THE ECONOMIC
AND SOCIAL
RESEARCH INSTITUTE
MEMORANDUM SERIES
NO. 105

The Method of Path Coefficients and OLS Regression

by

R. C. Geary

October 1975

The Method of Path Coefficients and OLS Regression*

By R. C. Geary.

The object of this paper is to study the relationship between the theory (and practice) of path coefficients and OLS regression, mainly in the hope that thereby some light might be shed on the still dark patches in the theory of relationship between random variables.

Path Coefficients

The fundamental paper+ on path coefficients, by Sewall Wright, over 40 years old, is greatly to be admired for its comprehensiveness and thoroughness. I am unaware that the method is used much nowadays, so a brief summary may be desirable, with special reference to two of Wright's telling applications. In the following exposé, different notation from Wright's is used.

Write

$$y = b_1 x_1 + b_2 x_2 + \ldots + b_k x_k, \tag{1}$$

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

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

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

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

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

---

* I am indebted to Sir Maurice Kendall for suggesting the problem discussed in this paper.

+ Sewall Wright: "The Method of Path Coefficients, "Annals of Mathematical Statistics, Vol. 5, 1934

with only 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, $b_1 = r_1$, the coefficient of correlation (c.c.) between y and x_1. In general the path coefficients are functions of the c.c.s 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 diagramatically. Thus Fig. 1. with its single and

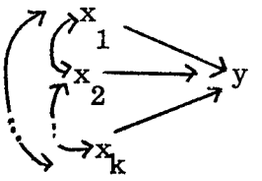

Fig. 1 Causal representation of formula (1)

double-headed arrows. With the single-headed, the head indicates the effect, the depvar, the other end the cause, the indvar. The double-headed arrows 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

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. We reduce the report to bare essentials. Let

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

Three c.c.s 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 $r_2 = +.56$, $r_3 = -.66$, $r_{23} = -.48$.

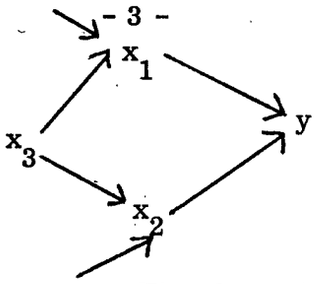

Fig. 2

Note that (1) x_1 is not precisely defined, (2) that x_1 and x_2 can have causative factors* 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

$$\text{(i)}\quad y = b_1x_1 + b_2x_2$$

(3) $$\text{(ii)}\quad x_1 = r_{13}\,x_3$$

$$\text{(iii)}\quad x_2 = r_{23}x_3$$

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

$$1 = b_1^2 + b_2^2 + 2\,b_1b_2\,r_{12}$$

$$\text{(i)}\quad = b_1^2 + b_2^2 + 2\,b_1b_2\,r_{13}r_{23} \text{ (using (3) (ii) and (3) (iii)}$$

(4) $$\text{(ii)}\quad .56 = b_1\,r_{13}\,r_{23} + b_2$$

$$\text{(iii)}\quad -.66 = b_1\,r_{13} + b_2\,r_{23}$$

(4) consists of three equations to determine three unknowns the path coefficents b_1 and b_2 and c.c. r_{13}, the only other quantity involved, namely r_{23}, $(= -.48)$ being given. Though the equations (4) are non-linear an unique solution is easily derivable.

(5) $$b_1 = .87,\quad b_2 = .30,\quad r_{13} = -.59,$$

to which we add $r_{23} = -.49$. Then, from (4) (iii), the c.c. between average weight at birth and size of litter, namely $r_3 = -.66$ breaks into two parts (on the right).

$$b_1\,r_{13} = -.51 \text{ and } b_2\,r_{23} = -.15.$$

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

---

* I.e. Indicated by arrows with provenance undefined.

introduced into the calculation and its 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 -.51 and -.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 careful 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

(6)
$$\text{(i)}\quad r_2 = .56 = c_3 + c_2 r_{23}$$
$$\text{(ii)}\quad r_3 = -.66 = c_3 r_{23} + c_2$$

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

> "The term $[c_3 = -.51]$ 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.*

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 precentage changes in quantity and price respectively and again assuming linearity

---

* It is true; see later

$$X_d = \eta Y + D$$

(7)

$$X_s = \epsilon Y + S$$

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

$$x = b_{11} d + b_{12} s$$

(8)

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

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

(9)

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

$$\text{(ii)}\quad 1 = b_{21}^2 + b_{22}^2 + 2\, b_{21} b_{22}\, r_{sd}$$

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

Causal relations are indicated on Fig. 3

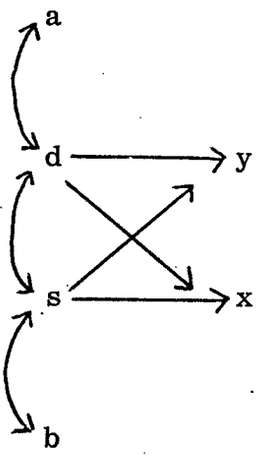

Fig. 3

(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 standardised, from (8) we easily derive four additional equations

(10)
$$\text{(i)}\quad r_{ax} = b_{11} r_{ad}$$
$$\text{(ii)}\quad r_{ay} = b_{21} r_{ad}$$
$$\text{(iii)}\quad r_{bx} = b_{12} r_{bs}$$
$$\text{(iv)}\quad r_{by} = b_{22} r_{bs}$$

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 c.c.s 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 $r_{sd} = o$. 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 is

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

Values of the price elasticities are $\epsilon = b_{11}/b_{21} = .192$ for supply and $\eta = b_{12}/b_{22} = -1.361$ for demands*.

It should be pointed out, in regard to this second example,

---

* These values calculated from Wright's formulae differ considerably from those given by Wright, namely $\epsilon = .133$ and $\eta = -.944$, for reason unknown.

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 application, not within the theory developed.

Again we see illustrated, in s and d, the possibility of deriving functions (coefficients and c.c.s) 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.

Its outstanding characteristic is that it is non-stochastic, except in the very minor degree that there is mention of asymptotic estimates of standard errors of means, c.c.s 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 c.c.s, 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.

The Nature of Linear Relationship between Variables

The OLS estimate of the coefficient b in the simple regression

(12) $$y = bx + v = y_1 + v$$

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

(13) $$\sum vx = o,$$

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 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

(14) $$y = bx + cv$$

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

(15) $$r_{yx} = b + c \sum vx/n$$
$$r_{yv} = b \sum vx/n + c$$

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

(14a) $$y = \sum_{i=1}^{k} b_i x_i + v = y_c + v$$

the standard equations for estimating the b_i may be written

(16) $$\sum vx_i = o, \quad i = 1, 2, \ldots, k.$$

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 the v 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 the Geary γ, for adjudging the completeness of y_c as estimates of data y, by the test for residual nonauto-regression.

From (14a) using (16),

$$1 = \frac{1}{n}\sum y_c^2 + \frac{1}{n}\sum v^2, \tag{17}$$

so that $\sum y_c^2/n = R^2$ (or its variant $\bar{R}^2$, i.e. R^2 corrected for d.f.) is the principal measure of the extent to which the k indvars represent the y. In the case of simple OLS regression $R^2 = r^2_{yx}$. It is somewhat unfortunate that R^2 (and not R) intervenes as we may feel that R is a truer measure of relationship, as, with appropriate sign, is the case in simple c.c analysis.

Though, as far as he knows, the writer's view,* reasonably closely argued, that the individual coefficients in multivariate OLS regression are meaningless (except in the trivial case of all indvars being uncorrelated) has never been formally refuted, it is probably not accepted by most statisticians. Nevertheless the writer still holds it to be true. It is the whole vector of coefficients that matters, mainly for forecasting, or at 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,

* R. C. Geary:

possible to estimate the contribution of individual variables to the total variance of y. It is possible only to assess the total effect, namely $\sum y_c^2/n = R^2$. 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

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

The Sewall Wright diagram would be as Fig. 4.

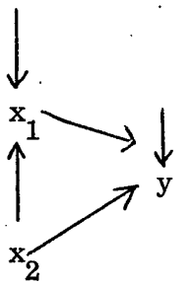

Fig. 4

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

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

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

$$\text{(20)} \qquad \sum x_1 v = o, \sum x_2 v = o, \sum x_2 w = o$$

Hence $\sum x_2 (v + b_1 w) = o$ 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 $b_1^2 + b_2^2 + 2 b_1 b_2 r_{12}$. Hence the contribution of x_1 alone is

$$\begin{aligned} &b_1^2 + b_2^2 + 2 b_1 b_2 r_{12} - (b_1 r_{12} + b_2)^2 \\ &= b_1^2 (1 - r_{12}^2). \end{aligned} \qquad \text{(21)}$$

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.

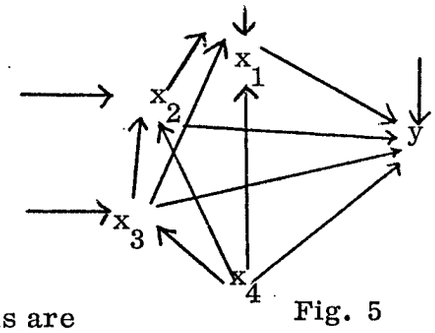

Fig. 5

The full set of equations are

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

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

Total sum squares in y is $\sum y_c^2 + \sum v_o^2$. 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 the OLS of $x_2, x_3, \ldots,$ x_k on y. The difference between $\sum y_c^2$ and $\sum' y_c^2$, sum squares for this 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 $\Sigma\, y_c^2$ 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 $v_1, v_2, \ldots v_{k-1}$. Substituion 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, (28) is the well known recursive set.

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*.

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 c.c.s

---

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

(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 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 two sections. The standardised variables* are

---

* 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.

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 c.c.s (given by Wright) required for solution are

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

Using the standard equations the y - coefficients are

$$b_1 = .3160; \quad b_2 = -.5083$$

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

Contribution of $x_1 = b_1^2 (1 - r_{12}^2) =$	.0769
" " $x_2 = (b_1 r_{12} + b_2)^2 = r_2^2 =$	.4356
" " x_1 and x_2	= .5125
" " disturbance	= .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.* 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.

Conclusion

The method of path coefficients might be described as OLS regression without a disturbance term. This is less of a disadvantage than might at first appear since in practice the method exploits mainly correlation, whereby the disturbance term would be eliminated even if it were introduced

---

* 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

into the system. (Our modest contribution is to so introduce it, with consequent emphasis on variance rather than correlation.) One valuable feature of the method lay in the estimation of c.c.s 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 Sewall 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. (One recalls the Bentzel-Wold theorem that in a recursive system with disturbances independently and normally distributed, the maximum likelihood solution of the estimation of all coefficients is obtained by solving each equation separately by OLS).

When the writer controversially maintained that individual coefficients in a multivariate regression are, in general, meaningless, he wishes he had recalled Wright's paper, for then he might have come up with the present proposal for making them meaningful. The recursive system, essentially due to Wright, when it can be used, goes far towards solving this problem, in its enabling the estimation of the contribution of each variable to the total variance of the depvar. Recursiveness is a condition <u>sufficient</u> in character; one would like to know if it is also <u>necessary</u>.

What Wright's work and this paper shows 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 underusing the data available to us. We are positively doing so if the variables can be made to observe Wright's principle of causal ordering.

Of course, in the social sciences we are familiar with the theory and practice of systems of linear simultaneous equations in endogenous and predetermined variables. In setting these up, one is accustomed to regarding each equation as a cause-effect statement with one endo as the effect. But the causal linear expression may contain one or more current endos. The system need not necessarily be recursive. Incidentally, in such case there would be no difficulty in devising a Wright-type diagram. (Hint: introduce two single-headed arrows between two variables but pointing different ways.)

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 programmes for solution of the general simulanteous 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?

22 October 1975. R. C. Geary