CAUSAL MODELLING : THE SIMON-BLALOCK APPROACH

D.G.Pringle

CATMOG

27
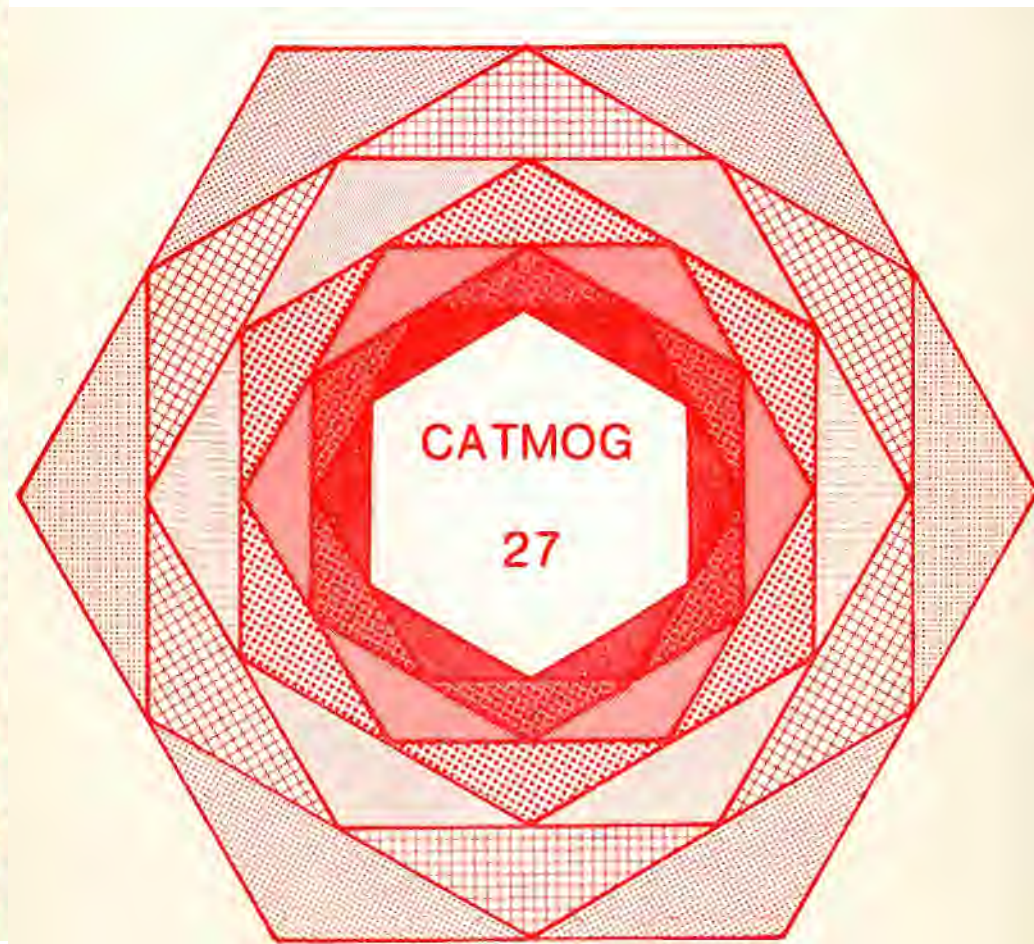
<u>CATMOG</u>

(Concepts and Techniques in Modern Geography)

CATMOG has been created to fill a teaching need in the field of quantitative methods in undergraduate geography courses. These texts are admirable guides for the teachers, yet cheap enough for student purchase as the basis of classwork. Each book is written by an author currently working with the technique or concept he describes.

*Continued on inside back cover*

<u>CAUSAL MODELLING</u> : <u>THE SIMON-BLALOCK APPROACH</u>

by

D. G. Pringle

(Maynooth University, Ireland)

<u>CONTENTS</u>

# I INTRODUCTION

## (i) Objectives and prerequisites

One of the principal objectives in geography is to develop explanatory theories of causes and effects, but in some instances alternative (and possibly conflicting) explanations may appear to explain the same phenomena. This monograph outlines a fairly simple approach which may be used to evaluate the extent to which alternative hypothesized causal models are consistent with empirically observed data. This approach is referred to as the Simon-Blalock technique.

The Simon-Blalock technique is one of a family of causal modelling techniques, some of which are briefly reviewed in Chapter V. It is associated, as its name might suggest, with the work of Hubert Blalock and Herbert Simon - two sociologists. Despite the familiarity of many geographers with Blalock's book Causal Inferences in Nonexperimental Research (Blalock, 1964), in which the technique is developed from the earlier ideas of Simon (1954;1957), geographical examples of the Simon-Blalock technique are few and far between. Notable exceptions are provided by Cox (1968) and Mercer (1975), although Cox's paper attracted a certain amount of adverse criticism (Cox, 1969; Kasperson, 1969; Taylor, 1969). Examples from other disciplines (e.g. Goldberg, 1966; Wilbur, 1964) suggest, however, that the Simon-Blalock technique, and indeed causal modelling techniques in general; deserve greater geographical attention.

The Simon-Blalock technique is preferred here to other causal modelling techniques due to its relative simplicity, although it often appears very complicated to those not familiar with it. However, it is basically a simple extension of correlation and regression techniques, therefore the only prior knowledge assumed of the reader is a basic understanding of correlation, regression, and partial correlation. Good introductions to these techniques are provided by Ferguson (1977), Johnston (1978), and Nie et. al. (1975, Chapters 18 to 20).

## (ii) Causality

It is useful to briefly consider what we mean by causality. Causality implies at least three conditions:

1. Covariance. If variable X is thought to be a cause of variable Y, variations in X over space or time should correspond with variations in Y. In other words, large quantities of X should (assuming a positive relationship) be associated with large quantities of Y. Areas with a high rainfall (variable X), for example, should be subject, all other things being equal, to more flooding (variable Y) than areas of low rainfall.

2. Temporal precedence. If X is a cause of Y, changes in X should occur before the corresponding changes in Y. In other words, flooding would be expected to follow higher rainfall, rather than vice-versa.

3. Production. In a causal situation a change in X is not simply followed by a change in Y, but rather the change in X should produce the change in Y. Wet periods are sooner or later followed by dry periods, but we do not speak of dry periods being caused by wet periods because wet periods do not produce dry periods. Flooding, however, is believed to be produced by rainfall (in conjunction with other factors), therefore the relationship is causal.

The third condition is the most difficult to prove. In an experimental situation we could manipulate a change in X to see what effect it has upon Y while holding the values of all other variables constant. In a non-experimental situation (which is usually the case in geography) we can neither control the other variables nor manipulate a change in X. All we can do is observe the system and attempt to figure out what is causing what from a knowledge of covariations and causal sequence.

The Simon-Blalock technique is useful in this respect by providing a systematic procedure for testing whether a hypothesized model of cause and effect for a sub-system of variables is consistent with observed data. The technique essentially entails identifying the covariances which could be expected if the hypothesized model is correct, and then testing to see if these covariances are found in reality. The approach is therefore deductive. In other words, the model is hypothesized on a priori reasoning before being empirically tested. Blalock (1964, 77) gives an example of how the technique may also be used inductively to derive an empirically consistent model which then has to be theoretically interpreted, but this approach is not recommended because several models may be equally consistent with the data.

Since one cannot prove production in a non-experimental context, one cannot prove that a given hypothesized model is correct even if it is consistent with empirical data. Rather, all one can do is to show that there is no empirical reason for believing it not to be correct. If, on the other hand, a given hypothesized model is not consistent with empirical data, we have strong grounds for believing the model to be incorrect. By ruling out incorrect models, the Simon-Blalock technique may be used to identify the 'true' explanation by a process of elimination. However, no matter how well a hypothesized model agrees with the data, one must always allow for the possibility that an even better model still remains undiscovered. Thus, all theories, even if consistent with empirical observations, must be regarded as working hypotheses rather than absolute truths.

It is important to note that the Simon-Blalock technique entails a different logic from that usually adopted in hypothesis testing. Normally one identifies a null hypothesis ($H_0$) which is the opposite of the hypothesis ($H_1$) which the researcher hopes to prove. The test is organized so that if it can be shown that $H_0$ is inconsistent with the data, it follows that $H_1$ must be correct, unless one has made a Type I error (Siegel, 1956, 9). In the case of causal modelling, however, there are an infinite number of possible alternatives, therefore the objective is not to prove that a null hypothesis (or model) is wrong, but rather that there is no reason to believe that the hypothesized model is wrong. As noted above, it is not actually possible to prove that it is correct.

(iii) Causal models

When the causal relationships between a number of variables are considered simultaneously, verbal statements of what is hypothesized as a cause of what become quite complex. It is therefore useful to represent a hypothesized causal model diagrammatically. A causal link is usually represented by a one headed arrow as in Figure 1(a). The direction of the arrow indicates the direction of the causal link. In this case X is believed to be a cause of Y.



(a)                    (b)

(c)                    (d)
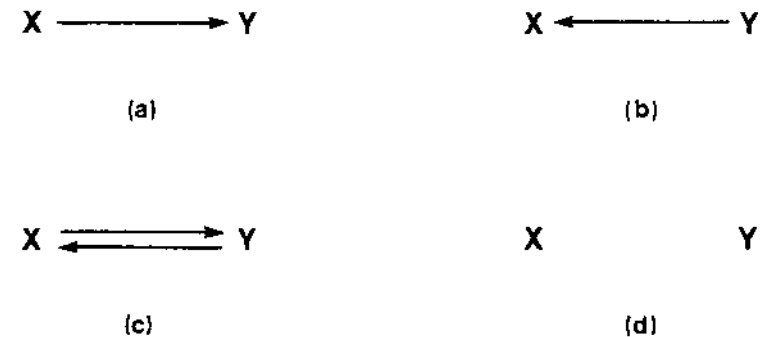
Fig 1. Possible causal relationships between two variables.

With two variables there are only four possible causal combinations:

1. X is a cause of Y, as in Figure 1(a).

2. Y is a cause of X. This would be represented by an arrow leading from Y to X (Figure 1(b)).

3. X may be a cause of Y, but Y may also be a cause of X. An example of this is provided by consumer prices and wages. Increased prices create a demand for higher wages. If granted, higher wages may necessitate a further increase in prices to cover production costs. Reciprocal causal relationships of this type are usually represented by two arrows, one in either direction (Figure 1(c)), but they are difficult to deal with and require data to be available for time series. It is consequently assumed for the purposes of this monograph that all causal relationships are unidirectional, but the possibility of feedback effects must also be kept in mind.

4. X and Y may be causally unrelated. This situation is simply represented by omitting the arrow (Figure 1(d)).

In practice we usually have to deal with more than two variables. Even though we may be primarily concerned with the relationship between X and Y, we must consider how the relationship is influenced by other variables. A high correlation between X and Y does not necessarily indicate a direct causal relationship. Two possible alternatives are shown in Figure 2.

Fig 2. Two three variable models.

In Figure 2(a) variable X causes variable Z and variable Z causes variable Y, but there is no direct link between X and Y. Thus, for example, areas with a high proportion of rented apartments (variable X) may attract higher percentages of young unmarried people (variable Z) than areas with a different tenure type. Areas with a high percentage of young unmarried people may have a higher rate of illegitimate births (variable Y). Although illegitimacy is correlated with tenure type, it would be very misleading to claim that rented apartments 'cause' illegitimacy, without at least elaborating upon the contention by identifying the intervening variable Z. One of the major objectives in science as a whole is to provide more complete explanations by correctly identifying the role played by intervening variables.

Figure 2(b) represents a more serious situation. In this case the observed correlation between X and Y is _spurious._ Neither variable has any effect upon the other, either directly or indirectly, yet they will be correlated due to common dependence on Z. For example, poor sanitation (variable Z) may cause a bad 'odour (variable X). It may also cause serious infectious diseases such as cholera (variable Y). Until one identifies the underlying variable Z, the high correlation between bad smells and disease might lead one to adopt totally ineffective preventive measures such as the use of air fresheners. In this example an underlying variable Z would probably be suspected fairly rapidly, but in other instances an underlying variable might not be suspected and one might be quite content to accept the erroneous hypothesis that X causes Y.

The only safeguard in these situations is to examine more variables than those of primary interest. Thus, for example, if we are interested in the relationship between rainfall and flooding it would be useful to consider the effects of other variables such as vegetation cover, topography, river capacity etc. However, increasing the number of variables also has drawbacks because the models become increasingly complex.

The number of possible causal links (ignoring direction) increases rapidly with increases in the number of variables. For two variables there is only one possible causal link, for three variables there are three possible links, and for four variables there are six possible links. In general, the maximum number of possible links P is given by

$$P = \tfrac{1}{2}V(V-1) \qquad (1)$$

where V is the number of variables.

If we assume that there are no reciprocal links, and if we assume that the causal sequence of the variables (and hence the direction of the links) is known, then for each possible link there are only two possibilities - either the link is present or it is absent. The number of possible model configurations C is therefore given by

$$C = 2^P \qquad (2)$$

The values of C for V equal to 2 to 6 are given in Table 1.

| Number of Variables V | Possible Links P | Possible Configurations C |
|---|---|---|
| 2 | 1 | 2 |
| 3 | 3 | 8 |
| 4 | 6 | 64 |
| 5 | 10 | 1024 |
| 6 | 15 | 32768 |

Table 1. Maximum number of configurations for V variables.

The rapid increase in the number of configurations is obvious, although some of the configurations with very few links are rather trivial. The 8 configurations for 3 variables are shown in Figure 3.

The actual number of possible models is much higher and depends upon the causal ordering of the variables. In Figure 3, for example, it is assumed that X precedes Y (i.e. if a link exists it is from X to Y rather than vice versa), and that Y precedes Z. However, there are V: (V factorial) ways in which V variables may be ordered, and for each ordering there are C configurations. The maximum number of models possible for V variables therefore approaches $CV'$. (The actual number is slightly less than CV: because models with no causal links will be equivalent irrespective of causal ordering). The chances of discovering the correct model decreases rapidly as the number of variables is increased, therefore one should try to keep hypothesized models as simple as possible whilst ensuring that no major variable has been inadvertently omitted.

(iv) Comparison with regression

The Simon-Blalock technique, as will be seen in Chapter II, is closely related to regression techniques. It is therefore useful to note some of the more obvious differences between the Simon-Blalock technique and multiple regression.

The first difference is that the Simon-Blalock technique examines a whole system of causes and effects, whereas regression is only concerned with the effects of a number of variables upon one variable (i.e. the dependent variable).

A second, and more fundamental difference, is that regression techniques assume a model, such as depicted in Figure 4, and then attempt to estimate the values of the parameters. The objective of the Simon-Blalock technique, on the other hand, is to test whether a hypothesized model is consistent

(a)



(b)



(c)



(d)



(e)



(f)



(g)



(h)
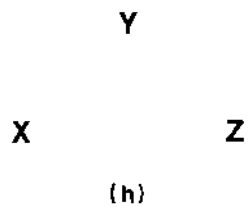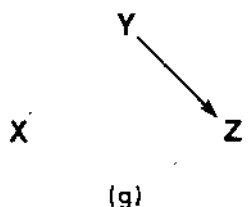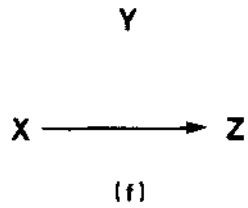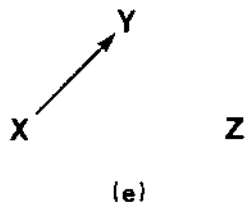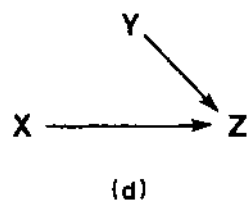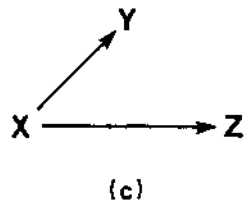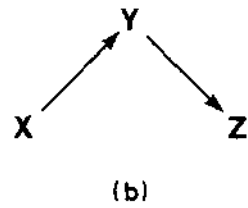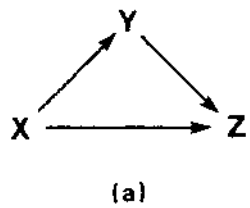
Fig 3. Possible configurations between three variables.

with empirical data. The values of the parameters are only of interest to the extent that they facilitate the testing of the model.
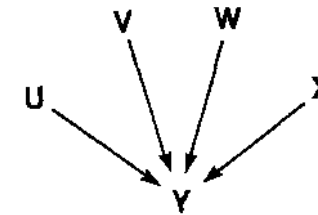


Fig 4. Multiple regression model with four independent variables.

## II THE SIMON-BLALOCK TECHNIQUE

Having outlined the general objectives of the Simon-Blalock technique in Chapter I, the technique is now explained in more detail in the present chapter.

(i)  Derivation of the prediction equations

The mathematical basis of the Simon-Blalock technique may be illustrated by considering a four variable example.
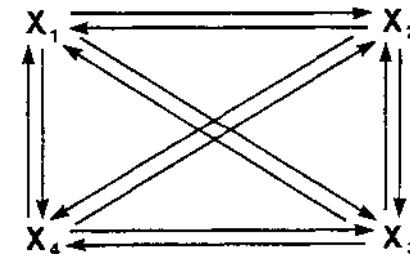


Fig 5. Four variable model with reciprocal causal links.

Figure 5 shows a model in which each of the four variables ($X_1$, $X_2$, $X_3$, and $X_4$) are causally dependent on each of the others. This situation is unlikely to arise in practice, but it is didactically useful. Each of the four variables may be represented by a multiple regression equation:

$$X_1 = b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 + e_1$$
$$X_2 = b_{21.34}X_1 + b_{23.14}X_3 + b_{24.13}X_4 + e_2$$
$$X_3 = b_{31.24}X_1 + b_{32.14}X_2 + b_{34.12}X_4 + e_3 \qquad (3)$$
$$X_4 = b_{41.23}X_1 + b_{42.13}X_2 + b_{43.12}X_3 + e_4$$

where the variables are expressed in standardized form to eliminate the regression constants, the b's are standardized partial regression coefficients, and the e's refer to stochastic elements and to the effects of variables not specifically included in the model. The partial regression coefficients measure the effects of the independent variables upon the dependent variable, if the effects of the other variables were controlled for. The first subscript, by convention, indicates the dependent variable, the second indicates the independent variable, and those after the dot indicate the control variables. Thus, for example, the coefficient $b_{12.34}$ measures the effect of $X_2$ upon $X_1$ controlling for the effects of $X_3$ and $X_4$. Equation (3) states that the values of each variable are the sum of the effects of the other three variables plus the unknown factors represented by e.

If we assume that there are no reciprocal relations, and that variable $X_1$ causally precedes $X_2$ which precedes $X_3$ which precedes $X_4$ (thereby eliminating feedback effects), the situation is considerably simplified. The model is shown in Figure 6, and the corresponding equations are:

$$X_1 = e_1$$
$$X_2 = b_{21}X_1 + e_2$$
$$X_3 = b_{31.2}X_1 + b_{32.1}X_2 + e_3 \quad\quad (4)$$
$$X_4 = b_{41.23}X_1 + b_{42.13}X_2 + b_{43.12}X_3 + e_4$$

Each variable is now only dependent upon variables which precede it, therefore the first variable is not dependent upon any of the other three variables in the model, but simply upon the external factors represented by $e_1$. It will also be noted that the b terms are not only fewer in number compared to equation (3), but, with the exception of those for $X_4$, are of a lower order. In other words, it is only necessary to control for variables which precede the dependent variable.
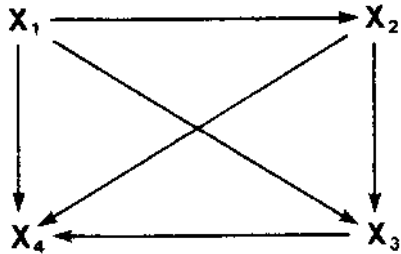


Fig 6. Four variable model with no reciprocal links.

The 6 partial regression coefficients in equation (4) have unknown values, as do each of the disturbance terms e, therefore the model is said to be underidentified because the number of unknowns exceeds the number of equations. However, if we assume that the e terms are independent of one another we obtain a further 6 equations, viz.

$$E(e_1e_2) = 0$$
$$E(e_1e_3) = 0$$
$$E(e_1e_4) = 0$$
$$E(e_2e_3) = 0 \quad\quad (5)$$
$$E(e_2e_4) = 0$$
$$E(e_3e_4) = 0$$

i.e. the expected value of the cross product of any two disturbance terms is zero. We now have the same number of equations as unknowns, therefore the model is exactly identified and we can solve to find a unique value for each of the unknowns using ordinary least squares (OLS) (Blalock, 1964, 63).

The model depicted in Figure 6 has a causal link connecting each pair of variables, and each link corresponds to one of the b coefficients in equation (4). If, however, one or more of the causal links are absent, then the corresponding partial regression coefficient(s) should have a value of zero. For example, the model depicted in Figure 7 is the same as that in Figure 6 except there is no direct link between $X_1$ and $X_3$. The corresponding regression equations are:

$$X_1 = e_1$$
$$X_2 = b_{21}X_1 + e_2$$
$$X_3 = b_{32.1}X_2 + e_3$$
$$X_4 = b_{41.23}X_1 + b_{42.13}X_2 + b_{43.12}X_3 + e_4$$

which is equivalent to saying that

$$b_{31.2} = 0 \quad\quad (6)$$
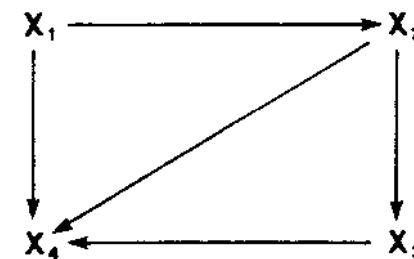


Fig 7. Four variable model with one missing link.

Equation (6) is known as a prediction equation because it predicts the value of $b_{31.2}$ if the model is correct. The model can be easily tested by calculating the value of $b_{31.2}$ to see if it is in fact zero. If it is not, then the data are not consistent with the model depicted in Figure 7, and the onus is on the researcher to find an alternative model which does provide accurate predictions. If, on the other hand, $b_{31.2}$ is close to zero, one has reason to believe that the hypothesized model may be correct. Due to measurement and sampling errors, it is very unlikely that the partial regression coefficient will be exactly zero even if the model is correct, so in practice

the researcher is usually content to hypothesize a model with prediction equations which are approximately accurate rather than totally accurate.

It will be noted that there is one prediction equation for every 'missing link' (i.e. pair of variables not directly connected by an arrow). Thus, for example, the model depicted in Figure 8 would be represented by the regression equations:

$$X_1 = e_1$$
$$X_2 = b_{21}X_1 + e_2$$
$$X_3 = b_{32.1}X_2 + e_3$$
$$X_4 = b_{42.13}X_2 + b_{43.12}X_3 + e_4$$

which give rise to two prediction equations:

$$b_{31.2} = 0$$
$$b_{41.23} = 0 \qquad (7)$$

This means that the simpler the hypothesized model (i.e. the fewer the number of direct causal links), the more prediction equations there are to test it. If a simple model is found to be consistent with the empirical data, then we may place more confidence in it than in a more complex model which is supported by a smaller number of prediction equations.
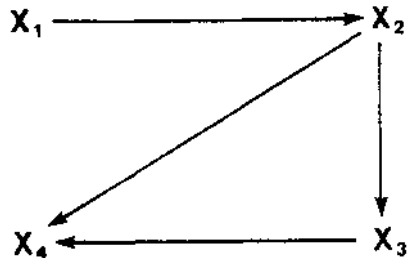


Fig 8. Four variable model with two missing links.

To simplify matters partial correlation coefficients may be substituted for partial regression coefficients. If a partial regression coefficient is zero, it can be shown that the corresponding partial correlation coefficient is also zero. For example, the formula for $b_{31.2}$ is

$$b_{31.2} = \frac{s_3}{s_1} \cdot \frac{(r_{13}) - (r_{12})(r_{23})}{(1 - (r_{12})^2)}$$

where s is the standard deviation. If $b_{31.2}$ is zero, the numerator must be zero. However, the formula for $r_{31.2}$ is

$$r_{31.2} = \frac{(r_{13}) - (r_{12})(r_{23})}{\sqrt{(1 - (r_{12})^2)}\sqrt{(1 - (r_{23})^2)}}$$

As this has basically the same numerator, it follows that if $b_{31.2}$ is zero, $r_{31.2}$ must also be zero. The same applies to other partial regression and partial correlation coefficients, therefore the prediction equations in equation (7) may be substituted by

$$r_{31.2} = 0$$
$$r_{41.23} = 0 \qquad (8)$$

Correlation coefficients do not change their value if the dependent and independent variables are interchanged, therefore equation (8) is equivalent to

$$r_{13.2} = 0$$
$$r_{14.23} = 0$$

Equation (7), however, is not the equivalent of

$$b_{13.2} = 0$$
$$b_{14.23} = 0$$

because the value of a regression coefficient is governed by which variable is the dependent variable and which is the independent.

(ii) The rule of thumb

The Simon-Blalock approach may be summarized by saying that for each missing link in a hypothesized causal model there is a prediction equation of the form

$$r_{ab.cde...} = 0 \qquad (9)$$

where a and b are the variables at either end of the missing link, and c,d,e etc. are the variables that causally precede either a or b. It is essential that c,d,e etc. should not include any variables which are causally subsequent to both a and b.

To correctly identify which variables need to be controlled for, trace back the paths depicted by the arrows in the causal diagram. In other words, beginning at variable a, note all the variables which are directly connected to a by an arrow pointing towards a. Then note all the variables connected by arrows leading to these variables, and repeat until all the causal paths have been traced back to their origins. Repeat the procedure for variable b. All of the variables noted in this process should, under the rule of thumb, be controlled for in the partial correlation between a and b, although some. may, under special conditions, be omitted (see computational simplifying rules below).

There should be one prediction equation for each missing link. To check that none have been inadvertently omitted, one should calculate the maximum number of possible links using equation (1), and then subtract the number of arrows in the model to find the number of missing links. This gives the number of prediction equations.

(iii) Computational simplifying rules

Although one could test any given causal model by simply applying the rule of thumb, some of the prediction equations would be of a very high order.

As high order partial correlation coefficients are more difficult to calculate than low order coefficients, it may be useful, if possible, to replace prediction equations which use high order coefficients by equivalent equations which use lower order coefficients. This simplifies the calculations. The disadvantage is that these modifications tend to confuse the logic of the technique, especially for beginners, with the result that there is a greater likelihood of a prediction equation being incorrectly identified. Five rules which may be used to simplify the calculations are given below, but readers who find them confusing are advised to ignore them and skip to the beginning of Chapter III. This is particularly true for those who have access to a computer program which is capable of calculating high order partial correlation coefficients (e.g. SPSS).

To facilitate discussion the following terminology is adopted to describe the variables associated with each missing link. If the link is missing between variables a and b, and variable b precedes variable a in the causal sequence, then variable a is referred to as the <u>dependent</u> variable and variable b as the <u>independent</u> variable. The other variables c,d,e etc., which causally precede either a or b, are referred to as control variables. The subset of c,d,e etc. which causally precede the dependen· variable but not the independent variable are referred to as <u>intervening</u> variables.

The five simplifying rules are:

Rule 1: A prediction that a first order partial correlation coefficient is equal to zero may be replaced by one involving only zero order coefficients, as follows:

$$\text{If } r_{ab.c} = 0,$$

$$\text{then } r_{ab} = r_{ac} \cdot r_{bc}$$

This may be easily verified by considering the formula for $r_{ab.c}$:

$$r_{ab.c} = \frac{(r_{ab}) - (r_{ac})(r_{bc})}{\sqrt{1 - r_{ac}^2} \sqrt{1 - r_{bc}^2}}$$

If $r_{ab.c}$ is zero, the numerator must be zero or else the denominator must be infinite. The denominator, however, has a maximum value of 1, therefore

$$r_{ab} - r_{ac} \cdot r_{bc} = 0$$

and therefore

$$r_{ab} = r_{ac} \cdot r_{bc}$$

Rule 2: If a direct causal link between two variables is reinforced by an indirect link through a third variable, it is not necessary to control for the third variable in prediction equations where either of the first two variables are the dependent or independent variables, providing that the third variable is not linked to any other variables.

This rule applies to the triangular situation found, for example, between X2,X3, and X4 in the model depicted in Figure 8. The prediction equations, using the rule of thumb, are given by equation (8), viz.

$$r_{13.2} = 0$$

$$r_{14.23} = 0$$

14

Variable $X_3$ is controlled for in the second of these equations, but as $X_3$ simply provides a route for the indirect link between $X_2$ and $X_4$, and is unrelated to any other variables, the second prediction equation may be simplified to

$$r_{14.2} = 0 \qquad (10)$$

The rule could not be applied, however, if $X_3$ was linked directly to $X_1$.

If in doubt, it may be useful to examine the formula for the higher order partial correlation to see if it may be reduced. In this example, the formula for $r_{14.23}$ is given by

$$r_{14.23} = \frac{r_{14.2} - (r_{13.2})(r_{34.2})}{\sqrt{1 - r_{13.2}^2} \sqrt{1 - r_{34.2}^2}}$$

If $r_{14.23}$ is zero, the numerator in this equation must also be zero, therefore

$$r_{14.2} = r_{13.2} \cdot r_{34.2}$$

If the model is correct, we know from the first prediction equation that $r_{13.2}$ is zero, therefore $r_{14.2}$ must also be zero as in equation (10). However, if there was a direct link from $X_1$ to $X_3$, $r_{13.2}$ would no longer be zero and therefore $r_{14.2}$ could not be zero either. The simplifying rule can only be applied if the third variable (in this case $X_3$) is unrelated to the other variables in the system.
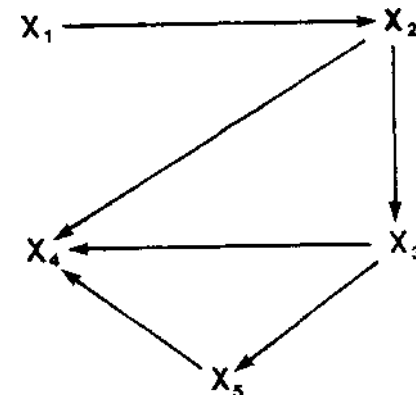


Fig 9. Five variable model with reinforcing links.

It should be noted that if the indirect route passes through two or more variables, which are unrelated to other variables, they may all *be* omitted as control variables. Applying the rule of thumb to the model depicted in Figure 9 (noting that $X_5$ causally precedes $X_4$) we get the predictions:

$$r_{13.2} = 0$$

$$r_{15.23} = 0$$

$$r_{14.235} = 0$$

$$r_{25.13} = 0$$

$$r_{34.125} = 0$$

15

Applying our simplifying rule, the third equation reduces to

$$r_{14.2} = 0$$

Rule 3: When two variables are indirectly connected by a causal chain through two or more intervening variables, but are not otherwise linked, one need only control for one of the intervening variables in the corresponding prediction equation.

Applying the rule of thumb to the model depicted in Figure 10 we get three prediction equations:

$$r_{13.2} = 0$$
$$r_{14.23} = 0$$
$$r_{24.13} = 0 \qquad (11)$$

Applying the second simplifying rule, the second equation may be reduced to

$$\text{or} \quad \begin{array}{l} r_{14.2} = 0 \\ r_{14.3} = 0 \end{array} \qquad (12)$$
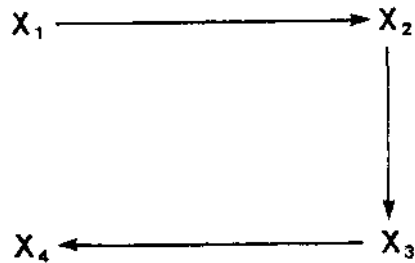


Fig 10. Four variable causal chain.

As before, the justification for this simplification may be verified by examining the formula for the higher order partial correlation:

$$r_{14.23} = \frac{r_{14.2} - (r_{13.2})(r_{34.2})}{\sqrt{1 - r_{13.2}^2}\,\sqrt{1 - r_{34.2}^2}}$$

If $r_{14.23}$ is zero,

$$r_{14.2} = r_{13.2} \cdot r_{34.2}$$

but as $r_{13.2}$ is zero if the model is correct (equation 11), it follows that $r_{14.2}$ must also be zero. A similar argument may be used to show that $r_{14.3}$ must also be zero.

It will be noted that the rule would not have applied if there was a direct link between $X_1$ and $X_3$, or between $X_2$ and $X_4$, because $r_{13.2}$ and $r_{24.3}$ would no longer be zero. The rule may apply in some cases when one or more of the intervening variables are linked to variables other than those before and after in the causal chain, but the researcher needs to be very

16

careful in such circumstances. If in doubt, one should either examine the formula of the higher order partial or simply use the equations given by the rule of thumb.

Rule 4: The zero order correlation between the variables at either end of a causal chain should be equal to the product of the zero order correlations between the variables at either end of each link in the chain. This rule is an extension, under special conditions, of the first simplifying rule.
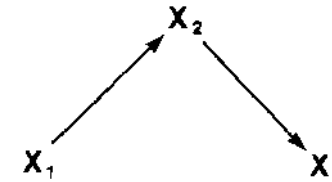


Fig 11. Three variable causal chain.

If we apply this rule to the prediction equation for the model in Figure 11,

$$r_{13.2} = 0$$

we get

$$r_{13} = r_{12} \cdot r_{23} \qquad (13)$$

This may be intuitively interpreted if one considers that the square of a correlation coefficient measures the proportion of variance explained by the independent variable. The proportion of the variance in $X_3$ explained by $X_2$ is $(r_{23})^2$, whereas the proportion of the variance in $X_2$ explained by $X_1$ is $(r_{12})^2$. $X_1$ therefore explains $(r_{12})^2$ of $(r_{23})^2$ of the variance in $X_3$, i.e.

$$(r_{13})^2 = (r_{12})^2 \cdot (r_{23})^2$$

Taking square roots on both sides gives equation (13).

The same logic applies to longer chains. The rule of thumb gave three prediction equations for the model in Figure 10, viz:

$$r_{13.2} = 0$$
$$r_{14.23} = 0$$
$$r_{24.13} = 0$$

Applying the simplifying rule to the second equation:

$$r_{14} = (r_{12})(r_{23})(r_{34})$$

As the first and third equations refer to two link causal chains they may also be simplified to give:

$$r_{13} = (r_{12})(r_{23})$$
$$r_{14} = (r_{12})(r_{23})(r_{34})$$
$$r_{24} = (r_{23})(r_{24})$$

These equations are equivalent to equation (11) if the model is correct.

17

If a variable is linked, either directly or indirectly, to either the dependent or independent variable, but not to both, it may be ignored as a control variable in partial correlation predictions between the dependent and independent variables.

The prediction equations given by the rule of thumb for the model depicted in Figure 10 are:

$$r_{13.2} = 0$$

$$r_{14.23} = 0$$

$$r_{24.13} = 0$$

However, $X_1$ is linked only to $X_2$, therefore it may be ignored as a control variable in the partial correlation between $X_2$ and $X_4$. The third prediction equation therefore reduces to:

$$r_{24.3} = 0$$

This may be verified if we examine the formula for $r_{24.13}$

$$r_{24.13} = \frac{r_{24.3} - (r_{12.3})(r_{14.3})}{\sqrt{1 - r_{12.3}^2} \ \sqrt{1 - r_{14.3}^2}}$$

If $r_{24.13}$ is zero, the numerator must be zero, therefore

$$r_{24.3} = (r_{12.3})(r_{14.3})$$

However, if the model is correct, $r_{14.3}$ is zero (equation 12), therefore

$$r_{24.3} = 0$$

This may be intuitively interpreted as follows. If we examine the model in Figure 10, the third prediction equation (i.e. $r_{24.13} = 0$) is basically saying that we would not expect any of the variance in $X_4$ to be directly explained by $X_2$. We obviously have to control for the indirect effect of $X_2$ on $X_4$ through $X_3$, but $X_1$ only influences $X_2$. All variations in the value of $X_1$ will therefore be subsumed in the variations of the value of $X_2$, and hence they will be taken into account by $r_{24.3}$. There is therefore no need to use the second order partial correlation $r_{24.13}$.
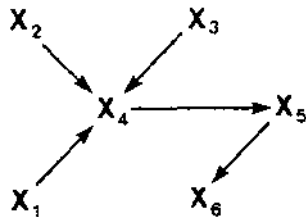


Fig 12. Six variable model I.

The same logic applies even if several variables are linked to the independent variable, but not to the dependent variable. In the model depicted in Figure 12, $X_1$, $X_2$, and $X_3$ are each linked to $X_4$, but not to $X_6$

(except through $X_4$). Thus the prediction equation

$$r_{46.1235} = 0$$

may be reduced to

$$r_{46.5} = 0$$

because $X_5$ is the only variable linked to both the independent variable $(X_4)$ and the dependent variable $(X_6)$. The effects of $X_1$, $X_2$, and $X_3$ are all incorporated in the variations of $X_4$.
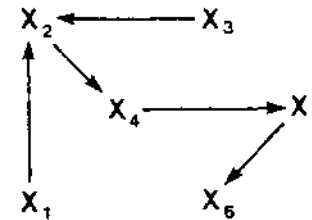


Fig 13. Six variable model II.

The model in Figure 13 is somewhat similar, except that the direct links between $X_1$ and $X_4$, and $X_3$ and $X_4$, have been replaced by indirect links through $X_2$. However, $X_1$, $X_2$, and $X_3$ still only operate on $X_4$, therefore the prediction equation

$$r_{46.1235} = 0$$

may again be reduced to

$$r_{46.5} = 0$$

Some of the other prediction equations for the model in Figure 13 will naturally be different to those for the model in Figure 12 because of the different configuration of links.
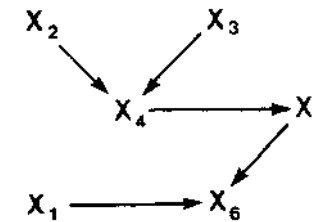


Fig 14. Six variable model III.

It does not matter whether the variables are linked to the independent or dependent variable. Providing that they are linked only to one they may be ignored as controls. Thus, the model in Figure 14 is similar to that in Figure 12, except that $X_1$ is linked now to $X_6$ rather than to $X_4$. The prediction equation

$$r_{46.1235} = 0$$

still reduces to

$$r_{46.5} = 0$$

In the model in Figure 15, however, $X_1$ is linked to both $X_4$ and $X_6$. The prediction equation therefore only reduces to

$$r_{46 \cdot 15} = 0$$
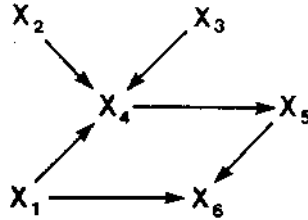


Fig 15. Six variable model IV.

Variables which are linked to an intervening variable may also be ignored as a control variable in certain circumstances. In the model in Figure 16, $X_3$ is linked to $X_5$ but not to $X_4$ or $X_6$. The prediction equation

reduces to
$$r_{46 \cdot 1235} = 0$$
$$r_{46 \cdot 5} = 0$$
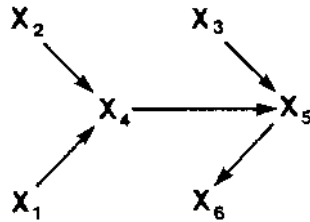


Fig 16. Six variable model V.

Because all of the effects of $X_3$ are subsumed in the variations of $X_5$, it is sufficient to control only for $X_5$. However, if $X_3$ was also linked to $X_6$, as in Figure 17, it would be necessary to control for the direct effects of $X_3$ upon $X_6$. The prediction equation would therefore be

$$r_{46 \cdot 35} = 0$$



Fig 17. Six variable model VI.

20

(iv)  Application of the simplifying rules

In some instances the researcher may have a number of options as to which simplifying rule to use. For example, the rule of thumb gives three prediction equations for the model in Figure 10, viz.

$$r_{13 \cdot 2} = 0$$
$$r_{14 \cdot 23} = 0$$
$$r_{24 \cdot 13} = 0$$

Using rule 3, the second equation may be reduced to

or $\quad \begin{aligned} r_{14 \cdot 2} &= 0 \\ r_{14 \cdot 3} &= 0 \end{aligned}$

whereas using rule 4 it may be reduced to

$$r_{14} = r_{12} \cdot r_{23} \cdot r_{34}$$

In this case the second alternative is probably preferable as it involves only zero order coefficients.
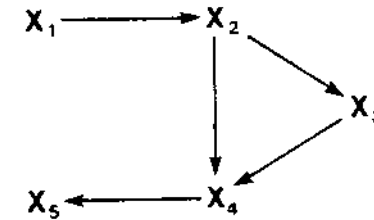


Fig 18. Five variable model to illustrate simplifying rules.

In other instances a prediction equation may be simplified using two or more rules simultaneously. For example, consider the model depicted in Figure 18. Using the rule of thumb we get five prediction equations:

$$r_{13 \cdot 2} = 0$$
$$r_{14 \cdot 23} = 0$$
$$r_{15 \cdot 234} = 0$$
$$r_{25 \cdot 134} = 0$$
$$r_{35 \cdot 124} = 0 \qquad (14)$$

Applying rule 1 the first equation reduces to

$$r_{13} = (r_{12})(r_{23})$$

Applying rules 1 and 2 the second reduces to

$$r_{14} = (r_{12})(r_{24})$$

Applying rules 2 and 4 the third reduces to

$$r_{15} = (r_{12})(r_{24})(r_{45})$$

Applying rules 1, 2, and 5, the fourth equation reduces to

$$r_{25} = (r_{24})(r_{45})$$

21

Applying rule 5, the last equation may be reduced to

$$r_{35.24} = 0$$

However, as $X_2$ only influences variables $X_3$ and $X_4$ this may be further simplified to

$$r_{35.4} = 0$$

and hence, using rule 1, to

$$r_{35} = (r_{34})(r_{45})$$

Thus, the five original prediction equations have been simplified to

$$r_{13} = r_{12}.r_{23}$$
$$r_{14} = r_{12}.r_{24}$$
$$r_{15} = r_{12}.r_{24}.r_{45}$$
$$r_{25} = r_{24}.r_{45}$$
$$r_{35} = r_{34}.r_{45}$$

This may be mathematically confirmed by inspecting the formulae for each of the partial correlations in equation (14).

Given the degree of choice open to the researcher regarding the form of the prediction equations, especially given that most of the simplifications are only valid if the model is correct, it is important to realise that all of the prediction equations must be satisfied. It is not sufficient for, say, four out of five equations to give correct predictions. Unless all of the predictions are correct (subject to possible measurement and sampling errors), the model must be rejected. However, if the model is correct all of the prediction equations will be satisfied irrespective of whether and how they have been simplified.

Further, although some of the simplifying rules assume that the model is correct, **if** the model is incorrect the simplified equations should still give poor predictions. However, if a set of simplified equations give good predictions, there is no harm in testing the equations in their unsimplified form as a double check that the results are not a statistical aberration.

### III EXAMPLES

Two case studies to illustrate the Simon-Blalock technique are detailed below. The first is a rather trivial example to demonstrate the basic principles using artificially created data, whereas the second is a real example using data collected for Belfast in 1971.

(i)   Study using simulated data

To demonstrate that the Simon-Blalock technique is in fact capable of identifying the correct causal model, this example analyses data artifically created with known relationships.

Four variables were created using the equations:

$$X_1 = e_1$$
$$X_2 = 2X_1 + e_2$$
$$X_3 = 0.5X_2 + e_3$$
$$X_4 = 2X_2 + 3X_3 + e_4 \qquad (16)$$

where $e_1$ to $e_4$ are four random series. The values of 30 cases of $X_1$ to $X_4$ are given in Table 2.

| $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|
| 0.6187 | 1.3146 | 1.1827 | 6.7452 |
| 0.9919 | 2.1572 | 1.1896 | 8.3299 |
| 0.3834 | 1.1119 | 1.4932 | 7.2725 |
| 0.3738 | 1.2564 | 1.2529 | 6.6655 |
| 0.7859 | 2.5315 | 1.5783 | 10.0405 |
| 0.3605 | 0.8807 | 0.6936 | 4.7538 |
| 0.0891 | 0.5008 | 0.9566 | 4.1576 |
| 0.2902 | 1.0781 | 1.4974 | 7.1614 |
| 0.9391 | 1.9607 | 1.3744 | 8.5502 |
| 0.0227 | 0.0614 | 0.7703 | 2.8478 |
| 0.6837 | 1.7219 | 1.7514 | 9.6353 |
| 0.8978 | 2.7788 | 2.0758 | 12.6812 |
| 0.2332 | 1.1756 | 0.6919 | 5.3695 |
| 0.3185 | 1.0446 | 0.9696 | 5.5860 |
| 0.8118 | 1.6866 | 1.5898 | 8.1900 |
| 0.0043 | 0.7184 | 0.8126 | 4.8656 |
| 0.7192 | 2.1307 | 1.0668 | 7.9836 |
| 0.2760 | 1.3186 | 1.5869 | 7.6105 |
| 0.1828 | 0.7347 | 0.9206 | 4.8107 |
| 0.6127 | 1.5408 | 1.7413 | 8.8690 |
| 0.0312 | 0.6325 | 1.1616 | 4.9177 |
| 0.6725 | 1.9278 | 1.2981 | 8.6850 |
| 0.7538 | 1.8743 | 1.3338 | 7.8508 |
| 0.4698 | 1.8959 | 1.3193 | 7.9394 |
| 0.0345 | 0.5059 | 0.9115 | 3.9787 |
| 0.9787 | 1.9724 | 1.5952 | 9.4176 |
| 0.5615 | 1.2813 | 1.3679 | 6.6973 |
| 0.5605 | 1.9356 | 1.8502 | 9.4249 |
| 0.3091 | 1.0815 | 1.2895 | 6.7733 |
| 0.8099 | 2.0690 | 1.5856 | 9.3158 |

Table 2. Data used in example **(i)**.

Normally one would not know the relationships between the variables and the objective would therefore be to test various hypothesized causal models to see which is consistent with the data in Table 2. To save space, however, we will only test the known 'correct' model.
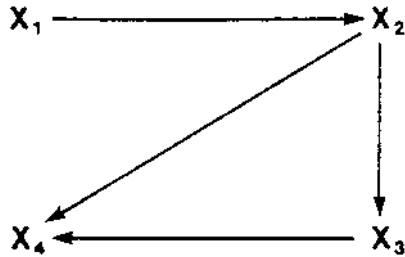
Fig 19. Causal model corresponding to equation (16).

It will be noted in equation (16) that neither $X_3$ nor $X_4$ depend upon $X_1$. The causal system represented by equation (16) is therefore as depicted in Figure 19. Applying the rule of thumb there are two prediction equations, viz.

$$r_{13.2} = 0$$
$$r_{14.23} = 0$$

The prediction equations are comprised of one first order partial correlation and one second order partial correlation. However, to calculate these two partial correlations it is first necessary to calculate all the zero order correlations between the four variables. The zero order product-moment correlation coefficients calculated for each pair of variables in the usual manner are given in Table 3.

|       | $X_1$  | $X_2$  | $X_3$  | $X_4$  |
|-------|--------|--------|--------|--------|
| $X_1$ | 1.0000 | 0.8992 | 0.6221 | 0.8431 |
| $X_2$ | 0.8992 | 1.0000 | 0.6900 | 0.9380 |
| $X_3$ | 0.6221 | 0.6900 | 1.0000 | 0.8790 |
| $X_4$ | 0.8431 | 0.9380 | 0.8790 | 1.0000 |

Table 3. Zero order correlation matrix for example (i).

The formula for the first order partial correlation $r_{13.2}$ is

$$r_{13.2} = \frac{r_{13} - r_{12}.r_{23}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{23}^2}}$$

Substituting the values from Table 3:

$$r_{13.2} = \frac{0.6221 - (0.8992)(0.6900)}{\sqrt{1 - 0.8992^2}\ \sqrt{1 - 0.6900^2}}$$

$$= 0.0052$$

The formula for the second order partial correlation $r_{14.23}$ is

$$r_{14.23} = \frac{r_{14.2} - (r_{13.2})(r_{34.2})}{\sqrt{1 - (r_{13.2})^2}\ \sqrt{1 - (r_{34.2})^2}}$$

We have already calculated the value of $r_{13.2}$, but to calculate $r_{14.23}$ we must first calculate the values of two other first order correlations, viz. $r_{14.2}$ and $r_{34.2}$. These may be calculated in a manner analogous to the calculation of $r_{13.2}$. Substituting the values into the formula for $r_{14.23}$ we find that

$$r_{14.23} = -0.0186$$

The calculated values of $r_{13.2}$ and $r_{14.23}$ are both fairly close to the predicted values, therefore one would be justified in assuming that the hypothesized model is correct.

To avoid having to calculate a second order partial correlation coefficient we could apply the simplifying rules discussed in Chapter II. For example, the sole function of $X_3$ in the model is to serve as an indirect route for the link between $X_2$ and $X_4$, which are also linked directly. Therefore, applying the second simplifying rule, the second prediction equation reduces to

$$r_{14.2} = 0$$

The calculated value in this instance is

$$r_{14.2} = -0.0024$$

If we further apply the first simplifying rule, we need only consider zero order coefficients. The prediction equations reduce to

$$r_{13} = r_{12}.r_{23}$$
$$r_{14} = r_{12}.r_{24}$$

The actual values would again give us reason to assume that the hypothesized model is correct:

$r_{13} = 0.6221$    $r_{12}.r_{23} = (0.8992)(0.6900) = 0.6204$

$r_{14} = 0.8431$    $r_{12}.r_{24} = (0.8992)(0.9380) = 0.8434$

The reader might like, as an exercise, to confirm that alternative hypothesized models do not give as good predictions, using the zero order correlation coefficients given in Table 3.

(ii) Unemployment and religious affiliation in Belfast, 1971

Unemployment is a serious social problem everywhere, but it has been particularly serious in Belfast since the 1920's. To further compound the problem, there is a marked discrepancy in the unemployment rates for different religious affiliations. In 1971, for example, the unemployment rate for Catholics was slightly more than twice that for non-Catholics (mostly Protestants). Due to its political ramifications this fact has been interpreted by different commentators in various ways, but basically two major hypotheses may be identified:

1. Religious discrimination. The most obvious explanation is that unemployment is higher amongst Catholics because of discrimination against Catholics by the predominently Protestant employers. In some versions of this hypothesis discrimination is regarded as a deliberate ploy by Protestants to maintain a higher rate of Catholic outmigration, thereby offsetting the higher Catholic birthrate and maintaining the Protestants as the majority. In other versions discrimination has less sinister connotations and may arise from a 'jobs for the boys' mechanism whereby existing employees use their 'pull' to fix up jobs for their friends (who are usually of the same religious affiliation). In either event there is a direct causal link between religious affiliation and employment.

2. Family size. The second hypothesis is that the higher rate of unemployment amongst Catholics is due to their higher birth-rate. The crude birth-rate for Catholics in Northern Ireland in 1971 was 25.4 per 1000 women, whereas for non-Catholics it was 18.0 (Compton, 1976). This is believed to influence unemployment in various ways. Children in large families, for example, are thought to be more prone to educational deprivation. Also, given that the job market tends to be religiously divided (Barritt and Carter, 1972), proportionately more new jobs are required for Catholics than for Protestants in a period of population increase to prevent the Catholic share of unemployment becoming even worse. In this hypothesis no direct link is envisaged between religious affiliation and unemployment. Rather, the observed association is believed to operate through the intervening variable 'family size'.

To clarify the situation a number of alternative causal models are tested using data originally collected in a study of social malaise in the city (Boal, Doherty and Pringle, 1974; 1978). These data were used to calculate the values of five variables for each of 97 subdivisions of the Belfast Urban Area:

1. Unemployment (U) : percentage of adult males unemployed.
2. Religious affiliation (R) : percentage Catholic.
3. Family size (F) : percentage of households with more than six people.
4. Social status (S) : percentage of households in Social Class V
   (i.e. unskilled manual workers).
5. Persons per room (P).

The zero order correlations are given in Table 4.

| | | U | R | F | S | P |
|---|---|---|---|---|---|---|
| Unemployment | (U) | 1.000 | 0.504 | 0.788 | 0.753 | 0.693 |
| Religious Affiliation | (R) | 0.504 | 1.000 | 0.627 | 0.304 | 0.468 |
| Family Size | (F) | 0.788 | 0.627 | 1.000 | 0.584 | 0.764 |
| Social Status | (S) | 0.753 | 0.304 | 0.584 | 1.000 | 0.554 |
| Persons Per Room | (P) | 0.693 | 0.468 | 0.764 | 0.554 | 1.000 |

Table 4. Zero order correlation matrix for Belfast data.

To simplify matters let us begin by considering only the first three variables. Three models are shown in Figure 20. The first two models are consistent with the religious discrimination hypothesis (i.e. a direct link is hypothesized between religious affiliation and unemployment). The prediction equation arrived at by applying the rule of thumb is indicated under each diagram and the calculated value is given in parentheses. In both instances the prediction equations are seen to be highly inaccurate.



$$r_{FU.R} = 0 \,(0.70) \qquad r_{RF} = 0 \,(0.63) \qquad r_{RU.F} = 0 \,(0.02)$$
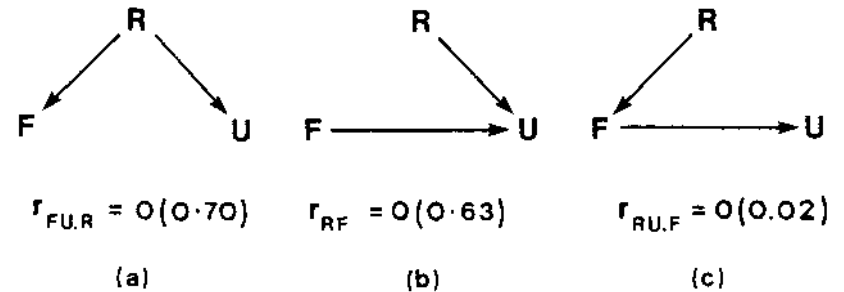
(a)       (b)       (c)

Fig 20. Three variable models of unemployment in Belfast.

The third model in Figure 20 is in agreement with the family size hypothesis. It is hypothesized that religious affiliation influences family size, possibly due to attitudes on contraceptives, and family size influences the chances of unemployment. The prediction equation is found to be extremely accurate, indicating that the model is in high agreement with the empirical data. Likewise, if one applies the first simplifying rule the prediction equation reduces to

$$r_{RU} = r_{RF} \cdot r_{FU}$$

Substituting in the values from Table 4:

$$r_{RU} = 0.504 \qquad r_{RF} \cdot r_{FU} = (0.627)(0.788) = 0.494$$

If it is accepted that it is highly unlikely that religious affiliation is determined by either family size or unemployment, no other three variable model gives as good predictions.

If we expand the analysis to four variables by including social status, the best four variable model is that shown in Figure 21. This retains the basic structure of the family size hypothesis (i.e. Figure 20(c)), but it also hypothesizes that low social status is also a function of family size and that this in turn reinforces unemployment. The two prediction equations derived by the rule of thumb are
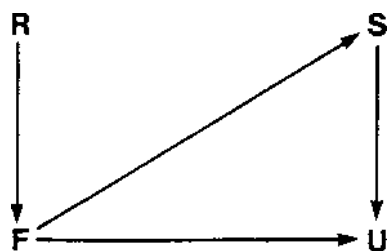
$$r_{RS.F} = 0$$
$$r_{RU.FS} = 0$$

$$r_{RS.F} = 0 \, (-0.09)$$

$$r_{RU.SF} = 0 \, (0.09)$$

Fig 21. Four variable model of unemployment in Belfast.

Both equations give fairly good predictions, as can be seen from the calculated values shown in parentheses in Figure 21, but the predictions are not quite as accurate as those found in the three variable analysis.

The two prediction equations can be simplified. Applying rule 2 the second equation reduces to

$$r_{RU.F} = 0$$

The calculated value of $r_{RU.F}$ is 0.02. Both equations may be further reduced by applying rule 1 to give

$$r_{RS} = r_{RF} \cdot r_{SF}$$
$$r_{RU} = r_{RF} \cdot r_{UF}$$

The second equation has already been tested above. Substituting the appropriate zero order correlation coefficients into the first we get:

$$r_{RS} = 0.304 \qquad r_{RF} \cdot r_{SF} = (0.468)(0.554) = 0.259$$

The results using the simplified equations therefore largely confirm the results using the rule of thumb equations.

If we incorporate the fifth variable (persons per room), the best five variable model yet found by the author is that shown in Figure 22. This retains the basic structure of the four variable model in Figure 21. The three prediction equations derived by the rule of thumb, together with the calculated values in parentheses, are indicated in Figure 22. In each case the prediction is found to be reasonably accurate. The reader may like to confirm that the simplified versions of these equations also yield accurate predictions.
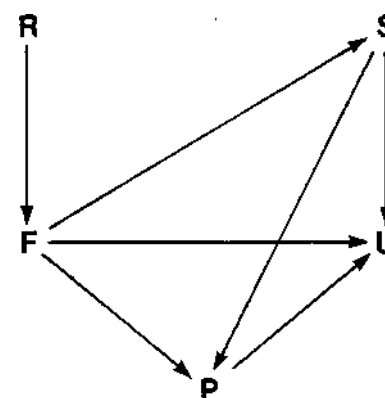


$$r_{RS.F} = 0 \, (-0.09)$$

$$r_{RU.SFP} = 0 \, (0.09)$$

$$r_{RP.FS} = 0 \, (0.00)$$

Fig 22. Five variable model of unemployment in Belfast.

It will be noted that the model shown in Figure 22 hypothesizes that religious affiliation is only related to the other variables in the model through its effects upon family size. In other words, there is no link corresponding to religious discrimination. This is not to say that religious discrimination does not exist - numerous case studies have shown that it does. Rather, the models indicate that religious discrimination is not a factor in explaining religious differences in unemployment. It may be, for example, that religious discrimination by Protestants against Catholics is to some extent counterbalanced by religious discrimination by Catholics against Protestants.

This case study does not prove that the higher rate of unemployment amongst Catholics can be explained by the larger average size of Catholic families, but it does indicate that this contention is supported by empirical data when the correlations between the various variables are taken into consideration, whereas the discrimination hypothesis is not supported. Family size has also been found to be an important correlate o = unemployment in studies elsewhere. McGregor (1978), for example, found a high correlation between family size and unemployment in Paisley, Scotland. This suggests that it may be 'normal' for the adverse effects of unemployment to be more common amongst larger families, and that Belfast is no exception in this

respect. However, given that the larger families in Belfast are often Catholic, this effect is reflected by a higher rate of unemployment amongst Catholics. This case study suggests that discrimination is a relatively minor factor in explaining the distribution of unemployment in Belfast, and that the key to a fuller understanding lies in examining the role of family size as a factor limiting opportunity in a market economy.


## IV ASSUMPTIONS AND LIMITATIONS

The Simon-Blalock technique, like all quantitative techniques, entails a number of implicit assumptions. These are made explicit below. Also, the Simon-Blalock technique has a number of limitations. These are listed to act as a cautionary warning in the interpretation of results. Some assumptions and limitations have already been mentioned in the course of the text, but they are restated here for completeness.

(i)  Assumptions

The first four assumptions arise because the Simon-Blalock technique is based upon the linear regression model. Readers who are unfamiliar with this model are advised to consult Ferguson (1977) or Johnston (1978) for a more detailed explanation of these assumptions.

1.  Linearity. All relationships are assumed to be linear, or at least approximately linear, i.e. a scattergram of one variable against another should approximate a straight line. Strictly speaking the linearity refers to the parameters, therefore it may be possible to convert a non-linear relationship into a linear one using data transformations.

2.  Additivity. The effects of independent variables upon a dependent variable are assumed to be additive.

3.  The mean value of the disturbance terms for each variable in the model (i.e. the e terms in equation (4)) should be zero. This assumption cannot be tested, but it is required in order to use OLS to solve for the parameters.

4.  Each of the disturbance terms is assumed to be uncorrelated with each of the variables in the model.

The next four assumptions also arise because of the linear regression model, but are only necessary if population parameters are to be estimated from sample data.

5.  The disturbance terms are assumed to be randomly drawn from a probability distribution with a mean of zero for every set of values of the independent variables. This assumption will be violated if an important explanatory variable has been overlooked in the model.

6.  Homoscedasticity. The variance of the probability distribution of the disturbance terms should be constant for all values of the independent variables.

7.  Serial independence. Values of the disturbance terms for each variable should be uncorrelated with other values for the same variable. This assumption will be infringed if there is a high degree of spatial or temporal autocorrelation in the data, but this is less likely if no important explanatory variables have been omitted in the model.

8.  A low degree of multicollinearity. Multicollinearity is said to exist if independent variables in a multiple regression are correlated. This is always the case, otherwise there would be no need to use multiple regression (Ferguson, 1977, 28), and therefore it is strictly speaking incorrect to refer to the independent variables as being 'independent' (although the terminology is retained here because of its widespread use elsewhere). Multicollinearity only becomes a serious problem if the degree of correlation is very high (say, 0.8 or higher) as it then results in unstable estimates of the population regression coefficients. However, high correlations suggest that the two variables may be essentially measures of the same thing, in which case it may be possible to simply omit one for causal modelling purposes.

The next two assumptions arise because the Simon-Blalock technique is based upon a set of regression equations, rather than simply one multiple regression equation.

9.  No feedback loops. If variable X is a cause of variable Y, it is assumed that variable Y is not a cause of variable X, either directly or indirectly. This means that all causal links must be unidirectional (rather than reciprocal) and that it should be possible to list all the variables in a model in such a way that no variable is a cause of one preceding it in the list.

10. The disturbance term for each variable in the model should be uncorrelated with the disturbance terms for all other variables in the model. This is likely to be violated if an important causal factor has been omitted by the model.

Finally, two more obvious assumptions should be added:

11. Data should be measured on an interval or ratio scale, otherwise Pearson's r cannot be calculated.

12. The variables should be measured without error. In regression measurement errors in the dependent variable can be tolerated if it can be assumed that they are subsumed by the disturbance term. In the Simon-Blalock technique most of the dependent variables are also independent variables in other equations, therefore measurement errors are more serious.

It might also be noted that normal distributions are not essential to the Simon-Blalock technique. These are only required in        regression if significance tests are to be applied, but, as noted below, significance tests cannot be applied in the Simon-Blalock approach.

Apart from the various limitations imposed by the assumptions (e.g. the restriction that models may not involve feedback loops) a number of other limitations should be recognised.

1. The Simon-Blalock technique may only be used to test models in which there is at least one missing link, because models with no missing links do not provide any prediction equations. Conversely, as there is one prediction equation for each missing link, models with a large number of missing links can be tested with more confidence than more complex models with fewer direct links missing.

2. The technique becomes extremely cumbersome once more than five or six variables are considered. As the number of variables is increased the number of possible models increases very rapidly in accordance with equation (2). Also, the prediction equations rapidly increase in number, and, unless the simplifying rules are used, they are of a higher order making verification more tedious. Computer programs may circumvent some of these problems, but it is recommended that models should be kept as simple as possible, at least in the initial stages of an inquiry.

3. Having identified what appears to be the 'best' model, there are no tests of significance. One can test whether a given correlation coefficient is significantly different from zero at the 1-a confidence level (i.e. whether there is a probability of less than a that the observed correlation could be so high just by chance if the true value is zero), but there is no way of testing whether a correlation coefficient is significantly zero. One could assume in the absence of any evidence that it is significantly non-zero that it was in fact zero, but this would only serve as an approximate guide. Besides, the objective is to find the best model rather than to find one which is 'significant' so significance testing is not really required. The decision as to what value for a correlation coefficient might be regarded as being close enough to zero, given measurement and possible sampling errors, is basically intuitive, but depends upon the number of cases. For a sample of 100 cases, a maximum of 0.1 might be regarded as a possible (if somewhat lax) rule of thumb.

4. Different models sometimes produce the same prediction equations. This means that even if the researcher hypothesizes a model that is consistent with the data, there may be other models which are equally consistent. For example, the two three variable models depicted in Figure 2 both produce the same prediction equation (viz. $r_{XY.Z} = 0$), but in Figure 2(a) the relationship between X and Y is indirect via Z, whereas in Figure 2(b) it is spurious. Simon (1954) discusses ways in which these and other situations may be distinguished, while Blalock (1964) notes that in general models giving similar predictions may be distinguished by the addition of an extra variable. However, it is important to realise that even if a model's prediction equations are satisfied, it does not necessarily mean that the hypothesized model is correct. Rather, the main strength of the technique lies in indicating which hypothesized models are wrong.

5. The Simon-Blalock technique suffers the same limitations as all correlation and regression methods in geography. One of the most important of these is the problem of ecological inference (Robinson, 1950). This arises if data for spatial units, or other aggregates, are used to make inferences about the behaviour of individuals. It may be that relationships observed at a spatial level do not apply at the individual level, therefore special care must be taken when interpreting the results. The example of unemployment in Belfast discussed in the previous chapter is a case in point. The data here referred to spatial subdivisions of the Belfast Urban Area, hence the models discussed are only valid at the ecological (i.e. spatial) level. Thus, while no evidence could be found to support the hypothesis that the spatial distribution of unemployment is caused by religious discrimination, there may be discrimination at the individual level.

Similarly, conclusions found at one scale level need not necessarily be replicated if a different set of spatial units are used as the data collection units.

## V OTHER CAUSAL MODELLING TECHNIQUES

The Simon-Blalock technique is only one of a number of causal modelling techniques. Some of the others are briefly reviewed to place the Simon-Blalock model in context.

(i) <u>Path analysis</u>

Path analysis was first developed by the geneticist Sewall Wright in the 1920s, but it has recently become quite popular in sociology (e.g. Duncan, 1966; Pickvance, 1974). A sociological technique known as <u>dependence analysis</u> (Boudon, 1965) is also closely related.

Path analysis is similar in many respects to the Simon-Blalock technique. Both, for example, utilise causal diagrams, but standardized partial regression coefficients (known as path coefficients) are preferred in path analysis to partial correlation coefficients. A path coefficient can be calculated for each direct causal link in a model by OLS regression. Each path coefficient provides a quantitative measure of the relative strength of the causal link to which it refers, consequently the technique is often used to <u>calibrate</u> an accepted causal model rather than to test a hypothesized model. For example, having found the most acceptable model using the Simon-Blalock approach, one could use path analysis to quantify the relative strengths of the direct and indirect effects of one variable upon another. Path analysis can also be used to test hypothesized models, but the Simon-Blalock technique is probably simpler     this purpose.

Path analysis differs from the Simon-Blalock approach in several minor respects. For example, the effects of unknown (i.e. unmeasured) variables are always explicitly shown on a path diagram. Also, variables in a path model may be correlated without being causally linked, either directly or indirectly, and are represented in a path diagram by a curved double headed arrow. By and large, however, path analysis has much in common with the Simon-Blalock approach, and both make similar assumptions, for example, About one way causality and uncorrelated disturbance terms.

Good expositions of path analysis are provided by Wright (1934), Kerlinger and Pedhazur (1973) and Asher (1976).

## (ii) Simultaneous equation techniques

In more complex causal situations one may be forced to abandon the assumption of one way causality, and the model may consequently incorporate reciprocal causal links or circular links. If a regression equation is written for each variable, it will be found that independent equations in some equations will depend elsewhere upon the dependent variable. This means that the disturbance terms cannot possibly be uncorrelated, and therefore the model is underidentified with the result that OLS techniques cannot be used to estimate the regression parameters. Fortunately alternative techniques, including two stage least squares (2SLS), have been developed in recent years for such purposes, mainly by econometricians. Nonrecursive models are discussed by Asher (1976) in some detail, whereas Todd (1979) provides an excellent introduction to 2SLS.

## (iii) Causal inferences from dichotomous variables

It has been assumed in this monograph that the data are measured on the interval or ratio scale, but alternative techniques may be used for data on a lower scale level. If they are measured on the ordinal scale, Kendall's tau, which can be partialled (Siegel, 1956), can be used instead of Pearson's r. Even if the data are measured on a nominal scale they may still be used for causal analysis, providing that they are dichotomous, using Yule's Q as the coefficient of association. In fact, as Davidson (1976) explains, quite a high degree of interpretation is possible by comparing the values of partial coefficients with differentials (measures which do not appear to have any equivalents at higher scale levels).

## VI SUMMARY

The Simon-Blalock technique may be used to test if a hypothesized causal model is consistent with empirical data. It is not possible to test the significance of the degree of correspondence between a model and the data, hence the researcher must use his/her discretion as to whether the degree of correspondence is sufficient. Even if it is, this does not prove that the model is necessarily correct, but simply that there is no evidence to suggest that it is wrong. The technique is therefore used to best effect within a framework whereby the correct model is envisaged as being revealed by a process of elimination of the incorrect alternatives (i.e. models which are not consistent with the empirical data).

The Simon-Blalock technique is a relatively simple technique. More complex techniques may be used if the objective is to calibrate an accepted model or to test a more complex model. However, the Simon-Blalock technique is particularly suited to the earlier stages of model testing and, as it takes more inter-relationships into account than multiple regression, it is a technique which would appear to be particularly useful in geography at its current stage of development.

## VII BIBLIOGRAPHY

The Simon-Blalock Technique

Blalock, H.M., (1962) Four variable causal models and partial correlation. *American Journal of Sociology,* 68, 182-194.

Blalock, H.M., *(1964)Causal inferences in non-experimental research.* (University of North Carolina Press, Chapel Hill).

Simon, H., (1954) Spurious correlation : a causal interpretation. *Journal of American Statistical Association,* 49, 467-479.

Simon, H., (1957) *Models of man.* (John Wiley, New York).

Related Techniques

Asher, H.B., (1976) *Causal Modelling.* Sage university paper on quantitative applications in the social sciences, 3, Beverly Hills.

Blalock, H.M., (1967) Causal inferences, closed populations, and measures of association. *American Political Science Review,* 61, 130-136.

Blalock, H.M., ed. (1971) *Causal models in the social sciences.* (Macmillan, London).

Boudon, R. (1965) A method of linear causal analysis : dependence analysis. *American Sociological Review,* 30, 365-374.

Davidson, *N., (1976)Causal inferences from dichotomous variables.* Concepts and techniques in modern geography, 9, (Geo Abstracts Ltd. Norwich).

Duncan, O.D., (1966) Path analysis : sociological examples. *American Journal of Sociology,* 72, 1-16.

Kerlinger, F.N. & Pedhazur, E.J., (1973) Multiple *regression in behavioral research.* (Holt, Rinehardt and Wilson, New York).

Todd, D., (1979) *An introduction to the use of simultaneous-equation regression analysis in geography.* Concepts and techniques in modern geography, 21, (Geo Abstracts Ltd. Norwich).

Turner, M.E. & Stevens, C.D., (1959) The regression analysis of causal paths. *Biometrics,* 15, 236-258.

Wright, S., (1934) The method of path coefficients. Annals *of Mathematical Statistics,* 5, 161-215.

Wright, S., (1960) Path coefficients and path regressions : alternative or complementary concepts. *Biometrics,* 16, 189-202.

Other Technical

Ferguson, R., *(1977) Linear regression in geography.* Concepts and techniques in modern geography, (Geo Abstracts Ltd. Norwich).

Johnston, R.J., (1978) *Multivariate statistical analysis in geography.* (Longman, London).

Nie, N.H. et. al., (1975) *Statistical package for the social sciences.* 2nd edition. (McGraw-Hill, New York).

Robinson, W.S., (1950) Ecological correlations and the behavior of individuals. *American Sociological Review,* 15, 351-357.

Siegel, S., (1956) *Non-parametric statistics for the behavioral sciences.* (McGraw-Hill, New York).

Applications

Capecchi, V. & Galli, G., (1969) Determinants of voting behavior in Italy : a linear causal model analysis. In: Dogan M. & Rokkan S. (ed.) *Quantitative ecological analysis in the social sciences.* (M.I.T. Press, Cambridge, Mass.).

Cox, K.R., (1968) Suburbia and voting behaviour in the London Metropolitan Area. *Annals Association American Geographers,* 58, 111-127.

Cox, K.R., (1969) Comments in reply to Kasperson and Taylor. *Annals Association American Geographers,* 59, 411-415.

Goldberg, A.S., (1966) Discerning a causal pattern among data on voting behavior. *American Political Science Review,* 60, 913-922.

Kasperson, R.E., (1969) On suburbia and voting behavior. *Annals Association American Geographers,* 59, 405-411.

Mercer, J., (1975) Metropolitan housing quality and an application of causal modelling. *Geographical Analysis,* 7, 295-302.

Pickvance, C.G., (1974) Life cycle, housing tenure, and residential mobility : a path analytic approach. *Urban Studies,* 11, 171-188.

Taylor, P.J., (1969) Causal models in geographic research. *Annals Association American Geographers,* 59, 402-404.

Wilbur, G.L., (1964) Growth of metropolitan areas in the South. social *Forces,* 42, 489-499.

Belfast Case Study

Barritt, D.P. & Carter, C.F., (1972) *The Northern Ireland Problem.* 2nd edition. (Oxford University Press, London).

Boal, F.W., Doherty, P. & Pringle, D.G., (1974) *The spatial distribution of some social problems in the Belfast Urban Area.* (Northern Ireland Community Relations Commission, Belfast).

Boal, F.W., Doherty, P. & Pringle, D.G., (1978) Social Problems in the Belfast Urban Area : an exploratory analysis. *occasional Paper Series,* No. 12, Dept. of Geography, Queen Mary College, London.

Compton, P.A., (1976) Religion and population in Northern Ireland. *Transactions Institute British Geographers,* 1, 433-452.

McGregor, A., (1978) Family size and unemployment in a multiply deprived urban area. *Regional Studies,* 12, 323-330.