill and not a little patience. To assist the process in three variable problems a machine has been constructed. This machine is a board through which holes are bored in a checkerboard pattern. Through these numerous holes are passed rods which may be slipped through them to any given length. One dimension along the board is taken to represent values of one independent, the other dimension the other independent. The rods are pushed through the holes so that the length they protrude is proportional to the average value of the dependent for the associated values of the independents, as indicated by the location of the particular hole on the board. A diagram which will help to visualize the above described machine may be seen on page 246 of G. U. Yule's Introduction to the Theory of Statistics (Sixth Edition, 1922).

If in the process of smoothing, it is decided that a flat surface--a surface with slopes in two directions, such as might be illustrated by tipping a piece of board in two directions--is the most appropriate, these two slopes are nothing but the net regressions of X on the two independents, for manifestly, the position of such a "flat" slope is the plane completely defined by two straight lines intersecting at right angles). This is the plane discussed by Yule in connection with the above cited diagram.

But, if curved slopes are to be introduced of a nature which might be illustrated by a section of a trough or flume, or a section of pipe divided longitudinally, and tilted so as to have specified average slopes with reference to a base, then the curvilinear correlation methods are appropriate.

But if, finally, the slopes are to be illustrated as in the preceding paragraph except that the trough is twisted, as might be accomplished by holding one end steady while the other end were twisted several degrees around the longitudinal axis, then it is essential that the joint function be determined. It is impossible to define the slopes in the two directions independently, for these slopes change as we pass from one edge of the plane to another.

In the machine for assisting in the determination of these slopes, the ends of adjacent protruding rods are connected by threads, which then, to a degree, represent the unsmoothed surface indicative of the relation of the heights to the two base-board measurements. A different colored thread is then connected to the shanks of the rods in such a manner as to represent a more generalized or smoothed surface. The heights of

this surface above the base-board are then taken as the effect on X of the values of the two independents indicated by the location of the rod on the baseboard. The surface may be recorded by constructing a table somewhat similar to a double frequency table. The captions may be the values of A, the stubs the values of B, the body of the table then contains the heights of the surface for the related values of A and B.

The measure of correlation may be obtained by securing the residuals--differences between the heights of the surface and of actual observations of X for associated values of A and B--and using the standard deviations of residuals and of X to obtain the correlation index by formulae described previously.

If there are three instead of two independents in formula (88) the determination of the joint relationship is further complicated. It becomes necessary first to determine the surface showing the relation of A and B to X for a given value of C. Another surface is then determined showing the relation of A and B to X for the next value of C; yet another surface is then determined for the next value of C--and so on. The resulting surfaces may be visualized by imagining the roofs on a row of houses, each roof progressively differing from the preceding with changing values of C. In short, we have here a three dimensional figure taken at points as it has moved through space, with one of the dimensions--the heights--systematically changing as it moved. The lines traced by given points on the surface (defined by their vertical position above given points on the base of the figure), as the figure moves through space, should represent smoothed curves.

In like manner a representation of four independent variables to

the dependent may be had by imagining that the whole row of houses be moved progressively sideways with changing values of the fourth variable, until we have a solid block of roofs. The whole block can then be moved in two directions to represent changing values of two more independents, and so on. These representations of joint relationships are of more use as concepts than they are as methods. It is generally impractical and often impossible to work out such joint relationships involving more than two independent variables.

On the other hand, Mr. Ezekiel has developed a method whereby it is possible to discover such joint relationships of two variables to the dependent, in the presence of other independents. In short, by his method it is possible to determine F in the following:

$$X = F_o (A,B) + F_3 (C) + \ldots + F_n (n) \quad \ldots\ldots\ldots\ldots (89)$$

By this method F' values are first secured in the following by curvilinear methods:

$$X = F_1' (A) + F_2' (B) + F_3' (C) + \ldots + F_n' (n) \quad \ldots\ldots\ldots (90)$$

When the values of F' have been secured, a surface is constructed, either graphically or by means of the model or machine described previously, in which the slope with reference to the A dimension is made equivalent to F_1', and with reference to the B dimension, F_2'. The surface is, of course, a curved surface, but cross sections of the surface taken through any values of either independent are similar in form irrespective of the values of the other independent.

The residuals are next plotted (or represented) as altitude deviations from the surface at points of the surface located vertically above the intersection of associated A and B values on the base. The surface is then warped so as to give the best possible representation of the

location of residuals.

It will be seen that by this method the average slopes of the surface are first determined by means of curvilinear methods, which then serves to show the effect of A and B on X, in so far as these effects are independent of their joint relationship. The warping of the surface then provides for any additional effects due to the joint effect of the two.

Following the determination of this surface it is advisable to recompute residuals and again plot these residuals as deviations from the several ascribed functions to ascertain whether or not further slight modifications should be made in the functions. This approximation process parallels that described for curvilinear correlation, except that in the case of variables A and B a surface rather than a curve is to be smoothed. The final resulting curves and surface are than considered to be F in (89), F_0 being a surface and F_3 .... F_n being curves. Needless to say, several pairings of variables can be carried on simultaneously in the same analysis, if desired. The factors C and D, for example may be paired in a manner similar to A and B.

The measure of correlation, as in previous cases is secured by inserting the $\overline{\sigma_z}^2$ in

$$P^2_{X.(AB)CD...N} = 1 - \frac{\overline{\sigma_z}^2}{\overline{\sigma_x}^2}$$

Coefficients of part correlation and determination are computed by correlating the readings from the curves and surface with X as described for curvilinear correlation, A and B, of course, being considered as one since their combined effect is included by using the readings from the surface as one of the independents.

VIII Application to Time Series

A time series is a series of measurements which have been made on a given phenomenon at different (usually equidistant) points in time. Many series used in economic research, such as average monthly prices, are time series. Many time series have certain peculiar characteristics which require special technique for their statistical description, and enjoin caution in the application of analytical methods in the delineation of relationships.

In general, many time series are characterized by a "trend" movement. Thus there is a gradually increasing production of most commodities, paralleling increasing population, and diminishing costs of production. Some series show a downward trend, such as the manufacture of horse-drawn pleasure vehicles, timber resources, percentage of child mortality, etc. Most such trend movements are basically attributable to a gradually evolving civilization or environment and its multitudinous manifestations. In studying, then, the relation of the price of a commodity to various factors over a period of time, it is well nigh requisite that some account be taken of the influence of these "multitudinous manifestations." But it is obviously impossible to secure statistical measurements of all these influences, and thus obviously impossible to include direct measurements of them in the analysis of the factors influencing the price of the commodity, over the period of time. If, however, we make a certain assumption, this difficulty may in an indirect way be surmounted. This basic assumption is that, since we are in an environment gradually evolving as we pass through time, then the influence of this environment may be

considered statistically as some function of a numerical description of the passage of that time. Of course, to the degree that we are able to obtain direct measurements on significant, related factors, such as supply, consumption, costs of production, etc., to that extent the number of influences in the environment which must be thrown together and empirically measured statistically by the passage of time is diminished. In short the basic assumption is that the aggregate influence of such otherwise unmeasured factors as develop gradually or recurringly may be taken as some function of a numerical description of the passage of time.

If we have no measures of related factors, but only the price series under consideration, then the relation of the price to all related factors, as measured under the blanket assumption that their effect is some function of time, may be determined by finding the regression of price on time as described for gross correlation. This regression line, in time series analysis, is customarily styled the (linear) "trend" of the series. Since the numerical measurement of time is purely arbitrary it is convenient to let its value for the middle observation be zero, and numbering succeeding observations, progressively 1, 2, 3, etc., and preceding observations progressively -1, -2, -3, etc. This simplifies the computation of the regression since the average, M_t, of the time measurements, T, will be zero. The product moment, p_{xt}, is then

$$p_{xt} = \frac{1}{n}.S(XT) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (91)$$

The standard deviation squared of T is

$$\sigma_T^2 = \frac{1}{n}.S(T^2) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (92)$$

And the regression, b_{xt}, is, of course,

$$b_{xt} = \frac{p_{xt}}{\sigma_t^2}$$

$$= \frac{S(XT)}{S(T^2)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (93)$$

In time series analyses it is customary to call the regression of X on T the annual or monthly "increment" in the trend according to whether the series is an annual or monthly series.

Tables have been constructed (See Mills and Davenport "Problems and Tables in Statistics") so that the value of $S(T^2)$ may be read from the maximum value of T in the series.

If T has values as suggested, S(T) will be zero only when there are an odd number of observations. When there are an even number of observations, S(T) may be made equal to zero by assigning values 1, 3, 5, 7, 9, etc. going forward through the last half of the observations and -1, -3, -5, -7, -9, etc. going backward through the first half of the observations. Tables showing the value of $S(T^2)$ when so numbered have also been prepared.

This computation of the relation of the dependent (price) series to all other factors has been made upon the basis that we have no measurement of those factors other than the passage of time. If, now, we have measurements of the supply, we have two sets of measurements on independent factors, the supply series, A and the <u>otherwise unmeasured</u> factors, T. And it is to be particularly noted that T in this case represents a different group of factors than in the earlier case where we had no measurement of A, since A has been removed from the "otherwise unmeasured" group. The effect of T, therefore, is not necessarily

the same. We now have a multiple correlation problem in which X is the dependent and A and T the independents. Usual multiple correlation technique is applicable. The values of $\underline{b}$ are determined in the following,

$$X = \underline{b}_1 A + \underline{b}_2 T + \underline{k} \quad \ldots\ldots\ldots\ldots\ldots\ldots (94)$$

and interpreted with strict reference to not only the implicit assumptions of the equation, but also to the assumptions involved in using T as a statistical measurement.

Not only are many time series characterized by what may be termed "trend" movements, but many are also characterized by what may be termed "seasonal" movement. Just as trend movements arise from gradually changing environment, seasonal movements arise from a changing and recurring environment. Thus, crops are marketed only in certain times during the year; building is more active in certain months than in others, retail sales have pronounced increases in holiday seasons, prices often reflect changes of a systematic nature as we pass through the year. Just as in the case of trend, it is equally impossible to measure all the factors which contribute to seasonal movements; and hence, in analagous fashion, the influence of such factors is arbitrarily said to be some function of the proportion of the year which has passed--is some function of a series with values from one to twelve, recurring each year. And, just as it was found that in the analysis of the relation of a given dependent series to an independent time series it was desirable to include a trend measurement, so also is it desirable to include a seasonal measurement.

Since, however, the curve representing the influence of the season upon the dependent must come back to its starting level, linear correlation

methods are inadequate. Curvilinear correlation methods should be employed. Thus to (94) should be added a term in S, seasonal, so that it reads,

$$X = \underline{K} + \underline{b}_1 A + \underline{b}_2 T \ldots + S \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (95)$$

This method of handling trend and seasonal in the analyses of time series [16] may be termed a simultaneous determination of trend and seasonal. In passing it may be pointed out that this method differs from procedure followed by many investigators, with whom it is customary to "extract trend and seasonal" from all series and then correlate residuals. In reconciliation of the two methods it may be pointed out, however, that absolute errors of estimating from regression equations will in both cases be identical.

There is another problem which is more acute in the analysis of time series than in many other analyses: that of a changing relationship as we pass through time. Thus the relation of supply to price of a commodity may have changed considerably during the past decade because of the introduction of substitutes, shifts in styles, major changes in costs, etc. Again, there may have been a pronounced change in the normal seasonal curve, owing to the introduction of storage facilities or other numerous factors. This type of change can be adequately delineated by application of the methods cited in the discussion of joint relationships, for it is essentially a problem of the joint effect of T and the given variable. Thus, all that is required is to obtain the surfaces represented

---

16/ See also Smith, B. B. The Error in Eliminating Secular Trend and Seasonal Variation before Correlating Time Series. Amer. Statis. Assoc. Jour. 20: 543-545. 1925.

by F in the following:

$$X = F_1 (A,T,) + F_2(S,T) + F_3 (T) + \ldots + F_n (n,T) \quad \ldots\ldots\ldots (96)$$

To accomplish this first solve for F in

$$X = F_1' (A) + F_2' (S) + F_3' (T) + \ldots + F_n' (n) \quad \ldots\ldots\ldots\ldots (97)$$

by pre-described methods, and then by conjoining the values of T with those of each of the other variables, as described in connection with joint relationships, determine the surfaces, F in (96).

In the analysis of historical price series, it is sometimes desirable to use "undeflated" prices as the dependent variable, and find the best adjustment for price level by using the index of price level as one of the independent variables. After this relation has been determined it may be desired to find the correlation of the price adjusted for price level according to the relation found. That is, instead of finding the multiple correlation of the independent factors (including price level) with price, the object is to take $\frac{\text{Price}}{\text{f (Price level)}}$ as the dependent variable.[17/] In this and in similar cases it may be desirable to correct the dependent for the influence of but one independent as shown by the net regression equation, and then ascertain the multiple correlation between the dependent so treated and the remaining independents.

Thus, having found $\underline{b}$ in

$$x = \underline{b}_1 a + \underline{b}_2 b + \underline{b}_3 c + \ldots + \underline{b}_n n \quad \ldots\ldots\ldots\ldots\ldots\ldots (98)$$

it is desired to find the correlation between

$$(x - \underline{b}_1 a) \text{ and } b, c, + \ldots + n$$

17/ Ezekiel, Mordecai. The Assumptions Implied in the Multiple Regression Equation. Amer. Statis. Assoc. Jour. 20: 405 - 408. 1925.

In a solution for $\underline{b}$ in

$$(x - \underline{b}_1 a) = \underline{b}_2 b + \underline{b}_3 c + \ldots + n, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (99)$$

the $\underline{b}$ constants will obviously take the same values as in (98) above, since the $\underline{b}_1 a$ term having been subtracted from x prior to the new correlation, these values of $\underline{b}$ are the only ones which will give minimum squared residuals. The residuals will thus obviously be the same whether evaluated from (99) or (98) since, in effect, in (99) all the terms are subtracted from x, just as in (98).

The correlation $R_{(x - \underline{b}_1 a) .bc \ldots n}$ may then be determined as follows:

$$\sigma_z^{-2} = \sigma_x^{-2} (1 - R^2_{x.abc \ldots n})$$

and $$\sigma^{-2}(x - \underline{b}_1 a) = \frac{1}{n} . S(x - \underline{b}_1 a)^2$$

$$= \frac{1}{n} [S(x^2) - 2\underline{b}_1 . S(ax) + \underline{b}_1^2 . S(a^2)]$$

$$= \sigma_x^{-2} - 2\underline{b}_1 \rho ax + \underline{b}_1^2 \sigma_a^{-2}$$

Then by a familiar theorem

$$R^2_{(x - \underline{b}_1 a).bc \ldots n} = 1 - \frac{\sigma_z^{-2}}{\sigma^{-2}(x - \underline{b}_1 a)}$$

$$= 1 - \frac{\sigma_x^{-2} (1 - R^2_{x.abc \ldots n})}{\sigma_x^{-2} - 2\underline{b}_1 \rho ax + \underline{b}_1 \sigma_a^{-2}} \quad \ldots\ldots\ldots (100)$$

All values necessary to the evaluation of the coefficient are procured by the original multiple correlation solution.

There is yet another problem which becomes more significant in the correlation of time series than in other types, owing to limited numbers of observations. This is the problem of the reliability of results obtained. In general the coefficient of multiple correlation has been taken as a measure of this reliability. But if we should take a case in which there were as many independent variables as there were observations, it would obviously be possible to find values of net regressions which would result in a perfect multiple correlation. The results, however, would be worthless as an interpretation of underlying economic laws. Thus the true, underlying correlation is confused with the probable correlation that might result from the pure accident of numbers and the possibility of adapting a certain number of functions to purely random series.

A correlation coefficient may be developed which eliminates this purely mathematical probability, and which includes the number of independent variables, m, as related to the number of observations, n, in its expression.

Thus from least square theory (See Merriman, "Method of Least Squares") Z being a residual, and e the error of estimate,

$$e^2 = \frac{S.(Z^2)}{n - m} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (101)$$

But by inserting the equivalent of $S.(Z^2)$ in (98), (98) becomes

$$e = \frac{n\sigma_x^2 (1 - R^2)}{n - m}$$

and

$$e = \sigma \sqrt{\frac{1 - R^2}{1 - m/n}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (102)$$

or a modified error of estimate which may be written, $\bar{E}$. If $\bar{E}$ becomes greater than σ_x it means that the error of estimating from the regression equation for new cases is greater than the standard deviation, and hence worse than merely taking the average of the dependent as the estimate.

If we substitute $\bar{E}$ for the standard deviation of residuals in a familiar formula, then we may secure a coefficient of correlation, $\bar{R}$, modified for the ratio of m to n

$$\bar{R}^2 = 1 - \frac{\bar{E}^2}{\sigma_x^2}$$

$$= 1 - \frac{1 - R^2}{1 - m/n} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (103)$$

which is the requisite formula. The application of this formula to time series analyses will often give the operator pause, and check unwarranted enthusiasm.