The Method of Path Coefficients and OLS Regression

R. C. GEARY

SOME time ago Sir Maurice Kendall suggested to the writer that a study of the theory and practice of path coefficients might be rewarding as shedding light on the still dark patches in the theory of relationship between random variables.

For the writer this is a personal matter. In holding (after a reasonably close train of argument—see Geary, 1963) that, in multivariate OLS regression, the individual coefficients, in general, are objectively meaningless, he is probably still in a minority amongst statisticians, though he is unaware of any systematic refutation of his position.

Attention is confined here to parts of Sewall Wright's (1934) seminal paper. The writer is aware (mainly through Dr D. E. Chambers) that there is a fairly large subsequent literature and that recently the path coefficient method has been used in social research. He has not read these papers; accordingly he is less concerned to claim novelty for any results in this paper than to find out if the path coefficients approach leads to a modification of the somewhat negative personal view expressed in the last paragraph. We shall find that in certain conditions it does. There are also some comments on Wright's analysis.

Path Coefficients

Wright's paper, over 40 years old, is greatly to be admired for its comprehensiveness and thoroughness. For those not familiar with the subject a brief summary may be desirable, with special reference to two of his telling applications. In the following exposé, notation different from Wright's is used.

Write

$$y = b_1x_1+b_2x_2+ \ldots +b_kx_k, \qquad (1)$$

each of the $k+1$ variables being standardised from n sets,[1] i.e., with

$$\bar{y} = 0, \Sigma y^2 = n$$

$$\bar{x}_i = 0, \Sigma x_i^2 = n, \quad i = 1, 2, \ldots, k$$

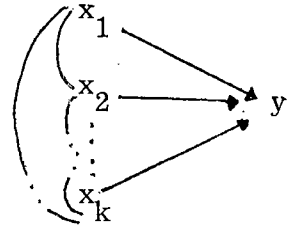
There is no disturbance term but it is clear that Wright had OLS regression in mind. He contemplates additional variables "u", so that it is possible that the ultimate representation is deemed to be *exactly*

$$y = \sum_{i=1}^{k} b_i x_i + \sum_{j=k+1}^{K} b_j x_j, \tag{2}$$

with only k (out of a possible K) variables known, so that nowadays we would write "y_c" for the "y" on the left of (1). The b_i are, by definition, the *path coefficients*. Because of standardisation, when $k = 1$, $b_i = r_1$, the coefficient of correlation (cc) between y and x_1. In general the path coefficients are functions of the ccs of the system, through the standard OLS equations for determining the b_i.

Essential in Wright's theory is also the notation of causal sequence, represented diagrammatically. Thus Fig. 1 with its single-headed arrows and connecting lines[2] is a causal representation of formula (1).

FIG 1



With the single-headed arrows the head indicates the effect, i.e., the depvar y, the other end the cause, the indvars x_i. The other lines indicate that the variables are possibly correlated, but without specification of direction of causation. Incidentally, throughout this paper we use the same algebraic notation, e.g., y and the x_i, for both description of variables and their measure.

Two Examples[3]

Wright's "simplest application" was in connection with the factors which determine the average weight of guinea-pigs at birth. Very full and clear data are given resulting from thousands of experiments.

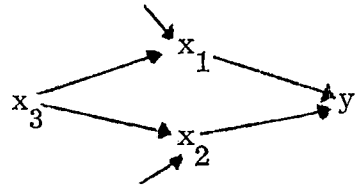
1. For simplicity of notation, throughout we omit cursive set subscript, say t, ranging from 1 to n. Unless otherwise indicated. Σ is summation according to t.
2. Wright uses double-headed arrows. The writer thinks this basically illogical and possibly misleading, as suggesting causation where none is intended.
3. All figures quoted in the examples are Wright's except as otherwise indicated.

We reduce the report to bare essentials. Let

$$y = \text{Average weight at birth}$$
$$x_1 = \text{Pre-natal rate of growth}$$
$$x_2 = \text{Length of gestation period}$$
$$x_3 = \text{Size of litter}$$

Three ccs are given r_2 for (y, x_2), r_3 for (y, x_3) and r_{23} for (x_2, x_3). Causation sequences are shown in Fig. 2. In fact $r_2 = +0{\cdot}56$, $r_3 = -0{\cdot}66$, $r_{23} = -0{\cdot}48$.

FIG 2



Note that (1), x_1 is not precisely defined, (2), that x_1 and x_2 can have causative factors[4] other than x_3, (3), y is completely determined by x_1 and x_2. Functions pertaining to x_1 are determined from the following equations

$$\text{(i)} \qquad y = b_1x_1+b_2x_2$$
$$\text{(ii)} \qquad x_1 = r_{13}x_3+z_1 \qquad (3)$$
$$\text{(iii)} \qquad x_2 = r_{23}x_3+z_2,$$

Disturbances z_1 and z_2 assumed to be uncorrelated with x_3 and with one another.

In succession, mean square of (3)(i) and mean products of (3)(i)$\times x_2$ and (3)(i)$\times x_3$ are set down, as follows

$$1 = b_1^2+b_2^2+2b_1b_2r_{12}$$
$$\text{(i)} \qquad = b_1^2+b_2^2+2b_1b_2r_{13}r_{23} \text{ (using (3) (ii) and (3) (iii))} \qquad (4)$$
$$\text{(ii)} \qquad 0{\cdot}56 = b_1r_{13}r_{23}+b_2$$
$$\text{(iii)} \qquad -0{\cdot}66 = b_1r_{13}+b_2r_{23}$$

(4) consists of three equations to determine three unknowns, the path coefficients b_1 and b_2 and cc r_{13}, the only other quantity involved, namely, r_{23}, $(= -0{\cdot}48)$ being given. Though the equations (4) are non-linear an unique solution[5] is easily derivable—

$$b_1 = 0{\cdot}87,\ b_2 = 0{\cdot}30,\ r_{13} = -0{\cdot}59, \qquad (5)$$

4. i.e. Indicated by arrows with provenance undefined.

5. This solution is Wright's. The Referee kindly points out that the solution of (4) should be $b_1 = {\cdot}86$, $b_2 = {\cdot}32$, hence slightly different.

to which we add $r_{23} = -0{\cdot}48$. Then, from (4) (iii), the cc between average weight at birth and size of litter, namely, $r_3 = -0{\cdot}66$ breaks into two parts (on the right).

$$b_1 r_{13} = -0.51 \text{ and } b_2 r_{23} = -0.15.$$

So far the argument is unexceptionable, indeed it has fascinating aspects. Characteristic of the method is the fact that (as we shall also see in the second example) variables, objectively undefinable, can be introduced into the calculation and their statistical functions calculated. This is the character of x_1 in equation (3) (i). Sewall Wright calls it "rate of growth". This is quite unnecessary: rate of growth, one would think, is Y/X_2 (Y and X_2 being the absolute values of y and x_2) but Wright carefully refrains from such definition. In fact x_1 is simply a standardised variable introduced to make (3)(i) an identity, and thus enabling the derivation of the crucial (4)(i). This is the true character of the variable x_1. It has nothing necessarily to do with "rate of growth", unless by definition. Nevertheless, from the previous figures $-0{\cdot}51$ and $-0{\cdot}15$, Wright states

> The result is an analysis of the correlation between birth weight and size of litter into two components whose magnitudes indicate that size of litter has more than three times as much linear effect on birth weight through the mediation of its effect on growth as through its effect on the length of the gestation period . . .

The wording is as cautious as the method is ingenious, but one suspects that Wright may have had qualms about the introduction of x_1, for he goes on to set up the standard OLS regression equations of estimation of coefficients c_2 and c_3 of y on x_2 and x_3

$$\text{(i)} \qquad r_2 = 0{\cdot}56 = c_3 + c_2 r_{23} \tag{6}$$

$$\text{(ii)} \qquad r_3 = -0{\cdot}66 = c_3 r_{23} = c_2$$

which he describes as "mathematically identical" with the earlier analysis. He finds $c_3 = -0{\cdot}51$ as before and states

> The term [$c_3 = -0{\cdot}51$] can be interpreted as measuring the influence of size of litter on birth weight in all other ways than through the gestation period.

Again the wording is careful and the truth of the assertion remains to be seen. (It is true: see later.)

The second example pertains to Sewall Wright's treatment of supply-demand (in which he acknowledges the collaboration of P. G. Wright) applied to the corn-hog problem. With X and Y representing year-to-year percentage changes in quantity and price respectively and again assuming linearity

$$X_d = \eta Y + D \tag{7}$$

$$X_s = \varepsilon Y + S$$

D and S representing demand and supply factors, not otherwise defined, η and ε are the demand and supply price elasticities. At transaction level $X_d = X_s = X$ and on standardisation and solution

$$x = b_{11}d + b_{12}s \tag{8}$$

$$y = b_{21}d + b_{22}s$$

Now $\varepsilon = b_{11}/b_{21}$ and $\eta = b_{12}/b_{22}$. On mean-squaring and mean-producting from (8)

(i) $$1 = b_{11}^2 + b_{12}^2 + 2b_{11}b_{12}r_{sd}$$

(ii) $$1 = b_{21}^2 + b_{22}^2 + 2b_{21}b_{22}r_{sd} \tag{9}$$

(iii) $$r_{xy} = b_{11}b_{21} + b_{12}b_{22} + (b_{11}b_{22} + b_{12}b_{21})r_{sd}$$

Causal relations are indicated on Fig. 3

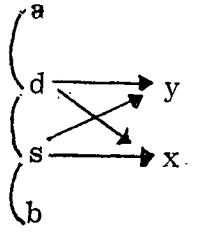


FIG. 3: *Supply-demand relations.*

(9) consists of three equations in five unknowns, i.e., the four path coefficients b and r_{sd}. For determination, two additional relations are necessary, pertaining respectively to demand and supply. These are indicated by a and b on the diagram, these being respectively assumed of the supply and demand situations. With a and b also deemed standardised, from (8) we easily derive four additional equations

(i) $$r_{ax} = b_{11}r_{ad}$$

(ii) $$r_{ay} = b_{21}r_{ad}$$

(iii) $$r_{bx} = b_{12}r_{bs} \tag{10}$$

(iv) $$r_{by} = b_{22}r_{bs}$$

since, by hypothesis, r_{as} and r_{bd} are zero. There are now seven equations to determine seven unknowns, namely, the four path coefficients and the three ccs r_{sd}, r_{ad} and r_{bs}. In theory the system is solvable, though (in the writer's view) one cannot be sure if the solution is unique in view of the non-linearity of the equations.

In the corn-hog application the possibly serious assumption is made that $r_{s\,d} = o$. As a factor of type b

> the most important single factor affecting the summer hog pack was shown to be the corn crop the preceding year. It is assumed that it is a factor [of type b] correlated with the supply situation . . . but not with the demand for pork . . .

The equation system and solution then are

Equations	*Solution*	
$1 = b_{11}^2 + b_{12}^2$	$b_{11} = 0{\cdot}132$	
$1 = b_{21}^2 + b_{22}^2$	$b_{12} = 0{\cdot}991$	
$-0{\cdot}63 = b_{11}b_{21} + b_{12}b_{22}$	$b_{21} = 0{\cdot}686$	(11)
$-0{\cdot}47 = b_{12}r_{bs}$	$b_{22} = -0{\cdot}728$	
$0{\cdot}64 = b_{22}r_{bs}$	$r_{bs} = 0{\cdot}646$	

Values of the price elasticities are $\varepsilon = b_{11}/b_{21} = 0{\cdot}192$ for supply and $\eta = b_{12}/b_{22} = -1{\cdot}361$ for demand.[6]

It should be pointed out, in regard to this second example, that, using modern terminology, the symbols d and s are "unidentified". As symbols they could have been reversed and then the demand price elasticity, in the corn-hog application, would have been found to be small and positive, the supply price elasticity large and negative; identification transpires only in economic interpretation, not within the statistical theory developed.

Again we see illustrated, in s and d, the possibility of deriving functions (coefficients and ccs) involving these, without defining them objectively.

Summary as to the Path Coefficient Method

The foregoing does not purport to be an adequate account of Wright's remarkable paper. For instance, only the two simplest of many applications have been mentioned and these have been briefly treated. Our object has been merely to reveal the bare statistical essentials of the method relevant to our main purpose.

Wright's approach is non-stochastic, except in the very minor degree that there is mention of asymptotic estimates of standard errors of means, ccs, etc. All that is involved is substitution and summing with exact linear equations (though Wright treats briefly of non-linearity). The approach to the study of relationship is essentially through ccs, while modern practice almost entirely favours single or simultaneous equation models with disturbance elements, treated as random variables, hence stochastic. As we shall see, there is less difference between the two approaches than might at first appear.

6. These values calculated from Wright's formulae differ considerably from those given by Wright, namely, $\varepsilon = 0{\cdot}133$ and $\eta = -0{\cdot}944$, for reason unknown.

Its outstanding characteristic from our point of view is that it implies an examination of relationships between the indvars. It involves a causal ordering of all the variables starting with x_k and culminating in the depvar y. It can be represented by something like

$$x_k \rightarrow x_{k-1} \rightarrow \ldots \rightarrow x_{i+1} \rightarrow x_i \rightarrow \ldots \rightarrow x_1 \rightarrow y,$$

meaning that x_i can be caused only by variables $x_j, j > i$ and all x_i may be causes of y. A sufficient condition for such a causal chain is that the variables can be so ordered in time. The whole system is thus *recursive* or, as it is sometimes called, a "Wold causal chain". There may be more than one such chain, as in the Wright-Wright demand-supply example above. In fact, examination may reveal a great variety of relationships between *all* the variables, not necessarily recursive, raising all the problems of simultaneity, identification and the rest.

In the study of relationships between stochastic variables in a particular case the most onerous part is the derivation of data—the computer will do most of the rest. Having gone to this trouble one should surely make full use of what one has.

The Nature of Linear Relationship between Variables

The OLS estimate of the coefficient b in the simple regression

$$y = bx + v = y_c + v \tag{12}$$

x and y standardised, v the disturbance, n pairs of (x, y) is found from

$$\Sigma vx = o, \tag{13}$$

with $(y - bx)$ substituted for v in (13) yielding, of course, $b = r_{yx}$. (13) can be written $r_{vx} = o$. We regard the form (13) as more "telling" than the more usual form of standard equation. It says that if x is to be regarded as the cause of y what remains after taking out bx should be unrelated to (literally uncorrelated with) bx. There is no point in the OLS operation at all unless y and x are related to start with. It is therefore natural that we should "purge" the y series until what remains is unrelated to what we have taken out. $b = r_{yx}$ is a path coefficient.

We can even find a path coefficient c for v, supposing (12) written

$$y = bx + cv \tag{14}$$

with v now deemed standardised. The standard equations for b and c are

$$\begin{aligned} r_{yx} &= b + c\Sigma vx/n \\ r_{yv} &= b\Sigma vx/n + c \end{aligned} \tag{15}$$

which, from (13), reduce to $b = r_{yx}$ (as before) and $c = r_{yv}$. Sewall Wright's omission of a disturbance term in (1) is therefore less serious than might at first appear.

In the multivariate OLS regression case, the argument is nearly identical. With

$$y = \sum_{i=1}^{k} b_i x_i + v = y_c + v \tag{15a}$$

the standard equations for estimating the b_i may be written

$$\Sigma v x_i = o, \quad i = 1, 2, \ldots, k. \tag{16}$$

If disturbance v also be standardised and endowed with a coefficient c, clearly $c = r_{yv}$, as before.

So far, therefore, there is no difference between path coefficient and OLS theory.

Contribution of Individual Causes to Total Variability

There is, however, a fundamental difference between the disturbance v regarded as a variable, and the other variables. The other variables (x, y in the simple case) are data known in advance, the v are known only in a formal way, *ex post*. The v summarises all we don't know about the system and is treated as a stochastic variable. In OLS regression the only functions we can usefully calculate about it are its variance and functions like the Durbin-Watson d or τ (tau) (see Geary, 1970) for adjudging the *completeness* of y_c as estimates of data y, by the test for residual non-autoregression.

From (15a) using (16),

$$1 = \tfrac{1}{n}\Sigma y_c^2 + \tfrac{1}{n}\Sigma v^2, \tag{17}$$

so that $\Sigma y_c^2/n = R^2$ is the principal measure of the extent to which the k indvars represent the y. In the case of simple OLS regression $R^2 = r^2_{yx}$.

As already stated, the writer holds that individual coefficients in multivariate OLS regression are meaningless (except in the trivial case of all indvars being uncorrelated). It is the whole vector of coefficients that matters, mainly for forecasting, or any rate the estimation of y_c, given indvar values. A corollary to this view would be that, with only the OLS regression available, it is not, in general, possible to estimate the contribution of *individual* variables to the total variance of y. It is possible only to assess the *total* effect, namely, $\Sigma y_c^2/n = R^2$. It may be otherwise if we have valid causative relations, i.e., OLS regressions, between the indvars.

Let us see what would happen in the simplest case of two indvars. Our treatment will be seen to be very similar to that of path coefficients, but with the introduction of disturbance terms v and w

$$\begin{aligned} &\text{(i)} && y = b_1 x_1 + b_2 x_2 + v \\ &\text{(ii)} && x_1 = r_{12} x_2 + w \end{aligned} \tag{18}$$

The Sewall Wright diagram would be as Fig. 4.

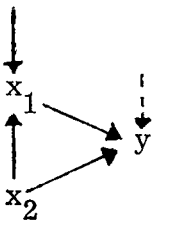


FIG. 4

Substituting for (ii) in (i) of (18)

$$y = (b_1 r_{12} + b_2)x_2 + (v + b_1 w) \tag{19}$$

Since (i) and (ii) of (18) are both OLS regressions

$$\Sigma x_1 v = o, \ \Sigma x_2 v = c, \ \Sigma x_2 w = o \tag{20}$$

Hence $\Sigma x_2(v+b_1 w) = o$ so that (19) is also an OLS regression. Hence the contribution of x_2 to the variance (which is unity) of y is $(b_1 r_{12}+b_2)^2 = r_2^2$. But from 18 (i) the contribution of x_1 and x_2 together is $b_1^2+b_2^2+2b_1 b_2 r_{12}$. Hence the contribution of x_1 alone is

$$b_1^2+b_2^2+2b_1 b_2 r_{12} - (b_1 r_{12}+b_2)^2 \tag{21}$$

$$= b_1^2 (1 - r_{12}^2).$$

In this particular case we have, therefore, succeeded in splitting up the total contribution (to the total variance of y) of the two indvars into the contributions of each, in what seems to be a consistent fashion. In particular, in the case (already mentioned as trivial) of $r_{12} = o$, the total contribution splits up into b_1^2, and b_2^2, as it should.

Generalisation involves the assumption that variables can be ordered in a causative fashion illustrated in Fig. 5 for $k = 4$.

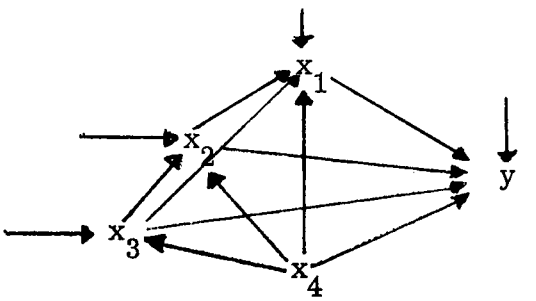


FIG. 5

The full set of equations are

$$
\begin{aligned}
&\text{(i)} && y = b_1x_1+b_2x_2+ \quad \ldots \quad \ldots +b_kx_k+v_o = y_c+v_o \\
&\text{(ii)} && x_1 = b_{12}x_2+b_{13}x_3+ \quad \ldots \quad \ldots +b_{1k}x_k+v_1 \qquad (22) \\
&\text{(iii)} && x_2 = b_{23}x_3+b_{24}x_4+ \quad \ldots \quad \ldots +b_{2k}x_k+v_2 \\
& && \cdot \qquad\qquad \cdot \\
& && \cdot \qquad\qquad \cdot \\
& && \cdot \qquad\qquad \cdot \\
& && x_{k-1} = b_{k-1\cdot k}x_k \qquad \ldots \quad \ldots \qquad +v_{k-1}
\end{aligned}
$$

All k equations in (22) as assumed to be solved by OLS regression. The causation chain is obvious.

Total sum squares in y is $\Sigma y_c^2+\Sigma v_o^2$. When x_1 in (22) (ii) is substituted in (22) (i), the disturbance is $(v_o+b_1v_1)$ which is uncorrelated with $x_2, x_3, \ldots x_k$ so that the equation is exactly the OLS of y on $x_2, x_3, \ldots, x_k$. The difference between Σy_c^2 and $\Sigma y'^2_c$ sum squares for y'_c, the regression of y on $x_2, x_3, \ldots, x_k$, is the contribution of x_1 to total sum squares. Incidentally, it is obvious that this difference must be non-negative. Using (22) (iii) we have the regression of y on x_3, $x_4, \ldots, x_k$ and so determine the contribution to total sum squares of x_2. And so on, to y regressed on x_k alone.

But is this breakdown of Σy_c^2 of (22) (i) unique? The answer is Yes. From the last k-1 equations of (22) each of the remaining indvars could be expressed as a linear function of one particular indvar and of $v_1, v_2 \ldots v_{k-1}$. Substitution for, say, x_2 in (22) (i) would yield an expression in x_2 and a residue a linear function of $v_o, v_1, v_2, \ldots v_{k-1}$, say v. But it would not necessarily follow that $\Sigma x_2 v = o$; hence this linear function for y in terms of x_1 alone would not necessarily be the OLS regression of y on x_2. Hence the $\Sigma y_c^2/n$ would not necessarily represent the contribution of x_2 to total variance. Similarly, it can be shown that only the strict sequence of causation, applied in the manner indicated will, *in general,* yield the contributions of each variable to total variance. Of course, (22) is the recursive set.

If one is interested in only a single depvar there is no need formally to construct the $(k-1)$ OLS regressions, (22) (ii), etc. All that one needs is awareness after due examination of the causal sequence

$$x_k \rightarrow \ldots \rightarrow x_1 \rightarrow y$$

One regresses y on $x_1 \ldots, x_k$, then on $x_2, \ldots, x_k$ etc and so split up sum squares of y_c into the separate contributions of $x_1, x_2, \ldots x_k$. In doing so, incidentally, we establish the coefficients b of the $(k-1)$ regressions in the indvars because of relations like b'_2 (from y'_c) $= b_2+b_1b_{12}$, to find b_{12}, knowing b_1, b_2, b'_2. Practically, this is not a point of much importance since the computer can so easily produce all the regressions. What is really important is the *ex ante* study of relations (with or without a Wright diagram).

As already remarked, there is an immense number of possibilities of relationships between *all* the variables, all of which are worthy of investigation, given the data. Time of occurrence may not be decisive of causation, perhaps because aggregation for a fixed time period (say a year) may impose simultaneity, therefore concealing causation, e.g., of current income as a part cause of value of food consumption.

Here we mention only a few possibilities, confining attention to the case of only one depvar (in general there may be many current endogenous variables each with its equation, endos being possibly part causes in some equations):—

(i) Some variables may be missing from the $(k-1)$ recursive equations between indvars because of *ex ante* considerations or insignificance of value by the t-test. While the *exact* relationships between coefficients referred to above no longer obtain, clearly the procedure for analysing sum squares of y_c is still valid. For theoretical (or algebraic) treatment one can restore the missing variables, and so obtain (22) in full version.

(ii) Relations between indvars are fewer than $(k-1)$. Clearly one can isolate the contribution to sum squares y_c in respect of each of the left hand variables for the equations one has, and amalgamate the effect of the rest.

(iii) The indvar recursive set may contain variables not in the prime depvar equations. The sensible course would appear to be to introduce them formally into the y-equation and proceed as before.

(iv) The case of non-recursive relationship between indvars may be exemplified by the five variable case.

$$\text{(i)} \qquad y = b_1x_1+b_2x_2+b_3x_3+b_4x_4+v_o$$

$$\text{(ii)} \qquad x_1 = b_{13}x_3+b_{14}x_4+v_1 \qquad (23)$$

$$\text{(iii)} \qquad x_2 = b_{23}x_3+b_{24}x_4+v_2$$

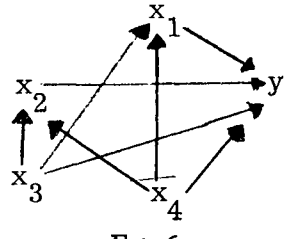


FIG. 6

All three equations are deemed solved by OLS regression, when the expressions for x_1 and x_2 by (ii) and (iii) are substituted on the right side of (i), y is an expression in x_3 and x_4 which is exactly the OLS regression of y on x_3 and x_4. So, we can consistently break up sum squares y into two, sum squares (x_3, x_4) and (x_1, x_2).

This seems the best we can do. If we substitute for x_1 alone we get an expression for y on (x_2, x_3, x_4) but this is *not* in general the regression of y on the three indvars.

But the contribution of x_1 to sum squares y can be formally calculated as above. Having worked the OLS regression of y on (x_2, x_3, x_4) and substituted therein for x_2 by (iii) we get the regression of y on (x_3, x_4) and so can calculate the contribution of x_2 to sum squares y. However, in general this result will not be consistent: the sum of the individual contributions will not in general add to the contribution of x_1 and x_2 together.

In general, it seems that when the indvars can be separated into two groups, one the causative and the others the effects, as in (23) we can only hope to express sum squares y in two classes (i), due to the causative variables (ii), due to the rest.

It would appear that (though the writer has no proof) it is only in the fully recursive case of (22) that one can consistently calculate the contribution of each indvar to sum squares y.

There are many other cases of types of relationship like those considered. The essential point is that in any particular application the causal chain is well worthy of investigation.

It would appear that, with the full causal order given one may infer the recursive relationship and so *uniquely* split up sum squares y into the contributions of each indvar.

Stochastic Aspects

As the fully recursive case is closely related to path coefficients one may recall the Wold-Bentzel theorem that (with disturbances independently and normally distributed) the maximum likelihood solution for the estimation of all coefficients—with asymptotic properties of consistency and efficiency—is obtained by solving each equation separately by OLS regression.

If the OLS regression (22) (i) is complete, i.e., if plausibly the population equation is

$$y = \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_k x_k + u, \tag{24}$$

with u regular and $Eu^2 = \sigma^2$, then σ^2 will be consistently estimated as

$$\Sigma v_o^2/(n-k-1) = s^2$$

It is evident that, in the full recursive case of (22), after each successive elimination of an indvar, the remaining regression, in say $k_p < k$ indvars is, in (22)(i), complete. Now, if to the existing regression of y on the k_p indvars we add a new standardised indvar, x_p, exactly uncorrelated with the existing k_p indvars and solving the (k_p+1) system by OLS regression, yielding a coefficient b_p, it is easy to show that $Eb_p^2 = \sigma^2$ with, of course, $Eb_p = o$.

Accordingly it is suggested that one sets up the F type statistic

$$F_p = b_p^2/s^2$$

with $(1, n-k-1)$ degrees of freedom, x_p now being the variable eliminated. Of course this is identical with the usual analysis for $F_p = t^2$. Generalisation of

procedure with a group of indvars is evident. This analysis will enable one to assess the contribution of each indvar or group to the magnitude of y and also indicate the variables making no significant contribution.

All this is very much as might have been expected. One may well ask why the same procedure could not have been adopted *without* any assumption about the recursive character of the indvars. Suppose four indvars are ordered x_1, x_2, x_3, x_4. Using the above elimination procedure assess the contribution of, say, x_2 to sum squares y. But if one orders as, say, x_4, x_3, x_1, x_2 (as of course, one can without doing violence to the y_c values of the regression) one gets, in general, a different value of the contribution of x_2. It is only because recursiveness imposes an order of causation that a unique solution transpires.

Empirical Treatment

With k possible indvars to start with, in theory there are (2^k-1) possible OLS regressions in all sets of indvars numbering from 1 to k. We conceive it our object to pick the "best", either as a single regression, or a small number of regressions. Our tests of "best" will be by reference to $\bar{R}^2$ as large as possible and a test of probable absence of residual autocorrelation. We are distrustful of regressions with large numbers of indvars (say for k exceeding five) as lacking objective reality, recalling that if k equalled number of sets of observations an exact fit, i.e., $y_c \equiv y$ could be attained even between $(k+1)$ sets of variables picked at random.[7]

Of course when k is large we never try to produce the full (2^k-1) number of regressions: with $k = 10$ this number would be 1023! Instead, using perhaps the full correlation matrix of $k(k+1)/2$ ccs (including the k involving the depvar) and with some speculation as to indvars most likely to be "influential" from the nature of the problem, we considerably reduce the number of regression experiments. Of course, we never lose sight of the fact that OLS regression is a statement of cause-effect, the indvars collectively the cause and the depvar the effect. In eliminating a variable from a regression we are not inferring that such variable is not causal in part but rather that its influence is taken up by the indvars we retain.

All this is rank empiricism. What the Sewall Wright approach does is to insist on sequential causal order in the elimination, one by one, of indvars. A change in the order will not result, in general, in the correct contributions of the eliminated variables to total variance. We have shown that this orderly elimination is associated with a recursive set of OLS regression equations in the $(k+1)$ variables.

Sequential ordering on Wright lines may not always be possible, especially when dealing with cross-section data. With time series, it may help to order indvars according to time of occurrence, assuming the earlier event to be causal. The time lag may be infinitesimal, as in the case of a consumption function with income as an

7. A statistician of old remarked "Give me five parameters and I will make the dog stand up and talk".

indvar: income is deemed to precede consumption. It is only when we have causally ordered the data as in Fig. 5 that we can calculate the contribution of individual variables to the total variance of the depvar.

Birth Weight of Guinea Pigs Reconsidered

This "simplest application" of Wright's admirably illustrates the theory developed in the last few sections. The standardised variables[8] are

y = average weight at birth

x_1 = length of gestation period

x_2 = size of litter

The causative sequence is shown on Fig. 4. The OLS regression equations are at (18). The ccs (given by Wright) required for solution are

$$r_1 = 0{\cdot}56;\ r_2 = -0{\cdot}66;\ r_{12} = -0{\cdot}48.$$

Using the standard equations the y-coefficients are

$$b_1 = 0{\cdot}3160;\ b_2 = -0{\cdot}5083.$$

The total variance of y is 1. Contributions of the variables and disturbance are as follows, using the formulae given earlier

Contribution of x_1	$= b_1^2(1-r_{12}^2)$	$= 0{\cdot}0769$
„ „ x_2	$= (b_1 r_{12} + b_2)^2 = r_2^2$	$= 0{\cdot}4356$
„ „ x_1 and x_2		$= 0{\cdot}5125$
„ „ disturbance		$= 0{\cdot}4875$

The contribution of x_2, size of litter, is over five times that of x_1, length of gestation period. Size of litter has a very much greater influence on average weight at birth than has length of gestation period, confirming broadly Wright's conclusion.[9] Wright's method, however, fails to reveal that the two causes together account for little more than half the total variance of y, average weight at birth.

Conclusions

The method path coefficients might be described as OLS regression together with relations between indvars all without disturbance terms. The latter is less of a disadvantage than might at first appear since in practice the method exploits

8. Notation has been changed from that used in the first example to bring application exactly into line with that of formula (18) and Fig. 4.

9. The fact that Wright's "over three times" and the "over five times" here is attributable mainly to the dimensions of the statistics on which the statement is based.

mainly correlation, whereby the disturbance terms would be eliminated even if they were introduced into the system. Our contribution has been to bring in disturbances. One valuable feature of Wright's method lay in the estimation of ccs involving variables for which data were not explicitly provided, though deemed necessary for analysis.

Related to the latter aspect is the main feature of the method. This is the use of a sequential (or ordered) causal chain involving all the variables, copiously illustrated in diagrams by Sewell Wright. In this paper it is shown that if all the indvars can be so ordered, there results a recursive system of $k-1$ equations, in addition to the original OLS regression. As each equation is a causal statement it may be solved by OLS regression.

What Wright's work and this paper show is that the solution of the single OLS multivariate equation is not enough, even when endowed with all the customary paraphernalia of t-values for coefficients, F, R^2, tests for absence of residual autocorrelation and even the full correlation matrix. With the single equation approach we may be under-using the data available to us. We are positively doing so if the variables can be made to observe Wright's principle of causal ordering.

In extending Sewall Wright's work we have shown here that, when the indvars are recursively related, the contribution of each indvar to the sum of squares of the depvar can be ascertained, in general uniquely. Heretofore, this was known to be true only when the indvars were uncorrelated, a trivial case from the practical viewpoint. Hence, we recommend that in multivariate OLS regression indvars should be examined for the possibility of the existence of recursive relations between them.

As a concluding remark: should we not, following Wright, in all cases examine our data, which usually are all that is available relevant to our problem, in the first instance to seek a complete or partial causal chain, or to set down the full system, whatever its character, for solution? Many computer systems have programs for solution of the general simultaneous equation system, so that difficulty of solution is no longer a consideration. May the single OLS regression system be on the way out, and should we practitioners not give it a gentle push on its way, while grateful for its services?

The Economic and Social Research Institute, Dublin

REFERENCES

GEARY, R. C., 1963. "Some remarks about relations between stochastic variables: A discussion document". *Review of the International Statistical Institute.*

GEARY, R. C., 1970. "Relative Efficiency of Count of Sign Changes for assessing residual autoregression in least squares regression", *Biometrika*, Vol. 57, No. 1.

WRIGHT, SEWALL, 1934. "The Method of Path Coefficients", *Annals of Mathematical Statistics*, Vol. 5.