CHAPTER II

ELEMENTARY OUTLINE OF THE METHOD OF CORRELATION ANALYSIS

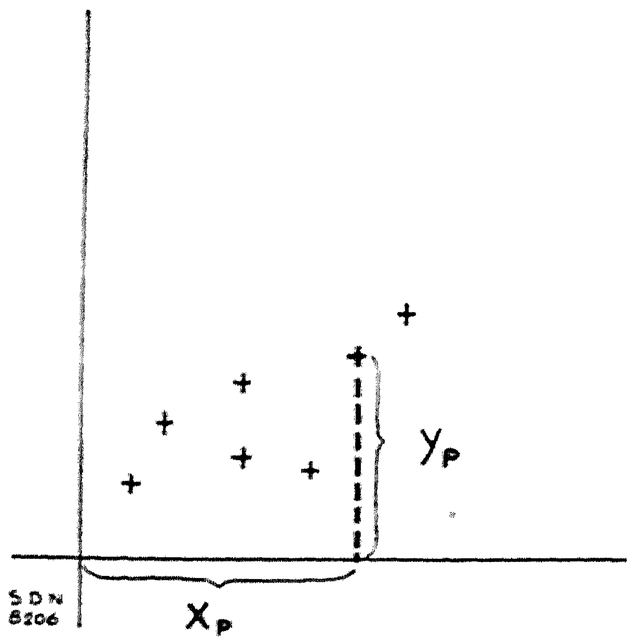
§ 4. Simple Correlation

As has been pointed out in the previous chapter, the object of correlation analysis is twofold: (1) to test whether some expected relation between two or more variates exists (verification) and, (2) if so, to find the strength of the influences exerted by each causal phenomenon (measurement). The exact meaning of these terms and the consecutive steps in the analysis will now be discussed. It seems useful to begin with simple correlation.

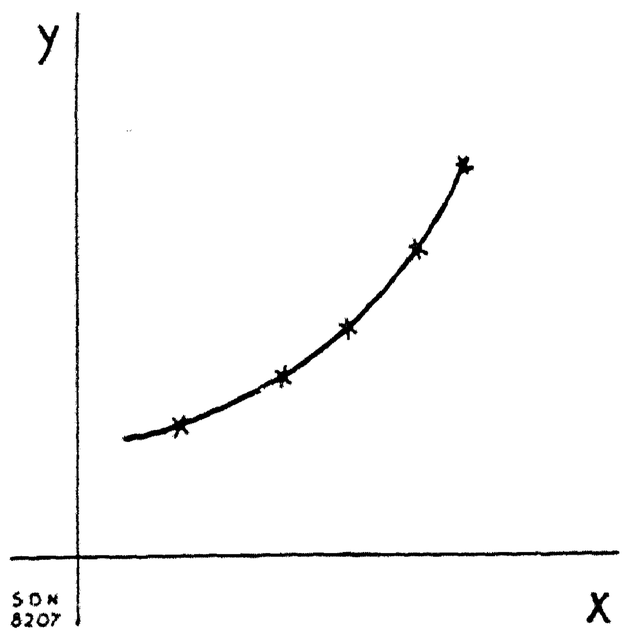
Simple correlation.

Simple correlation is expected to exist if the fluctuations in any series Y are supposed to be caused (or chiefly caused) by the fluctuations in only one other series X. The simplest type of analysis that can be made in this case is to draw a scatter diagram. In such a

Graph II. 1.
Scatter Diagram.



Graph II. 2.
Perfect Correlation.



Regression curves and coefficients.

The curve through the points is called the *regression curve.* If it is a straight line, its slope is termed the regression slope and may be measured by a "*regression coefficient*": this coefficient indicates the increase in Y which corresponds to a unit increase in X. In the table below, which indicates corresponding values for X and Y, arranged in ascending order, a unit increase in X clearly corresponds to an increase of 2 in Y. The regression coefficient is therefore 2. The relation between X and Y may also be described by the formula $Y = 10 + 2X$.

X	Y
10	30
11	32
12	34
13	36
14	38
15	40

Lags.

"Corresponding values" of X and Y will often be values for the same period. In some cases, however, the relation is between values of X and later values of Y. The time difference between corresponding values of X and Y is called the lag; Y lags behind X or X leads Y. It will be clear that if X is cause and Y effect, then X will lead Y. This fact may sometimes be used in order to find out which of two series is cause, which effect.[1]

The provisional determination of lags is best done with the help of an historical graph, showing the development in time of both series.

An example is to be found in Graph II.4, where two series have been drawn representing:

(A) Total volume of non-farm residential building in the United States, 1920-1935.

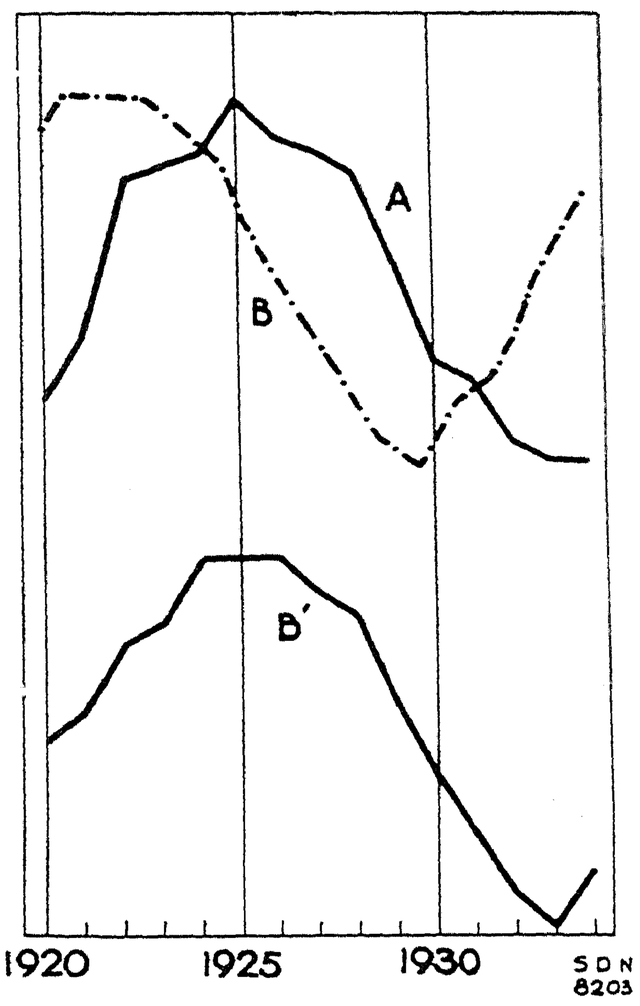
---

[1] One has, however, to be careful: it may happen, *e.g.*, that X leads Y, so that it would seem as if X were cause, Y effect. At the same time, however, $\dot{Y}$ (the rate of increase in Y) may lead X, and therefore $\dot{Y}$ may equally well be cause of X. Finally, it is possible that both causal connections exist: $\dot{Y}_t$ determining X_{t+1} and X_{t+1} determining Y_{t+2}.

(B) Total stock of houses, United States, deviations from trend (inverted).[1]

Graph II. 4.

EXAMPLE OF LAG.



It is at once clear that there is a lag of about three years and a-half between A and B. An immediate comparison of A with B′, where any value B′ equals the value of B three years and a-half earlier, confirms the existence of this lag.

Imperfect correlation.

As a rule, the scatter will not show perfect organisation. There may, however, still be a tendency for the points to group along a curve: then *imperfect correlation* is said to exist. That curve will now no longer be exactly determined. Various choices as to its type are possible, some of which will be discussed later. Once a choice has been made, the deviations between the actual points and the curve may be measured. Here also several methods of measurement may be chosen; but, after this second choice has been made, a measure for the degree of organisation can be given.

Correlation index; correlation coefficient.

The usual measure taken in the case of a general curve is the *correlation index*, which in the case of a straight line is reduced to a simpler measure called *correlation coefficient*. Both expressions have the property of being always less than or equal to unity; and they reach unity only if there is perfect correlation between the two variates (in the case of the correlation coefficient,

---

[1] The trend of a series is a series indicating its general tendency. Details as to calculation of trend will be found in Appendix A.

if there is perfect linear correlation). Conversely, if they are equal to unity, there is perfect correlation (in the case of the correlation coefficient, perfect linear correlation).

Regression curves and coefficients.

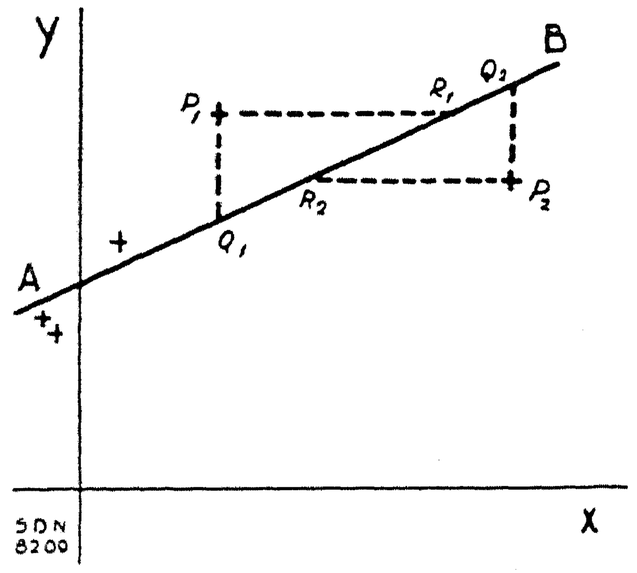
The notions of regression curve and regression coefficient, introduced above for the case of perfect correlation, are also used in cases of imperfect correlation; but they now depend on certain choices.

First, the type of curve has to be chosen. Usually a straight line is first tried. Secondly, a method of measuring deviations of the points from that line has to be devised. They may be measured in the direction of the Y-axis, in the direction of the X-axis or in other ways.

Graph II. 5 illustrates the procedure. The points representing the given observations are indicated by P_1, P_2, etc. As a regression curve, the line AB has been chosen. The deviations of P_1, P_2, etc., from AB, measured in the direction of the Y-axis, are indicated by P_1Q_1, P_2Q_2, etc. Those measured in the direction of the X-axis are indicated by P_1R_1, P_2R_2, etc.

Graph II. 5.

MEASUREMENT OF DEVIATIONS FROM A REGRESSION LINE.



The third step is to adopt some method for determining the curve in such a way that the deviations just defined will be as small as possible. Usually the "method of least squares" is taken: the sum of the squares of the deviations is made a minimum. In other words, that line is chosen as a regression line which shows the minimum sum of squares.

If deviations are measured in the Y direction—*i.e.*, in the direction of the dependent variate—the line obtained is called the *firs elementary regression line.* If deviations are measured in the X direction, the *second elementary regression line* is obtained. Each of the regression lines will be characterised by a regression slope and a regression coefficient.

In the case of perfect linear correlation, these two regression lines coincide, and no trouble arises as regards the choice indicated. When the correlation is not perfect, the difference between the two regression coefficients gives an idea of the degree of organisation of the scatter.

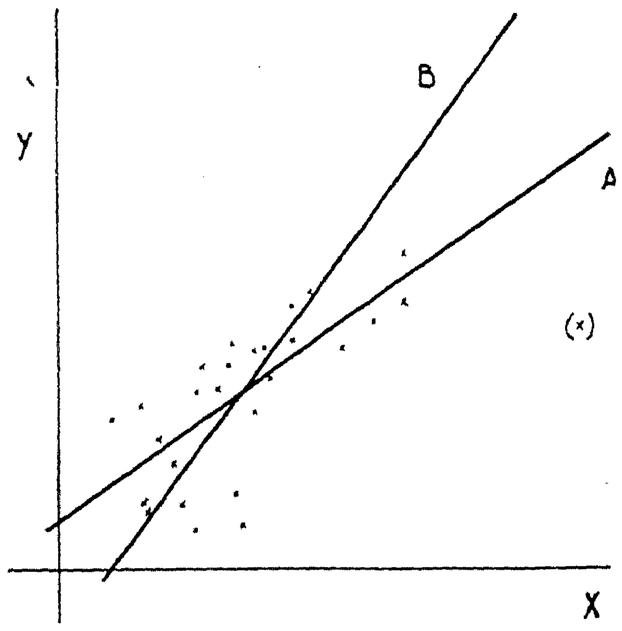
All that has been said applies as well to series x, y, etc., indicating the deviations which X, Y, etc., show from their average value $\overline{X}$, $\overline{Y}$, etc. over the period studied.

The correlation coefficient and the regression coefficients enable the two objects of the analysis to be substantially attained. The correlation coefficient tells whether or not the assumed relation between X and Y is exact, and therefore gives an answer to the verification problem. The regression coefficients indicate about how large a change in Y corresponds to a given change in X, and therefore answer the question of measurement. A first rough test of the economic significance of the coefficients is afforded by their signs, which may or may not be such as economic theory would lead one to expect.

Example. Graph II. 6 gives the scatter diagram between X, "value added" per ton of pig-iron, and Y, pig-iron production for Germany, 1881-1911. Value added per ton, which equals price minus raw-material cost, has been taken in order to eliminate the effect of the most important changes in production cost. Production has been measured in a somewhat unusual way, in order to eliminate influences of growth in productive capacity—viz., as the percentage deviation from trend. The relation is in its essence a supply relation, in which disturbing influences of cost and capacity changes have been eliminated by one of several possible methods. The scatter is moderately organised, and the only indication of curvilinearity is in the single point to the right, corresponding to the boom year 1900. Leaving aside this point, two elementary

Graph II. 6.

EXAMPLE OF SCATTER DIAGRAM.



linear regressions have been calculated and the corresponding lines drawn. The first elementary regression formula runs:

$$Y - \bar{Y} = 0.71\ (X - \bar{X})$$

in which $\bar{X}$ is the average value of all X's (except that for 1900) and $\bar{Y}$ is the average value of all Y's (except 1900). The meaning of the regression equation is that an increase of one point in X causes an increase of 0.71 points in Y. The second elementary regression would yield the figure 1.37 instead of 0.71 and shows that a rather high degree of uncertainty prevails here. The economic significance of these figures is closely connected with the elasticity of supply. In fact, it follows from the above definitions that an increase in prices by one unit, raw-material cost being supposed equal, would cause an increase in production of about 0.71% (the trend value of production being used as a basis). As the average price for the period was 59.7 Marks per ton, it may easily be deduced that the elasticity of supply was then 0.42. If the second elementary regression had been used, a figure of 0.82 would have been obtained.

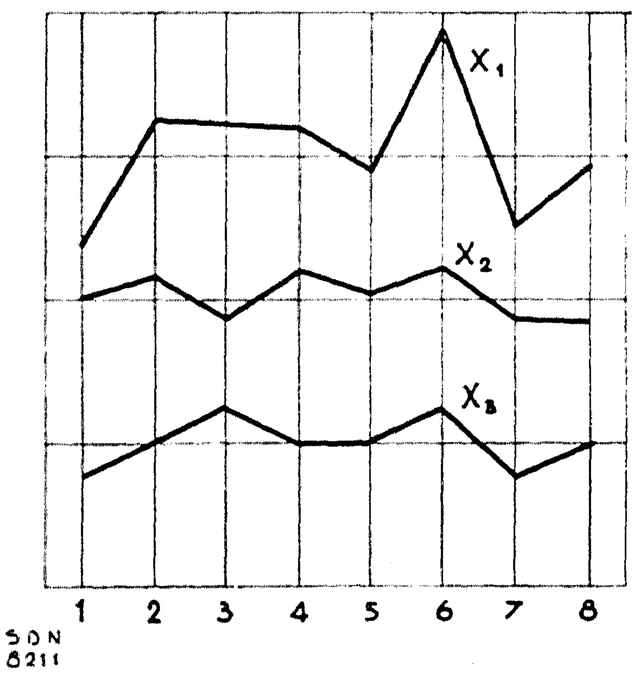
A first rough test of the economic trustworthiness of this figure is to see whether it has the right sign—*i.e.*, whether positive price changes are connected with positive changes in supply.

§ 5. Multiple Correlation

Multiple correlation.

As has already been said, in by far the greatest number of cases of economic importance, more than one cause is ordinarily assumed to have acted. Fluctuations in a series x_1 will have to be explained by the fluctuations in a number of other ("explanatory") series x_2, x_3, etc. To begin with, the nature of this problem may best be illustrated by an historical graph of all the series involved (*cf.* Graph II. 7). For

Graph II. 7.
Perfect Multiple Correlation.



the sake of simplicity, we may again start with a case of p
linear relationship.

Perfect linear multiple correlation.

The problem then is to find the figures b_2, b_3 by which the series x_2, x_3, etc., have to be mult in order that the sum $b_2x_2 + b_3x_3 + \ldots$, calculat each time-point, may equal the corresponding for x_1. In the theoretical example of Graph these numbers are 2 and 5. In fact, $2x_2 + 5x_3$ exactly x_1 for each observation. The graph shows some elemen features which are important for the carrying-out of the ana and may therefore be stated. The decline in year 5 is ca entirely by series x_2, x_3 showing no decline at all in that year. the other hand, the rise in year 8 can only be explained by x_3, does not rise in that year. These two examples clearly show only a combination of x_2 and x_3 can give the right result. Moreo it is the combination with coefficients 2 and 5 which g the best result, as is seen very clearly in year 3, where that combination will produce the absence of change in These elementary remarks are intended to demonstrate t considerations of this kind may be helpful in the study of tual relationships, since they may show, after a mere visual spection of the statistical material, whether or not success is to expected.

Regression coefficients.

As in the case of simple correlation, the coefficient and 5 in the above example are called regress coefficients. As before, the regression coeffici indicates the increase (or decrease) in x_1 caus by a unit increase of x_2 or x_3 respectively; and, as before, first rough test can be applied to this conclusion by enquiri whether the coefficient has the sign which economic theory wou lead one to expect.

Influence of x_2.

In the expression $2x_2 + 5x_3$, the term $2x_2$ (in gener b_2x_2) may be called "the influence of x_2" and 5 "the influence of x_3". In using these terms, on must, however, bear in mind that this expression justified only so far as the economic theory which has prompted th calculation is accepted as valid. The special value of such a ter

in year t may be called "the influence exerted in that year"; whereas the strength of that influence in a given period may be characterised by, *e.g.*, the standard deviation of the term—*i.e.*, $2\sigma_{x_2}$, etc. All these expressions are independent of the units in which x_2 or x_3, etc., is measured.

Partial scatter diagrams. In multiple correlation analysis, the scatter diagram may still be used, but with a somewhat different function. Plotting three or more variates in a plane is not easy; but, instead, two or more partial scatter diagrams may be considered. The first uses as co-ordinates x_2 and $x_1 - 5x_3$ (in general, $x_1 - b_3x_3$): *i.e.*, x_1 "minus the influence of x_3", or "corrected for changes in x_3". The diagram so obtained illustrates the relation between x_2 and x_1 "other things being equal" or, more exactly, "other relevant things being equal". A second diagram may be constructed comparing x_3 and $x_1 - 2x_2$ (in general, $x_1 - b_2x_2$).

Imperfect multiple correlation. The same technique [1] can be usefully employed in cases where no figures b_2, b_3, etc., can be found which make $b_2x_2 + b_3x_3 + \ldots$ exactly equal to x_1, for each time-point. This, in fact, is generally the case as long as the number n of series considered is smaller than the number N of time-points.[2] We must be satisfied if certain values for b_2, b_3, ... give a fairly good fit. As in the case of only two variates, such coefficients b_2, b_3, ... can be calculated after choosing the way in which deviations are to be measured and minimised. Again, b_2, b_3 are called regression coefficients, and the expression

$$x_1^* = b_2x_2 + b_3x_3 + \ldots$$

is called the *regression equation* of x_1 on x_2, x_3, etc.; x_1^* is often called the *calculated or theoretical value* of x_1. The differences $x_1 - x_1^*$ for each point of time are called *residuals*. If the line of best fit is chosen so as to make the sum of the squares of these residuals

---

[1] An example of this technique is found in Graphs III. 9-III. 11.

[2] If n equals N, then values b_2, b_3, etc., can always be found, as the number of unknowns b_2, b_3, etc., equals the number of relations which must be fulfilled.

as small as possible (*i.e.*, by application of the principle of least squares to the residuals), it is called the *first elementary regression*. The corresponding values for b_2, b_3, etc., will be written as

$$b_{12},\ b_{13},\ \ldots$$

The deviations might, however, have been measured in other directions—*e.g.*, in that of x_2, by trying to find an expression $x_2^* = b_{21}x_1 + b_{23}x_3 + \ldots$ which shows a minimal sum of the squares of $x_2 - x_2^*$. This is the second elementary regression. Of course, there are n such elementary regressions. In the calculations discussed in later chapters, the first elementary regression will generally be used; but information as to the other regressions will also be included.

The *total correlation coefficient* R between x_1 and x_1^* can be used as a measure of the degree of accordance between x_1 and x_1^*, and therefore, to some extent, as a measure of the success obtained.

The technique of partial scatter diagrams is again helpful to show whether or not the correlation obtained is satisfactory.

Multiple curvilinear correlation.

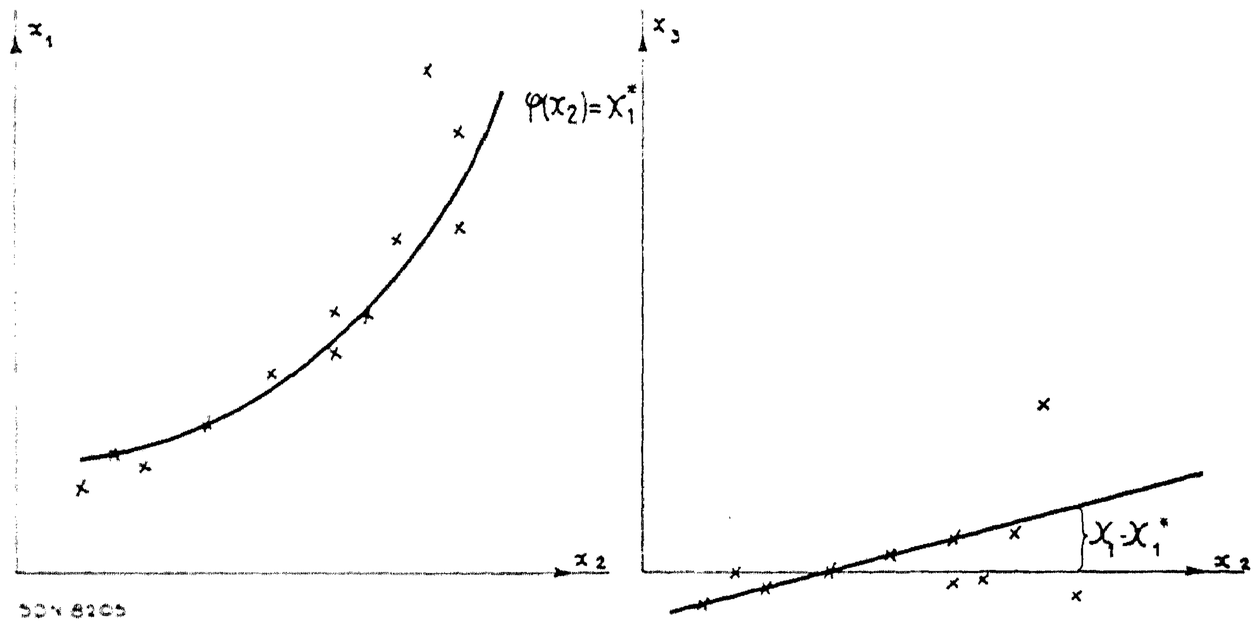
Partial scatter diagrams are especially helpful in order to test whether or not the assumption that the relation between x_1 and x_2, x_3 ... is linear, fits the facts.[1] If the partial scatters show curvilinearity, this assumption is no longer valid. Two ways are open for further attempts. First, more complicated algebraic formulæ can be tried and treated in a similar way to the linear ones; secondly, graphic methods can be used. These, however, can only start with a scatter between x_1 and one other variate (say, x_2), it being difficult to plot three or more variates in one chart. This scatter may show a tendency to a curvilinear relation, which may be drawn as a freehand curve through the cloud of dots. Let its ordinates (*cf.* Graph II. 8) be called $x_1^* = \varphi(x_2)$. Then for each point the difference between x_1 and the value x_1^* corresponding to its x_2 may be calculated, and this difference may be plotted again as x_3. If a close correlation—perhaps also curvilinear—is found, the curvilinear explanation may be more acceptable than the rectilinear one. Many alternatives are possible; to give details regarding them

[1] Graphs III. 9-III. 11 provide some examples.

and regarding the refinements of the method would, however, lead us too far.

Graph II. 8.

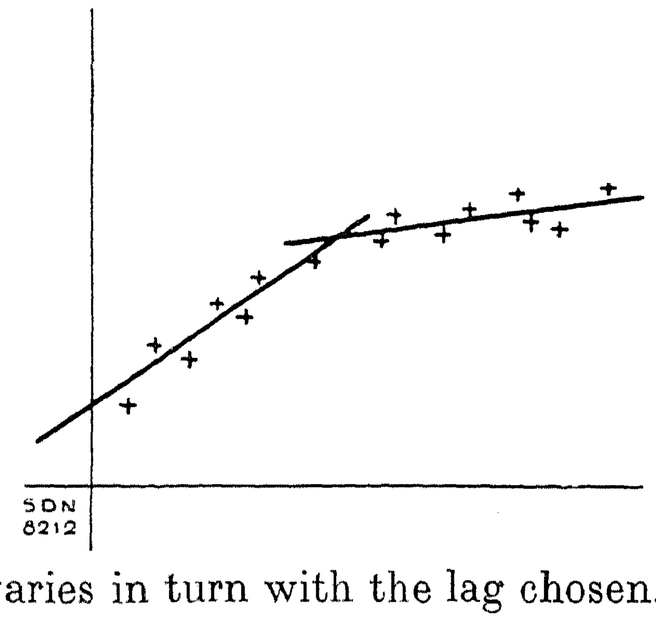
GRAPHIC ANALYSIS OF CURVILINEAR CORRELATION.



As a rule, curvilinear relations are considered in the following studies only in so far as strong evidence exists. A rough way of introducing the most important features of curvilinear relations is to use changing coefficients—for instance, one system of coefficients for the description of situations not far above normal and another for the description of extremely high levels. This amounts to approximating a curve by means of two straight lines (*cf.* Graph II. 9). Another way of introducing curvilinear relations is to take squares of variates, or still other functions, among the " explanatory series ".

Graph II. 9.

APPROXIMATION OF A CURVILINEAR RELATION BY TWO RECTILINEAR ONES.

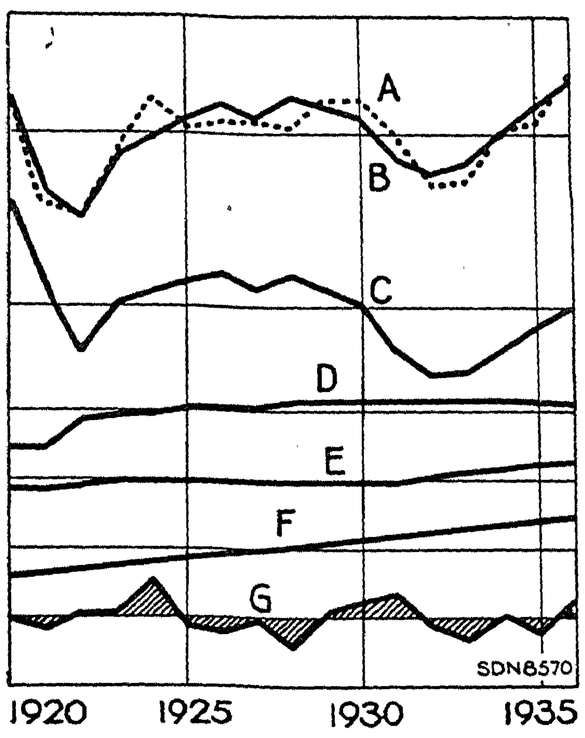


Lags may also be introduced in multiple correlation analysis. The best lag, however, can no longer be determined by mere examination of historical graphs, since it depends on the relative influence of the various explanatory factors, and this relative influence varies in turn with the lag chosen. In principle, all possible lags must be tried and the regression

coefficients calculated on each assumption. In practice, simplifications of procedure are possible if, for instance, one series only is of major influence and the others are secondary.

Graphic representation chosen.

As a rule, the results of multiple correlation calculations will be represented as in Graph II. 10. At the top, the actual series to be "explained" is indicated by dots, and on the same scale the theoretical values are indicated by a continuous line. Below the two lines, the various composing series b_2x_2, b_3x_3, etc., are drawn.



Graph II. 10.

GRAPHIC REPRESENTATION OF A CORRELATION CALCULATION. "EXPLANATION" OF IRON AND STEEL CONSUMPTION. UNITED KINGDOM 1920-1936.

A = Actual iron and steel consumption.
B = Calculated iron and steel consumption.
C = Influence of profits one year before.
D = Influence of interest rate ½-year before.
E = Influence of price of iron ½-year before.
F = Influence of time.
G = Residuals, *i.e.*, A-B.

The ordinates of these lines are proportional to—not equal to—x_2, x_3 (and even proportional only in cases of linear formulæ). They represent what have been called "the influence of x_2", "the influence of x_3", etc. The advantages of this procedure are, first, that the scale of these series is comparable with that of the first series, and secondly, that it can be seen at once which of the series are important (*a*) in general or (*b*) for the explanation of any particular feature.

Example.

For example, Graph II. 10 is the result of a calculation aiming at "explaining" the fluctuations in iron and steel consumption in the United Kingdom 1920-1936; the explanatory series are profits of all industries one year before, bond yield and iron price half-a-year before, and time. The regression equation found is

$$x_1^* = 1.17x_2 - 0.08x_3 - 0.24x_4 + 2.39x_5$$

where x_1^* represents the calculated physical volume of iron and steel consumption in percentage deviations from average;

x_2, profits, all industries, percentage deviations from average;
x_3 bond yield, deviations from average in hundredths of 1%;
x_4 iron prices, percentage deviations from average;
x_5 time, years.

Obviously, this relation may be interpreted as a demand relation for iron where the series x_2, x_3 and x_5 have been taken as the other chief causes for changes in demand, and where a lag of one year for x_2 and of half-a-year for x_3 and x_4 has been assumed to exist. The movements of the three series x_2, x_3 and x_5 are responsible for the shifts in the demand curve during the period under review. The partial scatter diagram between $x_1 - 1.17x_2 + 0.08x_3 - 2.39x_5$ and x_4 would give the usual representation of the demand curve, shifts having been eliminated. As both x_1 and x_4 have been measured in percentage deviations from average, it will readily be seen that the elasticity of demand for iron would amount to -0.24 for prices and quantities near to their average values. Economically, the negative signs of the coefficients of x_3 and x_4 are as they should be. In addition, it may be mentioned that the " influence " of x_3 and x_4 is only small.

§ 6. Statistical Significance of Results

Reliability of results.

The reliability of results may be judged by statistical as well as economic criteria. In general, the figures used are not exact. They are often derived from samples, or otherwise more or less inadequate for the problem under consideration. In addition, a number of minor explanatory causes are omitted; this seems to be the chief reason why observed and calculated values of x_1 in general do not coincide, and this lack of coincidence is responsible for a certain ambiguity in the results obtained. The question arises whether limits may be indicated for this uncertainty. As nothing is known about the factors omitted, it can be answered only if certain additional hypotheses are made.

Various methods of statistical testing have been worked out, using different hypotheses and leading, therefore, to different results. Some account of these methods will now be given. The non-mathematical reader should be warned that their comprehension will make somewhat greater demands on his attention than has the foregoing exposition of the method of multiple correlation analysis itself; and he may perhaps prefer to take the remainder of this chapter, together with Appendix A, on trust.

The classical method.

The classical method goes back to LAPLACE and GAUSS. It will be considered here in the final form that has been given to it by Professor R. A. FISHER.[1] According to this method, it is assumed that the unexplained parts—the residuals—are due to the circumstance that the "explained" variate, though essentially a linear function of the "explanatory" variates, contains an additional component representing the influence of neglected explanatory variates and may, moreover, be subject to errors of measurement. This so-called "*erratic component*" or "*disturbance*" in the explained variate not only gives rise to unexplained residuals, but also causes the regression coefficients calculated from the observations to differ from the coefficients of the true relation connecting the variates. The probable average magnitudes of these differences are derived from the assumption that the disturbances in subsequent time intervals are to be considered as "random drawings" from the "universe" of all possible values of these disturbances. In that "universe" there will be larger and smaller values of these disturbances, and these values are assumed to be normally distributed. This normal distribution means that the number of cases present in each class of magnitude will be determined by the so-called Gaussian law. In ordinary speech, small disturbances will be numerous and large disturbances will be few, their frequency obeying a simple law. The square root of the mean value of the squares of these disturbances is called their *standard deviation*, and is denoted by σ.

On certain further assumptions of a rather technical nature, it becomes possible to calculate what results with respect to the regression coefficients would have been obtained if another sample of disturbances had—by accident, so to say—been drawn. By comparing all possible results, one may say within what limits the results of the great majority of the possible cases will lie. These limits depend again on the choice one makes as to the "majority". Often 99% or 95% is taken. If b_2 is one of the regression coefficients calculated, and σ_{b_2} the so-called standard error of b_2, about 95% of the cases lie between

[1] Cf. *Statistical Methods for Research Workers*, London and Edinburgh, 1936; "The goodness of fit of regression formulæ and the distribution of regression coefficients", *Journ. Roy. Stat. Soc.*, **85**, 1922, p. 597; applications of "Student's" distribution, Metron, **5**, 3, 1926, p. 3.

$b_2 + 2\sigma_{b_2}$ and $b_2 - 2\sigma_{b_2}$, *i.e.*, in a range of width $4\sigma_{b_2}$ around b_2. About 99.7% lies between $b_2 \pm 3\sigma_{b_2}$.

This standard error σ_{b_2} is nothing else than the " standard deviation " of the differences between the calculated and the true regression coefficient in repeated samples. It depends—and with it the range of uncertainty in the calculated regression coefficients—on the following figures:

(1) The number (N) of observations containing mutually independent disturbances. The larger this number, the smaller σ_{b_2}. In economic problems, however, it is not always certain how large should be the time interval to which one observation refers in order to make successive values of the disturbances virtually independent.

(2) The number $(n - 1)$ of explanatory series. The larger this number, the larger σ_{b_2}. This will be understood if it is realised that, by $n = N$ (*i.e.*, if the number of explanatory series is one less than the number of observations), a perfect correlation can be obtained by any set of mutually independent explanatory series, even if they do not bear at all on the subject.

(3) The total correlation coefficient (R). The nearer to 1 this number is, the smaller is σ_b; for $R = 1$, σ_b becomes zero, except when there is perfect correlation between one of the explaining variates and a group of other explanatory variates.

(4) The correlations between two or more of the explanatory series. If at least one of these correlations is high, some of the regression coefficients show a larger σ_b (*i.e.*, are very uncertain). This, too, is easy to understand. In fact, in the extreme case, where two explanatory series were exactly parallel, it is clear that a substitution of one of them for the other would not change the correlation. The " best " fit could therefore be obtained with each of an infinite number of different combinations, in which one series would successively be substituted to a larger and larger extent for the other. The two regression coefficients of these two series would be entirely indeterminate; only some combination of them would be determinate.

Now even if the correlation between two explanatory series is not exact, small disturbances—which are always present—can change the result considerably, and therefore the various possible " samples " would show considerable differences. Hence σ_b will be large. The exact expression for σ_b and its computation are given in Appendix A, § 4.

Frisch's method.

Professor R. Frisch,[1] in his treatment of these problems, does not use the concept of some unknown " universe " from which a " sample " is drawn. He considers *every* variate as being built up of a systematic part and a disturbance. The relations assumed between the variates are supposed to hold good exactly between the systematic parts, and the regression coefficients in these relations

[1] Cf. *Statistical Confluence Analysis by Means of Complete Regression Systems.* Universitetets Økonomiske Institutt, Publ. Nr. 5, Oslo, 1934.

are called the true coefficients. The calculated coefficients may again show deviations from the true, and the object is to find these deviations or a limit to them.

On the further assumptions that there is no correlation (*i.e.*, that the correlation coefficient is zero) between: (i) the disturbances of different variates; (ii) the disturbances of one and the systematic part of another variate; and (iii) the disturbances and the systematic part of the same variate; it may be shown that, at least for problems of two variates the *true regression lies between the elementary regressions.*

Bunch-map analysis.

This is why Professor Frisch proposes to construct what he calls *bunch maps.* These indicate the regression slopes obtained for one pair of variates, if all possible elementary regression equations are solved. For a technical reason all variates are normalised—*i.e.*, expressed in their own standard deviation as units.

In order to explain the principle, a three-variate problem may be considered, where an endeavour is made to " explain " x_1 by x_2 and x_3. The first elementary regression equation provides an " explanation "

$$x_1' = b_{12\cdot3}\, x_2 + b_{13\cdot2}\, x_3 \qquad (1)$$

with a regression coefficient $b_{12\cdot3}$ for x_2 and $b_{13\cdot2}$ for x_3. Taking the second elementary regression, we obtain an " explanation " of x_2

$$x_2' = b_{21\cdot3}\, x_1 + b_{23\cdot1}\, x_3,$$

which may, however, be transformed into an " explanation " of x_1 by putting $x_2' = x_2$ and solving for x_1:

$$x_1'' = \frac{1}{b_{21\cdot3}}\, x_2 - \frac{b_{23\cdot1}}{b_{21\cdot3}}\, x_3 \qquad (2).$$

The two dashes have been added to indicate the second elementary regression as the origin of this estimate. Similarly, the third elementary regression

$$x_3' = b_{31\cdot2}\, x_1 + b_{32\cdot1}\, x_2$$

gives

$$x_1''' = -\frac{b_{32\cdot1}}{b_{31\cdot2}}\, x_2 + \frac{1}{b_{31\cdot2}}\, x_3 \qquad (3).$$

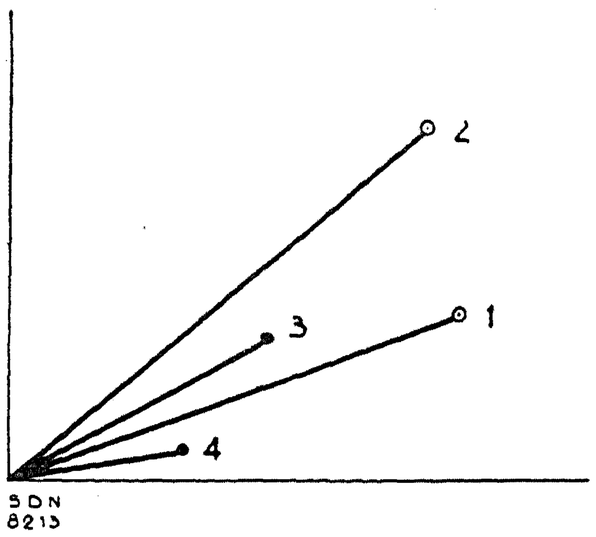
The equations (1), (2) and (3) are three estimates of the relation between the variates; two bunch maps are constructed to illustrate them. The first compares the three coefficients (in graphical representation, the slopes) obtained for the influence of x_2, viz. $b_{12\cdot3}$ from (1), $\frac{1}{b_{21\cdot3}}$ from (2) and $-\frac{b_{32\cdot1}}{b_{31\cdot2}}$ from (3). They are represented by three beams, numbered 1, 2 and 3 (being the numbers of the variates in whose direction the minimising has been performed). The beams 1 and 2 will be marked ⊙, indicating that the slopes are those between 1 and 2, 1 (the lower numbered variate) being considered as the variate to be " explained ". The second bunch map compares the three coefficients obtained for the influence of x_3 upon x_1 viz., $b_{13\cdot2}$, $-\frac{b_{23\cdot1}}{b_{21\cdot3}}$

and $\frac{1}{b_{31 \cdot 2}}$. The beams are again numbered 1, 2 and 3, but here 1 and 3 are marked ⊙.

Similar bunch maps are made for all conceivable combinations of variates, starting with the simplest and ending with the " complete set " including all variates. The bunch maps for a two-set are of course extremely simple: they always consist of two beams only, which, by the choice of units referred to above, are necessarily situated symmetrically with respect to the two axes.

In general, each bunch map consists of a number of beams, two of which —the " leading beams "—have their ends marked ⊙ (*cf.* Graph II. 11). The numbers at the ends of these two beams indicate the variates, the regression between which is being studied. In the cases considered in Chapters III, IV and V, the variate with the lower number will always be taken as the one " to be explained ". Every other beam bears a number, and all the numbers together represent the group of variates used. The number attached to any beam represents the variate used as the left-hand variate in the regression equation before transformation. In other words, it indicates the direction in which deviations have been minimised in constructing the regression formula studied.[1]

Graph II. 11.
SPECIMEN OF BUNCH MAP.



A case of perfect relationship without any ambiguity is provided by bunch maps where all beams coincide; for it cannot make any difference in what direction we decide to measure deviations, if there are no deviations to be measured. But if in any case one of the explaining variates has been omitted, perfect correlation cannot exist, and no perfectly closed bunches appear.

Useful, superfluous and detrimental variates.

If, therefore, in a given case the bunch is not closed, the aim of further research, and in particular of including further variates in the analysis, is to close the bunch. Any economically significant variate which helps to close the bunch, or brings about a distinct change in the various slopes in the bunch without making it less closed, is called a *useful variate*. Any new variate which only slightly changes the bunch is called *superfluous*. There is, however, a third possibility: the bunch may " explode "—*i.e.*, show a larger spread after a new variate has been introduced. This happens if there is a high correlation between the new series and one or several of the previous explanatory factors. We are then faced with a situation which is called " *multicollinearity* ". It has been shown that,

[1] An explanation of differences in the length of the beams would lead us into too much technical detail.

in such a case, some of the regression coefficients become very uncertain; it is therefore possible that quite different results will be obtained if the deviations are measured in different directions. The new variate, or one of the older variates, is then called *detrimental.* This should be interpreted as meaning that, if all variates are included at the same time, no trustworthy measurement can be made. This does not mean that the variate in question may not be economically significant, but only that, owing to some circumstance (fortuitous or systematic), complete measurement is impossible. A less ambitious measurement may still be possible. Because of the great importance which attaches to these cases, the following simple example may be given.

Suppose one tries to determine the demand function for butter; x_1, the quantity of butter sold, has to be explained by

x_2 the price of butter,
x_3 the price of margarine,
x_4 the income of consumers.

Now we find that butter and margarine prices (at least their annual averages) are fairly highly correlated. Hence in the proposed demand equation

$$x_1' = b_2 x_2 + b_3 x_3 + b_4 x_4, \qquad (1)$$

while b_4 may perhaps be readily determined, it will be impossible to find b_2 and b_3 separately with sufficient accuracy. One expedient, however, may be adopted. If x_3 is left out, the equation

$$x_1'' = b_2' x_2 + b_4' x_4 \qquad (2)$$

may be tried; it will be possible to determine b_2' and b_4' provided the correlation coefficient found is not too bad (b_4' will be approximately equal to b_4). Equation (2) may be used instead of (1) if it is kept in mind that x_2 now stands for the combined influence of x_2 and x_3; it will be found that approximately $b_2' x_2 = b_2 x_2 + b_3 x_3$. This holds good only as long as the correlation between x_2 and x_3 persists. Equation (2) may therefore be used in all problems in which this correlation does not fail. For example, if the price of butter is raised by State regulation, but the price of margarine is raised as well so as to maintain the correlation between the two prices, then the consequences of the policy on the amount sold may be calculated. If, on the contrary, the regulation does not maintain the correlation, the formula becomes useless for this purpose.

A combination of the two methods.

Dr. T. Koopmans [1] has pointed out that the classical method and that of Frisch are complementary rather than alternative. Each of them deals with a part of the margin of uncertainty which must be assigned to calculated regression coefficients. That part of this margin which constitutes the object of Fisher's argument could be called the *error of sampling.* According to Fisher's hypotheses, it is due to the fact that the disturbances in the explained variate may affect the calculated regression coefficients to an unpredictable

[1] *Linear Regression Analysis of Economic Time Series*, Haarlem, 1936.

amount, which can be dealt with only by means of laws of probability. Since FISHER does not assume disturbances in the explanatory series, he thus rules out the type of uncertainty studied by Frisch. For this additional uncertainty arises from the circumstance that we usually do not know to what extent the disturbances found to be present in the whole set of data must be ascribed to this or that variate entering into the relation; or, in more technical terms, since we do not know exactly, in calculating the regression coefficients, what relative weights should be applied to express the relative accuracy of each of the several statistical series representing the variates, we incur, by any choice of weights whatever, the risk of introducing an *error of weighting* in the calculated coefficients. On the other hand, the error of sampling is excluded from Frisch's argument by his somewhat restrictive assumptions which have been indicated above.

Koopmans therefore combines the two theories into one method which deals simultaneously with the error of sampling and the error of weighting in the calculated coefficients. His procedure is as follows: For any set of relative weights of the variates that we may choose—*i.e.*, for any numerical guess we may make about the *relative* strength of the disturbances in the several variates,— mathematical deductions lead to:

(1) A set of " best estimates " for the regression coefficients, which takes the place of the first elementary regression in the classical method;

(2) A set of " standard errors " indicating the degree to which each of these estimates may be subject to errors of sampling; these standard errors correspond to those of the classical theory;

(3) A set of estimates of the standard deviations of the disturbances in each of the statistical series employed, which estimates measure the *absolute* strengths of these disturbances.

Where normally the correct relative weights are unknown, it appears that, under certain conditions including mutual independence of disturbances in different variates, the estimates of the regression coefficients mentioned under (1) remain within certain limits for all *a priori* possible weights. These limits correspond to those found by Frisch for the case of two variates, and are given by the two ultimate beams (not always the two " leading " beams) in the bunch map for the corresponding coefficient in the complete set of variates. They constitute ultimate limits to the error of weighting.

In a number of cases, however, narrower limits can be established with the help of the estimates mentioned under (3). It is often very improbable that the disturbances in any variate are of a size comparable to that of the variate itself. If such a result were arrived at from any presumed set of relative weights, such weights could be discarded as being inacceptable. Thus, frequently, the elementary regressions corresponding to variates that exercise only a secondary influence on the explained variate are excluded by this rule. Interpreting this proposition in terms of the bunch-map analysis, it might be said that, in these cases, the beams corresponding to such series should be disregarded, or at least be assigned less importance than the others, even if they are " leading " beams.

(1) In the early phases of *statistical business cycle research*, attention was paid to somewhat superficial phenomena, such as the length of cycles, the degree of simple correlation between series and the relative amplitudes of their movements, the decomposition of series into trend, seasonal components, etc. Certainly all this work had its value, especially for the *negative* evidence it afforded on the validity of certain theories. For the purpose of applying more searching tests, however, it is necessary to dig deeper. An apparently simple relation, such as that between prices and production, is often not a direct causal relation at all, but a more or less complicated chain of many such relations. It is the object of analysis to identify and to test these direct causal relations: production, for instance, may be regarded as determined by the volume of orders; the volume of orders by the income of consumers and by prices; income by employment, wage rates and so on.

The part which the statistician can play in this process of analysis must not be misunderstood. The theories which he submits to examination are handed over to him by the economist, and with the economist the responsibility for them must remain; for no statistical test can prove a theory to be correct. It can, indeed, prove that theory to be incorrect, or at least incomplete, by showing that it does not cover a particular set of facts: but, even if one theory appears to be in accordance with the facts, it is still possible that there is another theory, also in accordance with the facts, which is the " true " one, as may be shown by new facts or further theoretical investigations. Thus the sense in which the statistician can provide " verification " of a theory is a limited one.

On the other hand, the rôle of the statistician is not confined to " verification ". As the above example illustrates, the direct causal relations of which we are in search are generally relations, not between two series only—one cause and one effect—but between one dependent series and several causes. And what we want to discover is, not merely what causes are operative, but also *with what strength each of them operates*: otherwise it is impossible to find out the nature of the combined effect of causes working in opposite directions. On this problem—the problem of " measurement ", as it may be called—the statistician can