The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

Historic, Archive Document

Do not assume content reflects current scientific knowledge, policies, or practices.

.QA278
2
F66
941

STATISTICS AND AGRICULTURE No. 1

APRIL 1941

The RELATIONSHIP OF THE METHOD OF GRAPHIC CORRELATION TO LEAST SQUARES

By

RICHARD J. FOOTE
J. RUSSELL IVES
BUREAU OF AGRICULTURAL ECONOMICS

UNITED STATES DEPARTMENT OF AGRICULTURE

This series is intended to be a medium for presenting the results of agricultural statistical research throughout the Department Edited in the Bureau of Agricultural Economics, with the aid of a Department advisory committee. the series will consist of monographs issued at irregular intervals as material is made available

THE RELATIONSHIP OF THE METHOD OF GRAPHIC CORRELATION TO LEAST SQUARES

By Richard J. Foote and J. Russell Ives [1]

INTRODUCTION

Considerable use has been made of the graphic method of correlation during the past 10 years, particularly in the field of agricultural economics. The method has been popular for several reasons: (1) It indicates graphically the "net" [2] relationships between the variables included in the study, (2) it has been thought to be an exceedingly simple and flexible method of studying curvilinear relationships, (3) it is a relatively simple mechanical substitute for mathematical calculations.

Although it is now more than 10 years since L. H. Bean's article on the graphic method of correlation was published [3], some confusion still exists among the users of the method regarding the correct interpretation of the results obtained in terms of standard mathematical coefficients. It is the purpose of this paper to examine the meaning of the various steps involved in the graphic correlation procedure from the point of view of the Least Squares method.

[1] The authors wish to acknowledge the encouragement given them by F. L. Thomsen, Bureau of Agricultural Economics, and G. S. Shepherd, Iowa State College, in the development of this paper. Suggestions and criticisms from A. Sturges, Bureau of Labor Statistics, and M. A. Girshick, Bureau of Agricultural Economics, have been particularly helpful. Mr. Girshick outlined many of the mathematical proofs, and materially assisted in the preparation of the manuscript.

[2] This term has been used in many discussions of the graphic methods. Its meaning will be explained in the body of the paper.

[3] "A Simplified Method of Graphic Curvilinear Correlation". Journal of the American Statistical Association, 24: 386-97, December 1929. (A mimeographed publication containing essentially the same material and issued by the Bureau of Agricultural Economics is no longer available except in libraries.) The method also is outlined in several textbooks including M. Ezekiel, Methods of Correlation Analysis, John Wiley and Sons, Inc., New York, 1930, and F. L. Thomsen, Agricultural Prices, McGraw-Hill Book Co., Inc., New York and London, 1936.

This article deals with the graphic method as applied to linear relationships or relationships which can be made linear by transformation. 4/ A brief summary of the conclusions presented is as follows:

The method of linear graphic multiple correlation suggested by L. H. Bean essentially is based upon two mathematical principles: (1) The multiple regression equation becomes the equation of a curve when all of the independent variables except one are held constant. In the case of linear regression, the curve is a straight line whose slope is equal to the partial regression coefficient between the dependent variable and that independent variable which is permitted to vary. For this reason the slopes of the drift lines indicate the partial regression coefficient. (2) The method of successive approximation as outlined by Bean is analogous to a mathematical iterative process which converges to the Least Squares solution. Thus, even if an error is made in the first approximations to the regressions, succeeding approximations will tend to yield more and more accurate results. However, the speed of convergence depends chiefly on the size of the error in the first approximation and the size of the correlation between the independent variables. The better the first approximation and the smaller the intercorrelation, the faster will the process tend to converge. The sizes of the intercorrelations are determined by the nature of the variables included in the analysis and hence, once the variables are chosen, very little can be

4/ Although the graphic method was developed primarily to handle curvilinear rather than linear relationships, only the linear case is discussed here for the following reasons: (1) by implication, the form of the relationship is given a priori; (2) the Least Squares method, as usually considered, is applicable mainly to linear relationships. A discussion of the graphic method as applied to curvilinear relationships would be complicated by the fact that its counterpart to the mathematical method would be difficult to exhibit. Consequently any attack on the problem of the graphic method as applied to curvilinear relationships must be made experimentally. Such an experiment has been started in the Division of Statistical and Historical Research.

done graphically to speed up the convergence. However, the accuracy of the first approximations may be greatly enhanced by the use of drift lines.

Mathematical Meaning of Partial Regression

The present discussion will deal with regression analyses as a method of estimating relationships between variables for purposes of prediction. The usual situation in such cases is as follows:

A set of Np quantities $X_{1i}, X_{2i}, \ldots X_{pi}$[5] $(i=1,2, \ldots N)$ are considered. It is assumed that (1) for all values of i, X_{1i} is a chance variable [6] and the quantities $X_{2i}, \ldots X_{pi}$ $(i = 1, 2, \ldots N)$ are known constants. (2) The <u>mean</u> value of X_{1i} is a linear combination of $X_{2i}, \ldots$ Xpi with unknown coefficients. Thus

mean value of $X_{1i} = b_1 + b_2X_{2i} + \ldots + b_pX_{pi}$ $(i = 1, 2, \ldots N)$

where the b's are unknown constants. (3) The variate X_{1i} is uncorrelated with X_{1j} for $i \neq j$.

Under the above assumptions, it can be shown that the best (linear) estimate of the constants $b_1, b_2, \ldots b_p$ is given by minimizing the quantity

$$\sum_{i=1}^{N} \left[X_{1i} - (b_1 + b_2 X_{2i} + \ldots + b_p X_{pi}) \right]^2 .$$

Here the term "best" is used in the sense that the estimates of the b's obtained in this way will have the smallest standard errors.

If in the equation

$$X_1 = b_1 + b_2 X_2 + b_3 X_3 + \ldots + b_p X_p, \qquad (1)$$

constant values were assigned to $X_3, \ldots X_p$, then $b_3X_3 + \ldots + b_pX_p$ would be equal to some constant which could be combined with the constant b_1 to give

[5] The first subscript refers to the type of variable, i.e. price, production, etc., and the second subscript refers to the observation.

[6] Roughly speaking this means that a variety of uncontrollable forces operate to give different values of X_{1i}, each value occurring with a certain, though possibly unknown, frequency.

a new constant K. Equation (1) could then be written as

$$X_1 = K + b_2X_2 \qquad (2)$$

which is the equation of a straight line having a slope equal to b_2. Here b_2, which may be written as $b_{12.34...p}$, is the regression of X_1 on X_2 when X_3,... X_p are constant.

If two or more observations in a scatter diagram of X_1 on X_2 had the same or approximately the same value of X_3,... X_p then an estimate of $b_{12.3...p}$ could be obtained by drawing a best fitting line through them. If this process were repeated for several groups of points having the same X_3,... X_p values, several lines whose slopes are estimates of the same partial regression coefficient, $b_{12.3...p}$, would be obtained 7/. Such lines have been called "drift lines" and their use is an integral part of the graphic method.

GRAPHIC METHOD AS APPLIED TO A LINEAR THREE VARIABLE PROBLEM 8/

Step. 1. The scatter diagram

The first step in the graphic method of correlation as applied to a linear 3-variable problem (data in table 1) is the plotting of the dependent variable, X_1, against one of the independent variables, say X_2, in an ordinary scatter diagram (dots in section A, fig. 1). Since succeeding steps in the process are carried out in terms of vertical deviations, the dependent variable, X_1, is represented along the vertical scale of the chart and the independent variable, X_2, along the horizontal scale. It is essential that each observation, that is, each dot, on this and following

7/ The process of obtaining estimates of $b_{12.3..p}$ from the slope of these lines is equivalent to breaking the total sample into selected sub-samples and obtaining from each of these an independent estimate of $b_{12.3..p}$.

8/ Much of the material given in the next 5 pages has been stated by Bean, Ezekiel and others. It is presented here as a background for the discussion which follows.

charts be labeled so that the corresponding observation for each variable may be identified throughout the process. As it stands, without further analysis, the scatter of the two variables plotted in this first section indicates the simple correlation between X_1 and X_2 (r_{12}).

Table 1.- Data used in figures 1 and 3

Observation	X_1	X_2	X_3	Estimated X_1	
				From mathematically calculated b's	From 2nd approximations 1/
1	52	50	60	53.9	54.5
2	20	38	24	25.6	25.7
3	62	30	50	58.8	57.2
4	30	42	26	24.7	25.5
5	36	50	44	37.2	38.2
6	40	40	40	40.8	40.7
7	42	24	26	38.4	37.0
8	50	38	50	52.7	51.8
9	20	48	22	15.9	17.5
10	60	42	58	58.0	57.5
11	22	30	20	27.5	27.0
Mean	39.4	39.3	38.2	39.4	39.3

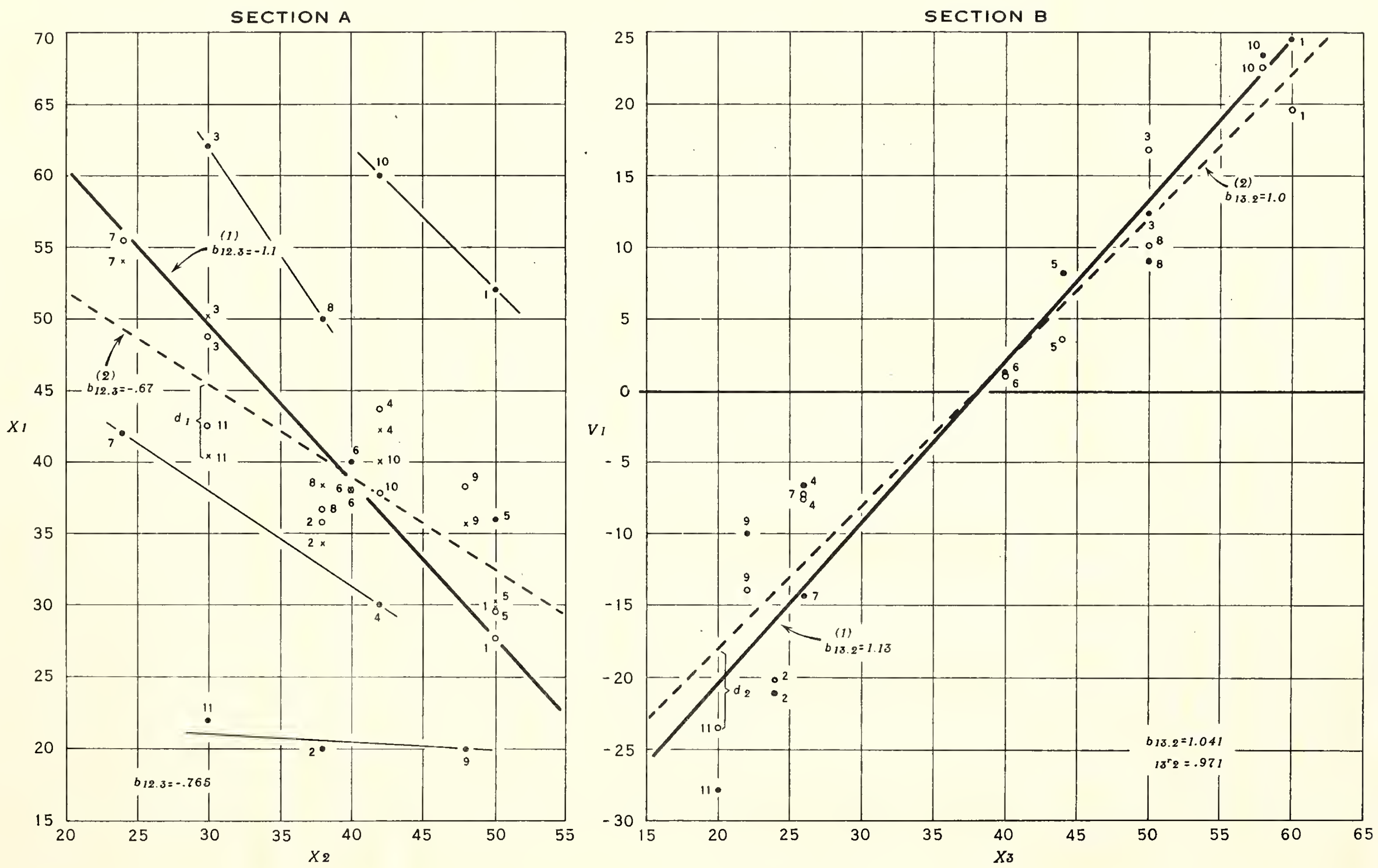
1/ Using graphic method for obtaining estimated X_1 values.

Step 2. The drift lines

As many drift lines as possible are next drawn in this section. For example, from table 1 it will be seen that the pair of observations (3, 8), which have been connected by light lines in section A, are those having identical X_3 values. Likewise, observations 4, 7; 1, 10; and 2, 9, 11 have approximately the same X_3 values and have been connected with light lines.

Step 3. The first approximation to $b_{12.3}$

The next step is to draw through the means of X_1 and X_2 in Section A a line having a slope equal to the average slope of all of the drift lines.



U. S. DEPARTMENT OF AGRICULTURE

NEG. 38822 BUREAU OF AGRICULTURAL ECONOMICS

Figure 1.- Graphic determination of first and second approximations to partial regressions for 'a three variable problem.

The slope of this line is the first approximation to the partial regression, $b_{12.3}$, giving the relation between X_1 and X_2 when X_3 is constant. The degree with which this average slope approximates the regression $b_{12.3}$ will depend on the stability of the slopes of the individual drift lines. In general, the amount of fluctuation that may be expected in the slopes of the drift lines will depend on (1) the number of observations that have the same or approximately the same value for X_3 on the basis of which any one of these drift lines are estimated, (2) the original variability in X_1, (3) the partial correlation of X_1 on X_2 when X_3 is constant.

It is not essential to the mechanical accuracy of the process or to the statistical meaning of the method that the regression line pass through the means of the two series of data. This was done for simplicity. In a later section the modification introduced by not passing the regression line through the means will be explained.

It should be pointed out that the accuracy of the first approximation should not be judged by the goodness of fit of the average drift line to the scatter of X_1 on X_2. This follows from the fact that $b_{12.3}$ is a partial regression and will equal the simple regression b_{12} only when the correlation between X_2 and X_3 is zero [9], which in a sample is an unlikely occurrence.

Step 4. Plotting deviations from the average drift line against X_3

Having obtained an approximation to $b_{12.3}$, the next step in the graphic procedure is the construction of a second scatter diagram in which

[9] This is apparent from the formula

$$b_{12.3} = \frac{b_{12} - b_{13}\, b_{32}}{1 - b_{23}\, b_{32}}$$

which reduces to $b_{12.3} = b_{12}$ when $r_{23} = 0$.

the vertical (perpendicular to the X_2-axis, not to the regression line) deviations of the observations in section A from the average drift line, that is, the first approximation to the line with slope $b_{12.3}$, are plotted against the second independent variable, X_3 (section B, fig. 1). A convenient method for plotting deviations is to take a small card or piece of paper, draw in a zero line near the center of one side, and then take off the plus and minus vertical deviations by moving the zero line on the card along the regression in the first chart and marking off the distance for each observation. If the average drift line passes through the means, then the sum of these deviations will be approximately zero.

A horizontal line, representing the zero value for the deviations, is drawn through the center of the second chart. The exact location of this line and the range for the vertical axis may be easily obtained by placing the marked card on the chart and observing the maximum plus and minus deviations. Then the zero line on the card is moved along the zero line on the chart and the deviation for each observation is inserted above the X_3 value for that observation. An alternative method is to use dividers to transfer the deviations from one chart to the other. This completes the plotting of the second scatter diagram and the best visual estimate of the simple regression for these dots may be fitted to this scatter. 10/

In fitting simple regressions graphically, it must be remembered that the sum of squares of the deviations should be reduced to a minimum. Thus, in general, more weight should be given to large deviations than to small ones. For example, suppose two observations deviated from one guessed regression by 0.4 and 1.0 units, respectively, and from a second guessed

10/ It would be possible to use drift lines based on constant values of X_2 to indicate the slope of this line. The advisability of using this method will be discussed in a later section.

regression by 0.6 and 0.8 units. Although the sum of the absolute deviations is the same for the two regressions, the sum of the squares of the deviations is smaller for the latter, and consequently, the second line is preferable to the first at least so far as these two points are concerned. With practice, the analyst will be able to approximate such a regression closely. If the first regression was put through the means of X_1 and X_2, the second regression should pass through the mean of X_3 at the zero line.

If the deviations from the regression in the first chart are considered as a new variable V_1, so that

$$V_1 = X_1 - b_{12.3}X_2 \tag{3}$$

and if the first approximation to $b_{12.3}$ is equal to the mathematically calculated $b_{12.3}$ for the sample of data, then it can easily be shown that the simple regression between V_1 and X_3 is equal to $b_{13.2}$, that is, the regression between X_1 and X_3 when X_2 is constant. 11/ If the first approximation to $b_{12.3}$ does not equal the mathematically calculated $b_{12.3}$, then the regression in the second chart may not equal $b_{13.2}$. However, the next section shows how the first approximations may be adjusted so that they will more nearly equal the mathematically calculated regression coefficients, $b_{12.3}$ and $b_{13.2}$.

Step 5. The process of successive approximation

The mathematical process of successive approximation is a systematic method of finding the linear regressions in section A and section B, which reduce the sum of squares of the deviations from the regression in section B to a minimum. Since the graphic method duplicates the steps of the mathematical method, successive corrections in the first approximations to $b_{12.3}$ and $b_{13.2}$ will tend to approach the mathematically calculated values.

11/ See note 1 in the Appendix.

The method may be described mathematically as follows: Some value is chosen as a first approximation to $b_{12.3}$.[12] This may be called $b_{12.3}^{(1)}$. Deviations of X_1 from a regression with this slope are related to X_3 and a regression obtained such that the sum of squares of the deviations about it is reduced to a minimum. The slope of this second regression may be called $b_{13.2}^{(1)}$. Deviations of X_1 from the line with slope $b_{13.2}^{(1)}$ are then related to X_2 and a regression obtained such that the sum of squares of the deviations about it is reduced to a minimum. The slope of this regression may be called $b_{12.3}^{(2)}$. Deviations of X_1 from the line with slope $b_{12.3}^{(2)}$ are then related to X_3 and a regression again obtained such that the sum of squares of the deviations about it is reduced to a minimum. The slope of this regression may be called $b_{13.2}^{(2)}$. It can be shown that if this process is continued, each succeeding approximation will be nearer the mathematically calculated value than the preceding one and $b_{12.3}^{(n)}$ and $b_{13.2}^{(n)}$ can be made to come as close to $b_{12.3}$ and $b_{13.2}$ as desired if n is made large enough, that is, if enough successive approximations are made.[13] The factors affecting the speed of convergence will be discussed as soon as the graphic equivalent of this method has been described.

The steps in the graphic method discussed in preceding sections have given first approximations to $b_{12.3}$ and $b_{13.2}$, the former being given by the average slope of the drift lines and the latter by the simple regression between V_1 and X_3. The next step is to take the vertical deviations from the regression in the second chart (fig. 1, section B) and plot them about the

[12] In the graphic method this value is often chosen on the basis of drift lines as discussed in step 3.

[13] See note 2 in the Appendix. In the remainder of the discussion the notations, $b_{12.3}$ and $b_{13.2}$, will be used only for the mathematical values obtained by the usual methods, and a superscript will be given if the value was obtained by some method of successive approximation.

average drift line, that is, the line with slope $b^{(1)}_{12.3}$, in the first chart so that the deviation for each observation is directly above or below the original dot for that observation, that is, retains the same X_2 value. It is advisable to use colored pencils for this purpose so that the deviations can be plotted in a different color than were the original observations. For purposes of discussion, however, it will be assumed that the original observations are indicated by black dots and the deviations from the regression line in section B by circles. Next, a simple regression is drawn through the means of X_1 and X_2 and the circles in such a way that it appears to reduce the scatter about itself to a minimum. This is the dashed line in section A and gives the second approximation, $b^{(2)}_{12.3}$.

It should be noted that the slope of the new regression, $b^{(2)}_{12.3}$, is exactly equal to the slope that would be obtained if X_1 were plotted against X_3, a regression drawn in having a slope equal to $b^{(1)}_{13.2}$, deviations from this regression plotted against X_2, and a simple regression drawn in. The scatter about such a regression would also be exactly equal to the scatter of the circles about the dashed line in the first chart (fig. 1, section A).

This point is so fundamental to the understanding of the graphic method that it seems advisable to give the proof here. The deviations from the line with slope $b^{(1)}_{12.3}$ in the first chart are given by

$$V_1 = X_1 - b^{(1)}_{12.3} X_2 \qquad \text{(Equation 3)}$$

Since in the second chart V_1 is related to X_3, deviations from the line with slope $b^{(1)}_{13.2}$ in this chart are given by

$$\begin{aligned} V_2 &= V_1 - b^{(1)}_{13.2} X_3 \\ &= X_1 - b^{(1)}_{12.3} X_2 - b^{(1)}_{13.2} X_3 \qquad (4) \end{aligned}$$

But the circles in the first chart are equal to the regression coefficient $b_{12.3}^{(1)}$, times X_2 for that observation, plus V_2, so that

$$\begin{aligned} \text{Circles} &= b_{12.3}^{(1)} X_2 + V_2 \\ &= b_{12.3}^{(1)} X_2 + X_1 - b_{12.3}^{(1)} X_2 - b_{13.2}^{(1)} X_3 \\ &= X_1 - b_{13.2}^{(1)} X_3 \end{aligned}$$

Thus, the results are the same as those which would have been obtained if the value of $X_1 - b_{13.2}^{(1)} X_3$ had been obtained in a separate chart and related to X_2 in a separate chart. In other words, the circles and the dashed line in section A give exactly the same picture as would be given in a second chart, with the independent variables reversed, if the longer method had been used.

Now that $b_{12.3}^{(2)}$ has been obtained, deviations of the circles from the line with slope $b_{12.3}^{(2)}$ (the dashed line) are plotted about the line with slope $b_{13.2}^{(1)}$ in section B as circles, keeping the same X_3 values. Then a new regression is drawn in on this chart (the dashed line) such that the sum of squares of the deviations of the circles from it appears to be reduced to a minimum. The slope of this regression is $b_{13.2}^{(2)}$. Deviations from this regression are then plotted about the line with slope $b_{12.3}^{(2)}$ (the dashed line in section A). These deviations may be plotted as X's and a simple regression drawn through them as a dotted line. The slope of this regression is $b_{12.3}^{(3)}$. (In the example used, the dotted line would have coincided with the dashed line and consequently was not drawn in. In this instance $b_{12.3}^{(2)}$ is the final approximation.) The process is once more repeated by taking deviations of the X's from the dotted line with slope $b_{12.3}^{(3)}$ and plotting them about the line with slope $b_{13.2}^{(2)}$. This process is continued until no further correction appears to be needed in $b_{12.3}^{(n)}$, that is, until the simple regression for the deviations from the line with slope $b^{(n)}_{13.2}$ plotted about the line with slope $b^{(n)}_{12.3}$ coincides with the line with slope $b^{(n)}_{12.3}$.

The above discussion shows that the method of successive approximation, as outlined by Bean, is a short-cut graphic method analogous to the mathematical method of successive approximation discussed on page 10. 14/

Speed of Convergence

The speed with which the successive approximations lead to stable results is of interest for two reasons: (1) It takes time to make successive approximations and the charts become messy after several sets of dots have been inserted on them and (2) if the convergence is too slow, the analyst may think that no further correction is needed in the line with slope $b^{(n)}_{12.3}$ when in reality its slope is still quite different from the mathematically calculated value.

By making algebraic substitutions in the mathematical method of successive approximation for a three-variable problem 15/, it can be shown that

$$b_{12.3} - b^{(n)}_{12.3} = r_{23}^{2n-2} \quad (b_{12.3} - b^{(1)}_{12.3}) \text{ and} \qquad (5)$$

$$b_{13.2} - b^{(n)}_{13.2} = -\frac{s_2}{s_3} r_{23}^{2n-1} \quad (b_{12.3} - b^{(1)}_{12.3}) \qquad (6)$$

$$= r_{23}^{2n-2} \quad (b_{13.2} - b^{(1)}_{13.2}) \qquad (7)$$

where $b^{(k)}_{12.3}$ is the kth approximation to $b_{12.3}$, $b^{(k)}_{13.2}$ is the kth approximation to $b_{13.2}$, $b_{12.3}$ and $b_{13.2}$ are the mathematically calculated values of $b_{12.3}$ and $b_{13.2}$ respectively, r_{23} is the correlation between the independent variables X_2 and X_3 and s_2 and s_3 are the standard deviations of X_2 and X_3, respectively.

14/ The fact that graphically determined successive approximations to $b_{12.3}$ will tend to converge toward the mathematically calculated value has been demonstrated in specific cases by various writers on the graphic method. However, so far as is known, no mathematical discussion of this convergence or of the factors affecting the speed of convergence has been presented.

15/ See note 2 in the Appendix.

Equation (5) states that the difference between the mathematically calculated $b_{12.3}$ and any given approximation is equal to a function of the correlation between the independent variables times the error which was made in the first approximation to $b_{12.3}$. It shows that the higher the correlation between the independent variables, the slower will be the speed of convergence. 16/

Equation (5) also indicates the importance of the drift lines, since if the error in $b^{(1)}_{12.3}$ is small, one or two iterations may be enough, but if the error is large and the correlation between the independent variables is also large, 6 or 8 or even more successive approximations may be required to bring the slope of the regression to within 0.1 of the correct value.

As mentioned in footnote 10, drift lines could be used in the second chart as well as in the first. It should be borne in mind, however, that the function of the drift lines is to give a good "guessed" value for $b_{12.3}$ or $b_{13.2}$. Whether they should be used in both charts or in only one would depend upon the availability of drift lines in sufficiently large numbers to give good approximations to the regression lines. It might be well to use them in obtaining first approximations to both $b_{12.3}$ and $b_{13.2}$ and then to use the method of successive approximation as outlined above in the remainder of the analysis.

Illustrations of Effect of Correlation Between the Independent Variables on Speed of Convergence

Although Bean and Ezekiel have always considered the use of drift lines an integral part of the graphic method, it is of interest to investigate what

16/ Using the size of the original error (that is $b_{12.3} - b^{(1)}_{12.3}$) as a base, it can be stated from equation (5) that the percent error left after the nth iteration is given by r^{2n}_{23} times 100. Thus, if $r_{23} = .2$, the error remaining after one iteration is 4 percent of the original error and after two iterations is 0.16 percent while if $r_{23} = .9$, the error remaining after one iteration is 81 percent and after two iterations is 65.61 percent.

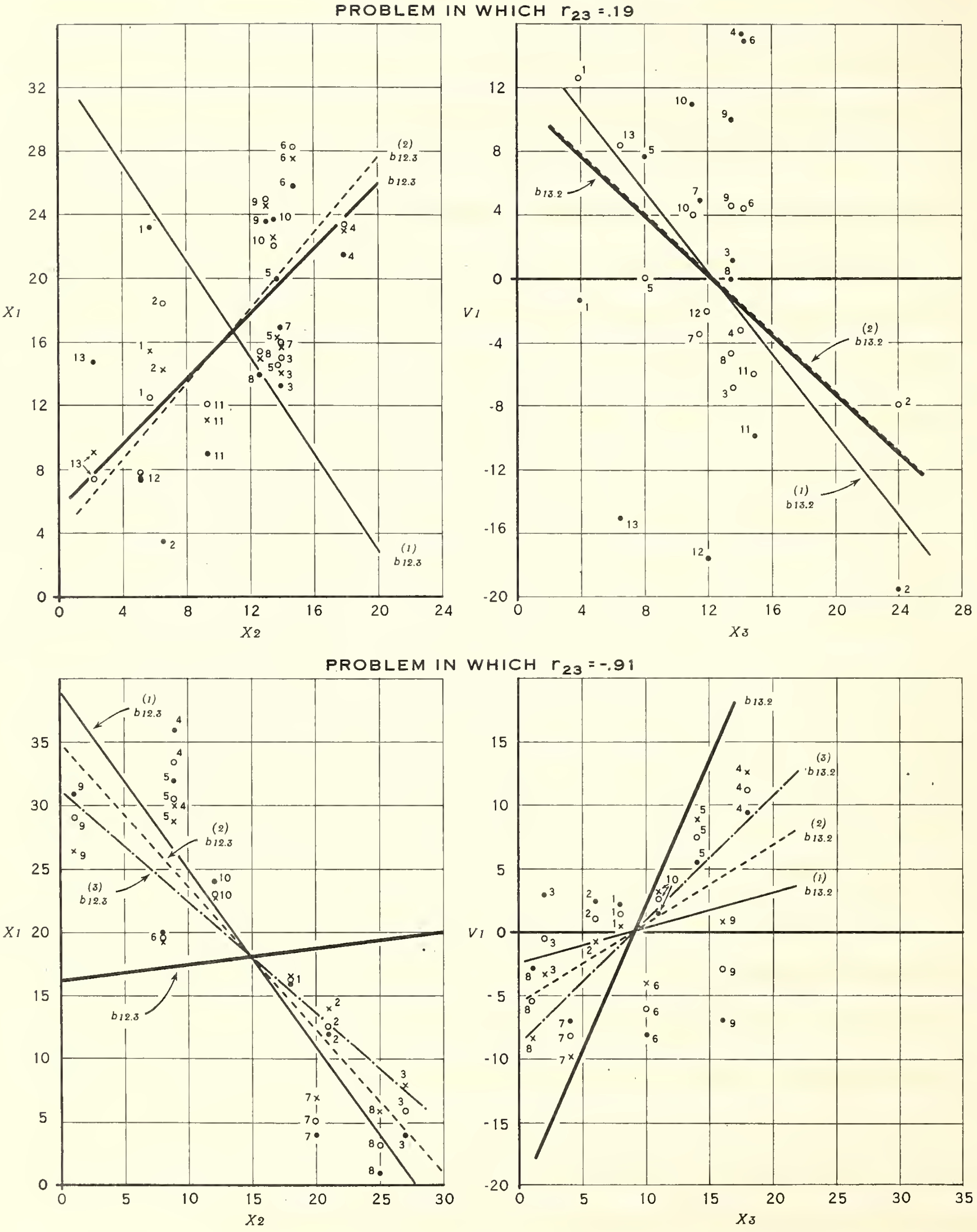
might happen in a particular example if drift lines were not used and completely arbitrary values were chosen as first approximations to $b_{12.3}$. This has been done in figure 2. For purposes of illustration, a line known to hav. a slope materially different from the mathematically calculated $b_{12.3}$ was drawn on the first chart in each example and the problems were then analyzed by the graphic method of successive approximation discussed above.

These examples should not be confused with the example in figure 1, which was handled by the usual graphic correlation technique. Fairly accurat results would have been gotten from each of the examples in figure 2 if drift lines had been used. They simply illustrate the results which may be obtaine if drift lines are not used or if reliable drift lines are not available.

Two examples are shown in figure 2. 17/ For one, r_{23} = .19; for the other, r_{23} = -.91. When r_{23} = .19, the second approximation to $b_{12.3}$ coincides almost exactly with the mathematically calculated $b_{12.3}$. But when r_{23} = -.91 the third approximation is still much closer to the first approximation than to the correct $b_{12.3}$. The same is true of the approximations to $b_{13.2}$. As a check, the mathematical method of successive approximation has been applied to these problems, using the same arbitrary first approximations to $b_{12.3}$. Table 2 shows the results of the analyses.

The above discussion of successive approximation throws light on several controversies that have arisen regarding graphic correlation. Certai users of the method have assumed that the first approximation to the line with slope $b_{12.3}$ could be drawn in without reference to the slopes of the drift lines. In case the correlation between the independent variables is low, this is permissible since the second approximation will correct most of

17/ The data for these problems are given in the Appendix.



PROBLEM IN WHICH r_{23} = .19

PROBLEM IN WHICH r_{23} = -.91

Figure 2.- Graphic illustrations of speed of convergence when drift lines are not used.

the error in the first. However, if the correlation between the independent variables is high, drawing in the first approximation without reference to the drift lines may lead to serious errors. In the first place, if the analyst does not employ drift lines, he is likely to make his first approximation to $b_{12.3}$ approximately equal to b_{12}. It has already been mentioned that $b_{12.3}$ may differ materially from b_{12}, particularly if r_{23} is large. Secondly, if the intercorrelation is high, the second approximation may be about the same as the first, so that the analyst may think he has reached a stable result when the regression still has a slope considerably different from $b_{12.3}$. As the graphic method gives little indication of the size of r_{23}, it is advisable to always draw in as many drift lines as possible.

Table 2.- Values of mathematically calculated successive approximations to the partial regressions for a three-variable problem, when arbitrary values are chosen for $b_{12.3}^{(1)}$

Approximation	When r_{23} = .19		When r_{23} = -.91	
	$b_{12.3}$	$b_{13.2}$	$b_{12.3}$	$b_{13.2}$
1st	- 1.50	- .47	- 1.50	.17
2nd	.86	- .91	- 1.23	.53
3rd	.95	- .93	- 1.00	.83
4th	.96	- .93	- .81	1.07
5th	.96	- .93	- .65	1.28
10th	---	---	- .19	1.90
15th	---	---	- .00	2.15
20th	---	---	.07	2.24
25th	---	---	.10	2.28
30th	---	---	.11	2.30
35th	---	---	.12	2.30
39th	---	---	.12	2.31
Mathematically calculated regression coefficients	.96	- .93	.12	2.31

On the other hand, some have assumed that the drift lines gave them such a close approximation to the correct regression that they have not used the successive approximation procedure, particularly if they were only making a rough analysis to be verified later by the mathematical method. This also appears to be objectionable, especially if the drift lines fluctuate considerably or if only one or two drift lines can be drawn.

In general, the chances for mechanical error are greater in the graphi than in the mathematical method and for this reason the safeguards of both drift lines and successive approximations are advisable. (Bean and Ezekiel have always considered them an integral part of the method.) If the wrong regressions are obtained, the scatter in the final chart may be so large that the particular analysis will be discarded, whereas if the correct or approximately correct regressions had been obtained the deviations from the regression in the final chart would have been smaller and the analysis would have been used or at least worked out mathematically.

The discussion so far also throws light on at least some of the effects of correlation between the independent variables on graphic correlation analyses. It has sometimes been thought that the partial regressions were completely indeterminate by the graphic method if the correlation between the independent variables was high. From the above it can be seen that the indeterminateness may be due entirely to the fact that for high correlations between the independent variables the speed of convergence is slow. With slo convergence the eye may not detect minute changes which should be made in the regressions in order to reduce the sum of squares of the deviations to a minimum and, consequently, the process of successive approximation may be stopped before the correct results have been obtained.

Effect of Not Passing Regressions Through the Means

One further point needs to be considered before the subject of regressions is left.

In Bean's description of the graphic method no mention was made of drawing the regressions through the means of the variables. In his original article, Bean states: "At this point it may be observed that the arbitrary placing of the approximation curves without reference to the average values of X_1 and of the other variables does not affect the values of X_1 computed from the curves. For example, had the approximation curve in section 1 been placed higher, the residuals in sections 2 and 3 would have been correspondingly decreased and the curves lowered." [18] F. L. Thomsen, on the other hand, suggests passing the regressions through the means of the variables in order that each regression will indicate the true net relation between each of the independent variables and the dependent variable in the analysis. [19]

It can be shown algebraically that if a line having a slope equal to $b_{12.3}$ is drawn in the first chart so that it passes through a point d units *above* the mean of X_1 at the mean of X_2, and if deviations from this line are plotted against X_3 in a second chart, the line which will reduce the sum of the squares of the deviations to a minimum in this chart will have a slope equal to $b_{13.2}$ and will pass through a point d units *below* the zero line at the mean of X_3. Moreover, the scatter about this regression will be the same as the scatter would have been if the first regression had been passed through the means of X_1 and X_2 and the second through the zero line and the mean of X_3. [20] Thus, so far as the graphic method itself is concerned, it makes little difference whether the regression in the first chart is drawn through the means or not. It may be preferable, however, in the graphic method, to pass the lines through the means, for when this is done the slope in the second chart is the only constant to be determined and no error is introduced in this chart in locating the intercept.

[18] Bean, *Op. cit.*, p. 393.
[19] Thomsen, F. L. *Op. cit.*, p. 229.
[20] See note 3 in the Appendix.

Step 6. Indicating the multiple correlation 21/

The scatter about the regression line in chart B, figure 1, represents the residual variation in X_1 which has not been accounted for by X_2 and X_3. It has been assumed by some that this final scatter is a direct measure of the multiple correlation. This is true only when the variation is related to the variation in X_1 around its mean. But to compare mentally the total fluctuation in X_1 with the variation remaining in section B is a difficult and untrustworthy method. The best method for obtaining an estimate of the multiple correlation coefficient is to calculate it from the formula

$$R^2_{1.23} = 1 - \frac{\Sigma d^2_{1.23}}{\Sigma X^2_1 - (\bar{X}_1)^2 n} \qquad (8)$$

in which d equals the deviation of any point from the final regression in section B, $\bar{X}_1$ equals the mean of X_1 and n equals the number of observations in the sample.

The multiple correlation may also be indicated graphically by constructing a chart in which the actual values of X_1 are plotted against the estimated values of X_1 obtained from the net relationships between X_1 and X_2 and X_3. Many of the published graphic correlation studies have included charts in which the two series are plotted in the form of a time series (section A, figure 3). A somewhat more reliable visual estimate of the multiple correlation may be gotten from a scatter diagram in which the actual value of X_1 is plotted against the estimated value of X_1 around a line drawn at a 45 degree angle through the origin (section B). The multiple correlation $R_{1.23}$ can be calculated from this scatter by the usual formula for a simple

21/ Most of the material in this and the following section, including equation (8), is given in Ezekiel, Op. cit.

correlation, using X_1 and the estimated value of X_1 as the variables. This, however, involves somewhat more work than the use of the formula given above.

Obtaining the estimated values of X_1

The mathematical formula for obtaining the estimated value of X_1 for any observation is

$$X_1' = \bar{X}_1 + b_{12.3} (X_2 - \bar{X}_2) + b_{13.2} (X_3 - \bar{X}_3). \qquad (9)$$

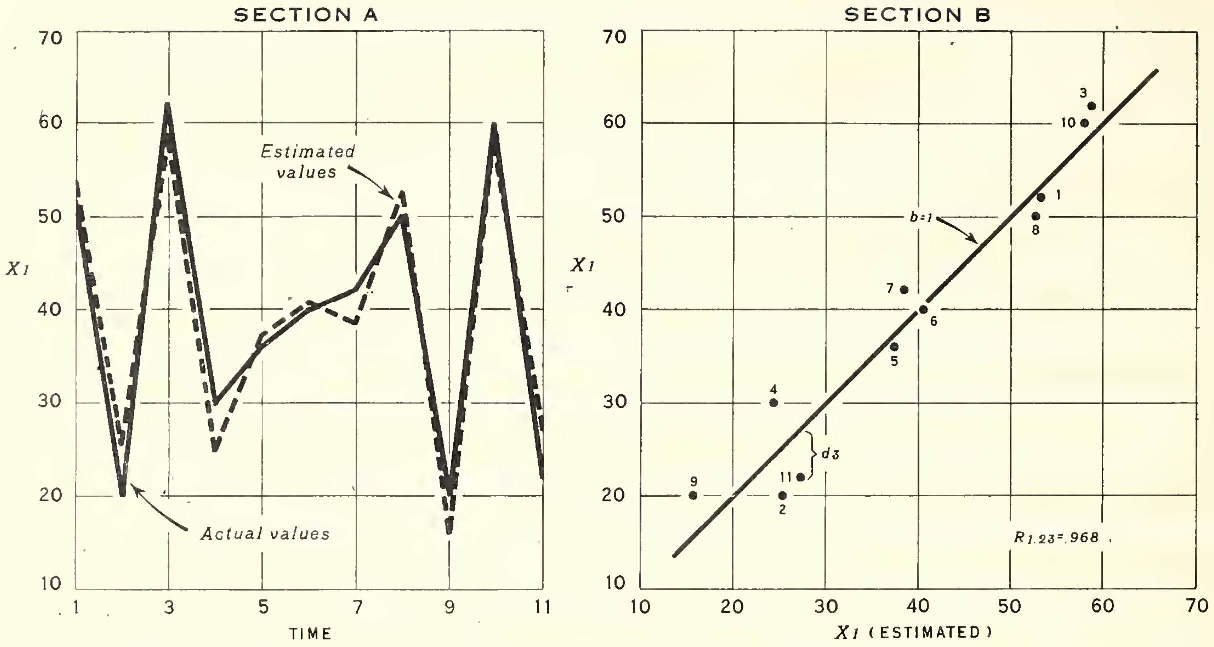
In order to obtain the estimated value of X_1 for any observation, say X_{1i}, graphically, the vertical distance from the zero line in section B, Figure 1, to the final regression line is obtained at the point X_{3i}. This distance, which equals $b_{13.2} (X_{3i} - \bar{X}_3)$, is added to the vertical distance in section A from the point X_{2i} on the X_2 axis to the final regression line in this chart. Since the latter distance equals $\bar{X}_1 + b_{12.3} (X_{2i} - \bar{X}_2)$, the sum of the two distances equals the estimated value of X_{1i} as can be seen from equation (9). These principles are illustrated graphically in figure 4.

Speed of Convergence of Multiple Correlation 22/

As was pointed out above, if the correlation between the independent variables is high, the speed of convergence for the regressions is reduced and regressions considerably different from the mathematically calculated ones may be accepted as the final regressions. Some factors which appear to affect the speed of convergence of the multiple correlation coefficient are discussed here.

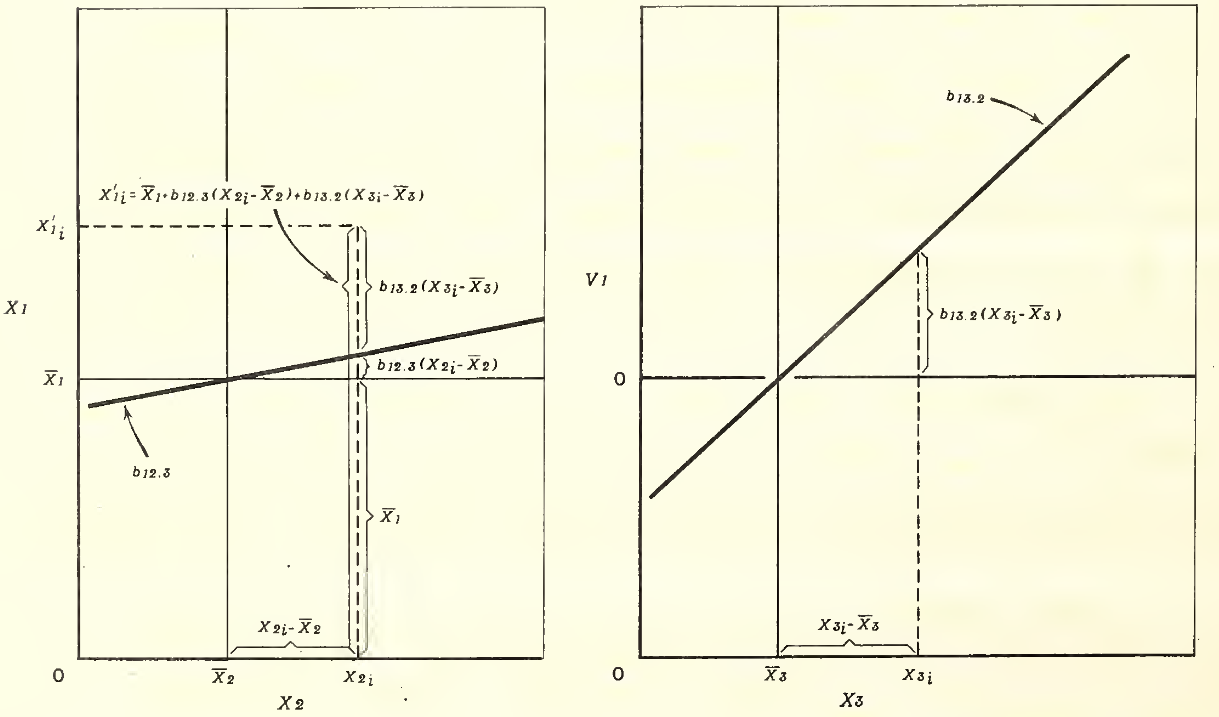
It has been pointed out that the regressions which will make the sum of squares of the deviations from the final regression in the second chart a minimum are the mathematically calculated partial regressions. From equation

22/ The general principles discussed here have been presented by others. See, for instance, Bean, Ezekiel, Malenbaum, and Black. The use of the Short-cut Graphic method of multiple correlation. Comments. Quarterly Journal of Economics. February 1940. p. 336.



U. S. DEPARTMENT OF AGRICULTURE NEG. 38824 BUREAU OF AGRICULTURAL ECONOMICS

Figure 3.- Graphic methods for representing the multiple correlation.


$X'_{1i} = \overline{X}_1 + b_{12.3}(X_{2i} - \overline{X}_2) + b_{13.2}(X_{3i} - \overline{X}_3)$
X'_{1i}
$b_{13.2}(X_{3i} - \overline{X}_3)$
X_1
$b_{12.3}(X_{2i} - \overline{X}_2)$
$\overline{X}_1$
$b_{12.3}$
$\overline{X}_1$
$X_{2i} - \overline{X}_2$
0
$\overline{X}_2$
X_{2i}
X_2
$b_{13.2}$
V_1
$b_{13.2}(X_{3i} - \overline{X}_3)$
0
$X_{3i} - \overline{X}_3$
0
$\overline{X}_3$
X_{3i}
X_3

U. S. DEPARTMENT OF AGRICULTURE NEG. 38825 BUREAU OF AGRICULTURAL ECONOMICS

Figure 4.- Graphic method for determining estimated values of X_1.

(8) it is evident that the size of the multiple correlation coefficient depends upon the size of the deviations from the final regression. Thus, the computed multiple correlation will depend on the regressions obtained, and if the regressions are inaccurate the computed multiple correlation coefficient will be inaccurate. The problem to be determined is how large this inaccuracy can be expected to be and what determines the size of the error.

In order to give a partial answer to the above questions, the multiple correlation coefficients have been computed from the data used in illustrating successive approximations to the regressions (table 2 and table 1 in the appendix). Each of the three methods of computation which have been outlined for obtaining the multiple correlation coefficient has been used, based on the mathematically calculated regressions and on the first four approximations to the partial regressions obtained by the mathematical method of successive approximation. The results are given in table 3.

Table 3.- Successive approximations to the multiple correlation coefficient

Approximation 1/	Problem in which $r_{23} = .19$			Problem in which $r_{23} = -.91$		
	Method A	Method B	Method C	Method A	Method B	Method C
R_1	Undefined	-.308	Undefined	.886	.902	.982
R_2	.804	.805	.782	.919	.926	.983
R_3	.806	.806	.805	.941	.945	.984
R_4	.806	.806	.806	.956	.958	.985
Mathematically computed R	.806	.806	.806	.990	.990	.990

1/ R_n equals the multiple correlation coefficient based on the n'th approximation to $b_{12.3}$ and $b_{13.2}$.

In the following formulas, the variables are taken as deviations from their respective means.

Method A. - Use of the formula, $R_n^2 = 1 - \frac{\Sigma d_n^2}{\Sigma x_1^2}$, in which the deviations have been computed by use of the formula,

$$d_n = x_1 - b_{12.3}^{(n)} x_2 - b_{13.2}^{(n)} x_3$$

Method B.- Calculating the simple correlation between x_1 and the estimated value for x_1, the latter being obtained by use of the formula,

$$\text{Est. } x_1 = b_{12.3}^{(n)} x_2 + b_{13.2}^{(n)} x_3.$$

Method C. - Use of the formula,

$$R_n^2 = \frac{b_{12.3}^{(n)} \Sigma x_1 x_2 + b_{13.2}^{(n)} \Sigma x_1 x_3}{\Sigma x_1^2}$$

(This formula would not be used in working with the graphic method but is frequently used in the mathematical method.)

The conclusions to be drawn are: (1) Some error will be made in the multiple correlation coefficient, if some error is made in the final approximations to the partial regressions, (2) the computed multiple correlation coefficients seem to be different, depending on which formula or method is used for computing them. (If the mathematically computed regressions are used, then the same results will be gotten by each method) and (3) the error seems to vary directly with the error made in estimating the regressions and inversely with the size of the correlation between the independent variables.

Step 7. Interpreting the Correlations Indicated in the Scatter Diagrams

In general, the regression lines obtained in the several charts of the Bean method have been interpreted correctly as "net", that is, partial, regressions between the dependent variable and the separate independent variables. Most of the confusion has come in interpreting the correlations indicated by the plotted observations in the scatter diagrams. This point can be cleared up if one is careful to note the exact meaning of each of the two variables represented by the horizontal and vertical scales of the charts, and considers the "visually indicated" correlation to be the simple correlation of those two variables.

In the first chart, (Section A, figure 1) the two variables represented by the vertical and horizontal scales are simply the dependent variable and one of the independent variables, X_2. Hence, this chart indicates the simple correlation, r_{12}.

With respect to the second chart (figure 1, Section B), if $X_1 - b_{12.3}X_2$ is considered as a variable, V_1, and the simple correlation between V_1 and X_3 is obtained, the resulting correlation will be equal to the part correlation $_{13}r_2$, as defined by Ezekiel. 23/ In this sense we can say that the second chart indicates the part correlation $_{13}r_2$. Likewise, if $X_1 - b_{13.2}X_3$ is considered as another variable, V_2, and the simple correlation between V_2 and X_2 is obtained, that correlation will be equal to the part correlation $_{12}r_3$. If X_2 were used as the second independent variable instead of X_3, the second chart would then indicate $_{12}r_3$. This explains why different correlations are indicated in the final charts when the order in which the

23/ Ezekiel. Op. cit. p. 181-183.

By making certain substitutions in Ezekiel's formula for part correlation, it can be shown that

$$_{13}r^2_2 = \frac{r^2_{13.2}}{1 - (1 - r^2_{13.2})\, r^2_{23}}$$

Since, in the denominator of this formula, the quantity $1 - r^2_{13.2}$ is non-negative and less than unity, the part correlation between two variables is always greater than or at most equal to the partial correlation between the same variables and the difference increases as r_{23} increases.

variables are used is changed. However, part correlations do not appear to have much meaning in the interpretation of an actual problem.

Deviations from the Regressions

Some investigators have been puzzled by the fact that the deviations from the regression lines in certain charts are exactly equal. In the problem outlined here, the deviation for any particular observation of V_1 from a line with slope $b_{13.2}$ in section B, figure 1 would be exactly equal to the deviation of X_1 from the regression in section B, figure 3. Also, if X_1 had been plotted against X_3, a line with slope $b_{13.2}$ inserted, deviations plotted against X_2, and the regression line with slope $b_{12.3}$ drawn in, the deviation from this line for any observation would be exactly equal to the deviation obtained in section B in either figure 1 or figure 3 for that observation. Thus, for example, if the mathematically calculated values for $b_{12.3}$ and $b_{13.2}$ had been obtained, the three quantities, d_1, d_2 and d_3 given in figure 1, section A and B and figure 3, section B, would be numerically equal. This follows from the fact that the deviation for the ith observation in each of these charts is given by

$$d_{1i} = X_{1i} - b_{12.3}X_{2i} - b_{13.2}X_{3i} \qquad (10)$$

APPENDIX

Table 1.- Data for figure 2

Observation	Problem in which $r_{23} = .19$			Problem in which $r_{23} = -.91$		
	X_1	X_2	X_3	X_1	X_2	X_3
1	23.2	5.7	3.9	16	18	8
2	3.6	6.6	24.0	12	21	6
3	13.2	13.9	13.6	4	27	2
4	21.5	17.8	14.1	36	9	18
5	20.0	13.7	8.1	32	9	14
6	25.7	14.7	14.3	20	8	10
7	16.9	13.9	11.5	4	20	4
8	14.0	12.6	13.4	1	25	1
9	23.6	13.0	13.5	31	1	16
10	23.7	13.5	11.0	24	12	11
11	9.0	9.3	14.8			
12	7.7	5.2	12.0			
13	14.7	2.2	6.5			
Mean	16.7	10.9	12.4	18	15	9
Mathematically calculated regression coefficients	$b_{12.3} = .956$ $b_{13.2} = -.927$			$b_{12.3} = .122$ $b_{12.3} = 2.307$		

MATHEMATICAL APPENDIX

In this appendix the following symbols, which may be new to some readers, will be used.

$$a_{12} = \frac{\Sigma(X_1 - \bar{X}_1)\ (X_2 - \bar{X}_2)}{n - 1}\ ,\ a_{13} = \frac{\Sigma(X_1 - \bar{X}_1)\ (X_3 - \bar{X}_3)}{n - 1},\ \text{etc.}$$

$$a_{11} = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n - 1},\ a_{22} = \frac{\Sigma(X_2 - \bar{X}_2)^2}{n - 1},\ \text{etc.}$$

where $\bar{X}_1$ is the mean of X_1, etc. and n is the number of observations in the sample.

The following transformations can be made.

$$a_{12} = s_1\ s_2\ r_{12},\qquad a_{13} = s_1\ s_3\ r_{13},\ \text{etc.}$$

$$a_{11} = s_1^2\quad ,\qquad a_{22} = s_2^2\quad ,\ \text{etc.}$$

where s_1 is the standard deviation of X_1 , etc.

It can be shown that in terms of the standard deviations and correlations

$$b_{12.3} = \frac{s1}{s2}\ \frac{r12 - r_{13}\ r_{23}}{1 - r_{23}^2} \qquad (1)$$

and

$$b_{13.2} = \frac{s1}{s3}\ \frac{r_{13} - r_{12}\ r_{23}}{1 - r_{23}^2} \qquad (2)$$

Note 1. It is desired to determine the simple regression coefficient of V_1 on X_3 when $V_1 = (X_1 - \bar{X}_1) - b_{12.3}\,(X_2 - \bar{X}_2)$.

$$\text{Now } b_{V_1X_3} = \frac{\sum V_1(X_3 - \bar{X}_3)}{\sum (X_3 - \bar{X}_3)^2} = \frac{\sum [(X_1 - \bar{X}_1) - b_{12.3}(X_2 - \bar{X}_2)]\ (X_3 - \bar{X}_3)}{\sum (X_3 - \bar{X}_3)^2}$$

$$= \frac{a_{13} - b_{12.3}\, a_{23}}{a_{33}}$$

Substituting the value of $b_{12.3}$ from equation (1) and simplifying,

$$b_{V_1X_3} = \frac{s_1}{s_3}\ \frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}$$

$\therefore$ By equation (2),

$$b_{V_1X_3} = b_{13.2}$$

Note 2. Iterative Process

In the method of least squares, the coefficients $b_{12.34}$, $b_{13.24}$ and $b_{14.23}$ are determined by minimizing the quantity

$$f(b_{12.34}, b_{13.24}, b_{14.23}) = [(X_1 - \bar{X}_1) - b_{12.34}(X_2 - \bar{X}_2) - b_{13.24}(X_3 - \bar{X}_3) - b_{14.23}\,(X_4 - \bar{X}_4)]^2$$

The solution yields the following three normal equations.

$$b_{12.34}\, a_{22} + b_{13.24}\, a_{23} + b_{14.23}\, a_{24} = a_{12} \qquad (3)$$

$$b_{12.34}\, a_{23} + b_{13.24}\, a_{33} + b_{14.23}\, a_{34} = a_{13} \qquad (4)$$

$$b_{12.34}\, a_{24} + b_{13.24}\, a_{34} + b_{14.23}\, a_{44} = a_{14} \qquad (5)$$

These equations can be solved by well known methods.

In the iterative process, values $b^{(1)}_{12.34}$ and $b^{(1)}_{13.24}$ are guessed for $b_{12.34}$ and $b_{13.24}$ respectively in equation (3) and a solution for $b_{14.23}$

(say $b_{14.23}^{(1)}$) is obtained from this equation. Then $b_{13.24}^{(1)}$ and $b_{14\ 23}^{(1)}$ are substituted in equation (4) for $b_{13.24}$ and $b_{14.23}$ respectively and a second approximation for $b_{12.34}$ (say $b_{12.34}^{(2)}$) is obtained. The values $b_{12.34}^{(2)}$ and $b_{14.23}^{(1)}$ are substituted in equation (5) for $b_{12.34}$ and $b_{14.23}$ respectively and a second approximation to $b_{13.24}$ (say $b_{13.24}^{(2)}$) is obtained. The values $b_{12.34}^{(2)}$ and $b_{13.24}^{(2)}$ are substituted in equation (3) for $b_{12.34}$ and $b_{13.24}$ respectively and a second approximation to $b_{14.23}$ (say $b_{14.23}^{(2)}$) is obtained. This process is repeated until the coefficients converge to stable values.

The iterative process outlined above is equivalent to the following steps: (1) Assign values $b_{12.34}^{(1)}$ and $b_{13.24}^{(1)}$ to $b_{12.34}$ and $b_{13.24}$ respectively in the function $f(b_{12.34}, b_{13.24}, b_{14.23})$ defined above. Find that value of $b_{14.23}$ which makes $f(b_{12.34}^{(1)}, b_{13.24}^{(1)}, b_{14.23})$ a minimum. Let it be $b_{14.23}^{(1)}$. (2) Find that value of $b_{12.34}$ which makes $f(b_{12.34}, b_{13.24}^{(1)}, b_{14.23}^{(1)})$ a minimum. Let this value be $b_{12.34}^{(2)}$. (3) Find that value of $b_{13.24}$ which makes $f(b_{12.34}^{(2)}, b_{13.24}, b_{14.23}^{(1)})$ a minimum. Let this value be $b_{13.24}^{(2)}$. (4) Find that value of $b_{14.23}$ which makes $f(b_{12.34}^{(2)}, b_{13.24}^{(2)}, b_{14.23})$ a minimum. Let this value be $b_{14.23}^{(2)}$, etc. It will be seen that the steps involved in the latter process are identical with those outlined on pages 10-12 for the graphic method involving three variables.

Speed of Convergence

The problem of the speed of convergence will be considered for three variables only. The normal equations for three variables are given by

$$b_{12.3}\, a_{22} + b_{13.2}\, a_{23} = a_{12} \qquad (6)$$

$$b_{12.3}\, a_{23} + b_{13.2}\, a_{33} = a_{13} \qquad (7)$$

If the iterative process is performed on these two equations, then the nth approximation to $b_{12.3}$ and $b_{13.2}$ respectively can be shown to be equal to

$$b_{12.3}^{(n)} = \frac{s_1}{s_2}\left(r_{12} - r_{13}r_{23} + r_{12}r_{23}^{2} - r_{13}r_{23}^{3} +- \dots - r_{13}r_{23}^{2n-3}\right) + b_{12.3}^{(1)}r_{23}^{2n-2} \qquad (8)$$

and

$$b_{13.2}^{(n)} = \frac{s_1}{s_3}\left(r_{13} - r_{12}r_{23} + r_{13}r_{23}^{2} - r_{12}r_{23}^{3} +- \dots + r_{13}r_{23}^{2n-2}\right) - \frac{s_2}{s_3}\, b_{12.3}^{(1)}\, r_{23}^{2n-1} \qquad (9)$$

$$= \frac{s_1}{s_3}\left(r_{13} - r_{12}r_{23} + r_{13}r_{23}^{2} - r_{12}r_{23}^{3} +- \dots - r_{12}r_{23}^{2n-3}\right) + b_{13.2}^{(1)}r_{23}^{2n-2} \qquad (10)$$

where $b_{12.3}^{(n)}$ and $b_{12.3}^{(1)}$ are the nth and the 1st approximations respectively to $b_{12.3}$ and $b_{13.2}^{(n)}$ and $b_{13.2}^{(1)}$ are the nth and the 1st approximations respectively to $b_{13.2}$.

But

$$b_{12.3} = \frac{s_1}{s_2}\left(r_{12} - r_{13}r_{23} + r_{12}r_{23}^{2} - r_{13}r_{23}^{3} + - \dots\right) \qquad (11)$$

and

$$b_{13.2} = \frac{s_1}{s_3}\left(r_{13} - r_{12}r_{23} + r_{13}r_{23}^{2} - r_{12}r_{23}^{3} + - \dots\right) \qquad (12)$$

which can be obtained by expanding the denominator of

$$b_{12.3} = \frac{s_1}{s_2}\,\frac{r_{12} - r_{13}r_{23}}{1 - r^{2}\,_{23}} \qquad (1)$$

and

$$b_{13.2} = \frac{s_1}{s_3}\,\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^{2}} \qquad (2)$$

in an infinite series.

Hence, comparing equations (8) with (11) and (9) or (10) with (12) it will be seen that $b_{12.3}^{(n)}$ and $b_{13.2}^{(n)}$ can be made to approximate $b_{12.3}$ and $b_{13.2}$ respectively as closely as desired by taking n sufficiently large.

The difference between the mathematically calculated value of $b_{12.3}$ and the nth approximation is given by $b_{12.3} - b_{12.3}^{(n)} = r_{23}^{2n-2}\left(b_{12.3} - b_{12.3}^{(1)}\right)$ (13)

and the difference between the mathematically calculated value of $b_{13.2}$ and the nth approximation is given by

$$b_{13.2} - b_{13.2}^{(n)} = - \frac{s_2}{s_3} r_{23}^{2n - 1} (b_{12.3} - b_{12.3}^{(1)}) \qquad (14)$$

$$= r_{23}^{2n - 2} (b_{13.2} - b_{13.2}^{(1)}) \qquad (15)$$

It is evident from equation (13) that the speed of convergence of successive approximations to $b_{12.3}$ varies directly with the error made in $b_{12.3}^{(1)}$ and inversely with the size of r_{23}. Likewise from equation (15), the speed of convergence of successive approximations to $b_{13.2}$ varies directly with the error made in $b_{13.2}^{(1)}$ and inversely with the size of r_{23}.

Note 3. It is desired to determine the effect on the partial regressions and the multiple correlation coefficient of passing the regressions through some points other than those determined by the means.

In a three variable problem, the regression equation is given by

$$X_1 = \bar{X}_1 + b_{12.3} (X_2 - \bar{X}_2) + b_{13.2} (X_3 - \bar{X}_3)$$

Let a line with slope equal to $b_{12.3}$ be drawn in the first chart of the graphic method. Let the line be drawn d_1 units above the mean of X_1 at the point $(\bar{X}_1, \bar{X}_2)$. This line is then given by

$$X_1 = d_1 + \bar{X}_1 + b_{12.3} (X_2 - \bar{X}_2)$$

[The constant d_1 would be zero if the line passed through the point $(\bar{X}_1, \bar{X}_2)$.]

Let $V_1 = X_1 - d_1 - \bar{X}_1 - b_{12.3}(X_2 - \bar{X}_2)$

Then in the second chart V_1 can be written as a regression on X_3 in the form

$$V_1 = d_2 + c (X_3 - \bar{X}_3)$$

If the least squares method is used for determining d_2 and c, it will be found that $c = b_{13.2}$ and $d_2 = - d_1$.

Thus it is seen that if the line with slope $b_{12.3}$ is passed through a point d_1 units above the mean of X_1 at the point $(\bar{X}_1 \bar{X}_2)$, the second regression will have a slope equal to $b_{13.2}$ and will pass through a point d_1 units below the zero line at the point $(0, \bar{X}_3)$. Thus

$$V_2 = V_1 - \left[- d + b_{13.2}(X_3 - \bar{X}_3) \right]$$

$$= (X_1 - \bar{X}_1) - b_{12.3}\,(X_2 - \bar{X}_2) - d + d - b_{13.2}\,(X_3 - \bar{X}_3)$$

$$= (X_1 - \bar{X}_1) - b_{12.3}\,(X_2 - \bar{X}_2) - b_{13.2}\,(X_3 - \bar{X}_3) \qquad (16)$$

This is the same as would have been obtained if the regressions had been passed through the means. Therefore, the multiple correlation is not affected by passing the regressions through points other than those determined by the means.