WORLD FERTILITY SURVEY

TECHNICAL BULLETINS

MARCH 1977

NO. 2/TECH. 414

Path Analysis and Model Building

M. G. KENDALL
AND
C. A. O'MUIRCHEARTAIGH

INTERNATIONAL STATISTICAL INSTITUTE
Permanent Office · Director: E. Lunenberg
Prinses Beatrixlaan 428
Voorburg, The Hague,
Netherlands

WORLD FERTILITY SURVEY
Project Director:
Sir Maurice Kendall, Sc. D., F.B.A.
35-37 Grosvenor Gardens
London SW1W OBS, U.K.

The World Fertility Survey is an international research programme whose purpose is to assess the current state of human fertility throughout the world. This is being done principally through promoting and supporting nationally representative, internationally comparable, and scientifically designed and conducted sample surveys of fertility behaviour in as many countries as possible.

The WFS is being undertaken, with the collaboration of the United Nations, by the International Statistical Institute in cooperation with the International Union for the Scientific Study of Population. Financial support is provided principally by the United Nations Fund for Population Activities and the United States Agency for International Development.

This paper is one of a series of Technical Bulletins recommended by the WFS Technical Advisory Committee to supplement the document *Strategies for the Analysis of WFS Data* and which deal with specific methodological problems of analysis beyond the Country Report No. 1. Their circulation is restricted to people involved in the analysis of WFS data, to the WFS depositary libraries and to certain other libraries. Further information and a list of these libraries may be obtained by writing to the Information Office, International Statistical Institute, 428 Prinses Beatrixlaan, Voorburg, The Hague, Netherlands.

PATH ANALYSIS AND MODEL BUILDING

WFS/TECH.414

By:

M.G. KENDALL and
C.A. O'MUIRCHEARTAIGH

ERRATA

Technical Bulletin No. 2: *Path Analysis and Model Building*

p. 19

First diagram:

p_{3u} should read 0.85 instead of 0.69

p_{4v} should read 0.96 instead of 0.84

Second diagram:

p_{3u} should read 0.85 instead of 0.69

p_{4v} should read 0.86 instead of 0.70

p_{43} should read -0.14 instead of 0.14

p_{03} should read -0.06 instead of 0.06

p. 22:

p_{03} should read -0.27 instead of 0.27

CONTENTS

1. INTRODUCTION

In a broad sense, all scientific analysis is model building. The scientist, whether in the physical or the social sciences, attempts to summarize the complexity of the phenomenal world in the form of simplified statements, laws, hypotheses or models; and for two main reasons, to understand and to control.

A *model* in this sense is not usually a physical replica of the system under study, though even this is possible, as when a newly designed aircraft is partially tested by observing the behaviour of a model in a wind-tunnel. More generally, by a *model* is meant a description of the relationships connecting the variables of interest: the rules of the game. The process of model-building consists of putting together a set of formal expressions of these relationships to the point when the behaviour of the model adequately mimics the behaviour of the system.

Models are built for specific purposes and do not necessarily attempt to prescribe in detail every facet of the system. In fact, the utility of some models resides as much in what they omit on grounds of irrelevance as in what is retained: an arterial road map serves its purpose by ignoring minor streets and unimportant topographical detail; and a model income-distribution may, for some purposes, merely present the actual distribution without concerning itself with the countless circumstances which determine the income of a particular individual person.

One of the basic problems in statistical analysis is the specification of the model to be used, i.e., the mathematical form of the population from which the data are regarded as a sample. The problem is that of drawing inferences from the probability distribution of the observed variables to the underlying structure which generated this observed distribution. In the social sciences, the observed data typically come from non-experimental situations; in the absence of experimental controls, statistical procedures must provide a substitute.

The models proposed frequently contain latent variables which, while not directly observed, have implications for relationships between observable variables. In the document *Strategies for the Analysis of WFS Data*[*] a distinction is drawn between *explanatory* and *intermediate* variables. The basic idea is that the explanatory variables, though more easily observable in general, are causally more remote from fertility than the intermediate variables and operate through them. Intermediate variables such as frequency of intercourse, contraceptive practice, and periods of lactation may be regarded as having a direct causal effect on conception. They are, in one sense, more explanatory than the explanatory variables. The latter, for example educational status and income, are considered as having an influence on fertility and may be described as causal; but their influence is brought to bear, in the main, through the effect which they themselves have on the intermediate variables. Some of the intermediate variables, such as age at marriage, may also be observed directly and may be included as predictor or explanatory variables in some of the equations.

The models are also generally built up of several equations or submodels which interact together and must be considered simultaneously. This interdepence of the relationships between the variables is the source of many of the difficulties which arise in attempting to describe a data set adequately using conventional statistical methods.

For the most part (though this is not laid down as an inviolable rule) fertility models seem to be most useful if they relate fertility itself, however measured, to 'explanatory' variables, verify or disprove hypotheses concerning that relationship and, if possible, quantify the contribution of particular variables to fertility behaviour. A simple example may illustrate some of the difficulties. Suppose we are interested in the relationship between fertility (y) (measured by number of children born) and the two variables age (x_1) and number of years of education (x_2). To simplify the exposition, we suppose that for each member of the sample satisfactory values of the x's and y are determined.

Once a dependent variable has been selected, the researcher identifies a set of variables which are related to it, in the sense that a change in the

[*] WFS Basic Documentation Series, No.9 (The Hague: International Statistical Institute, 1977)

value of one of these variables is believed to result in a change in the value of the dependent variable.

These other variables may be classified as predictors or controls. Some variables, such as level of education may be considered in either role. Certain other variables are consistently treated by demographers as controls. When the dependent variable is a measure of fertility, age is the best example of a control variable. The manner in which fertility varies with the age of a woman is well known, and thus we would not regard age as a variable which in itself advanced our understanding of fertility. Cumulative fertility, for example, increases monotonically with age by virtue of its definition.

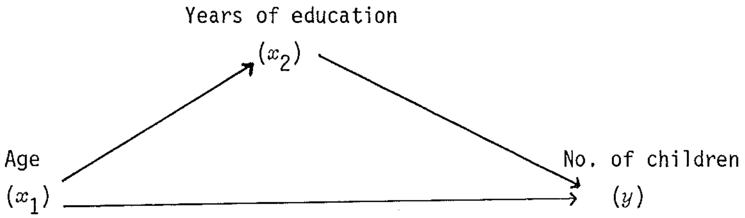
Nevertheless, it is not sensible to exclude age from an analysis of fertility. One of the following strategies should be adopted (a) the population should be subdivided into cohorts, or age groups, and each one analyzed separately; (b) the dependent variable should be defined in such a way that the impact of age is removed; or (c) age should be included explicitly as a control variable.

In the context of the path models described below, the third course is followed by including age as a variable in the equations in the model and thus by removing the effect of age from the other relationships estimated in the model. In the estimation, therefore, age is treated as an explanatory variable, though the purpose of its inclusion is largely that of control. In Section 6, some indication is given of the usefulness of cohort analysis.

The next question is whether we need to write into the model the intermediate variables. A decision depends on what we are trying to do with the model. From one point of view the explanatory variables can be regarded as stimuli evoking a response y through some mechanism which is not of immediate interest. The situation is then one of a *black box* linking stimulus and response and we are not concerned with how the relationships inside the black box operate. (Sooner or later we may want to examine them and take the lid off the box, but not for the present).

In such a case the causal links between x's and y are set up as direct, although it is recognized that the causality is more circuitous. There are a number of relationships which we may wish the model to contain. First, we expect that age will influence fertility directly (perhaps through some intermediate variables not included in the model). We would also expect that years of education will affect fertility directly. We can rule out the possibility that years of education will affect age, but we may postulate that age will be related to length of formal education and through this will affect fertility indirectly.

We represent this model diagrammatically below, using one-way arrows leading from each explanatory variable to each variable which it influences directly



EXAMPLE 1

We assume that the relationships are linear (or that the variables have been converted into a suitable form to justify linearity, as to which see Technical Bulletin No.1*). The relationships may be written as

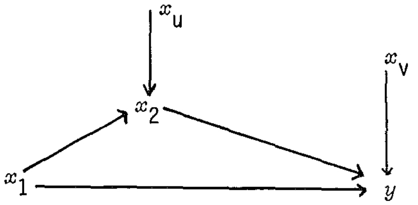
$$\begin{aligned} x_2 &= \beta_{21}\, x_1 \\ y &= \beta_{01}\, x_1 + \beta_{02}\, x_2 \end{aligned} \qquad (1.1)$$

These equations are not, however, exact in practice and it is necessary to make some allowance for departure from exactitude. This is usually done, as in the regression case, by adding a term on the right. But this is not necessarily a random variable. It stands for something we have purposely, or accidentally, omitted from the model but which we hope is not serious enough to impair the approximate representation provided by it. It is

*Sir Maurice Kendall, *Some Notes on Statistical Problems Likely to Arise in the Analysis of WFS Surveys* (The Hague: International Statistical Institute, 1976) not a standard disturbance term but represents a variety of unmeasured sources of variation. But to make any progress with the estimation we still need to assume something about this term. What we shall assume is that it has a mean value of zero and is uncorrelated with any of the immediate determinants of the dependent variable to which it pertains. The equations may now be written as

$$\begin{aligned} x_2 &= \beta_{21}\, x_1 + \beta_{2u}\, x_u \\ y &= \beta_{01}\, x_1 + \beta_{02}\, x_2 + \beta_{0v}\, x_v \end{aligned} \qquad (1.2)$$

The diagram may be modified to include the residual terms



Without loss of generality, we assume that all the variables are standardized to zero mean and unit variance. Conventionally, the coefficients in the equations with standardized variables are called *path coefficients* and are written as p_{ij} where the first subscript identifies the dependent variable, the second the variable whose direct effect on the variable is measured by the path coefficient. The system can therefore be written as

$$\begin{aligned} x_2 &= p_{21}\, x_1 + p_{2u}\, x_u \\ y &= p_{01}\, x_1 + p_{02}\, x_2 + p_{0v}\, x_v \end{aligned} \qquad (1.3)$$

This system is a recursive system - in other words, there are no feedback loops in the system whereby x_i can influence itself. In this Technical Bulletin we shall not consider models which include a direct or indirect feedback loop.

The estimation and interpretation of models such as (1.2) and (1.3) is called path analysis. The original formulation of the method by Wright (1921) was in terms of the decomposition of correlation coefficients. The alternative formulation in terms of regression analysis has however some advantages. The two methods are presented in the following sections.

2, ESTIMATION AND INTERPRETATION OF PATH COEFFICIENTS

The first method involves the use of the observed zero order correlations between the variables in the system together with the specified relationships between the variables in order to estimate the path coefficients. This method is described more fully in Duncan (1966). When dealing with sample data, the assumed zero correlations in the population between the disturbance terms and causally prior variables will not hold exactly. However, as part of the estimation procedure, the fact that the expected value of these correlations is zero is used in order to derive unbiased estimators of the coefficients.

The second method consists of applying ordinary least squares regression to the equations in the system one by one. If the variables are standardized (transformed to zero mean and unit variance), the estimates obtained are identical to those obtained by the first method.

The regression method of estimation is in general preferable on two counts. First, the fact that we are dealing with sample data is recognized more explicitly. Second, the regression estimation procedure provides automatically estimates of the precision of the coefficients and a framework in which hypotheses concerning the coefficients may be tested. The path approach does, however, provide an intuitively more appealing orientation and the diagrammatic representation makes the substantive assumptions in the model more apparent. Furthermore, as we show later, overidentification of the model is easier to detect when the full set of simultaneous equations is written down explicitly.

2.1 DECOMPOSITION OF CORRELATION COEFFICIENTS

Since the variables are standardized, the correlation coefficient r_{ij} can be written as

$$r_{ij} = \frac{1}{n} \Sigma\, x_i x_j .$$

Thus, from (1.3)

$$\begin{aligned} r_{12} = \frac{1}{n} \Sigma\, x_1 x_2 &= \frac{1}{n} \Sigma\, x_1\, (p_{21}\, x_1 + p_{2u}\, x_u) \\ &= p_{21} + 0 \text{ since } \frac{1}{n} \Sigma\, x_1^2 = 1 \qquad (2.1) \\ &\qquad \text{and } x_u \text{ is uncorrelated with } x_i . \end{aligned}$$

Similarly

$$\begin{aligned} r_{01} &= \frac{1}{n} \Sigma\, x_1 y = \frac{1}{n} \Sigma\, x_1\, (p_{01}\, x_1 + p_{02}\, x_2 + p_{0v}\, x_v) \\ &= p_{01} + p_{02}^{*}\, r_{12} \end{aligned} \qquad (2.2)$$

and

$$\begin{aligned} r_{02} &= \frac{1}{n} \Sigma\, x_2 y = \frac{1}{n} \Sigma\, x_2\, (p_{01}\, x_1 + p_{02}\, x_2 + p_{0v}\, x_v) \\ &= p_{01}\, r_{12} + p_{02} \end{aligned} \qquad (2.3)$$

Equations (2.2) and (2.3) enable us to solve for p_{01} and p_{02} in terms of r_{01}, r_{02} and r_{12}, giving

$$p_{01} = \frac{r_{01} - r_{02}\, r_{12}}{1 - r_{12}^2}$$

$$p_{02} = \frac{r_{02} - r_{01}\, r_{12}}{1 - r_{12}^2}$$

Thus, from (2.1), (2.2) and (2.3), the path coefficients p_{01}, p_{02} and p_{21} can be obtained directly from the correlation coefficients.

*The subscript 0 is used to denote the variable y.

With this simple model, the data from the Fiji Fertility Survey, with 4928 respondents, provide the following values for the correlation coefficients

r_{01} = 0.64 {correlation between age and number of children}

r_{02} =-0.34 { " " years of education and number of children}

r_{12} =-0.32 { " " age and years of education}

Substituting these values in (2.1), (2.2) and (2.3) gives

p_{21} = -0.32

p_{01} = 0.59

p_{02} = -0.15

The residual paths can be obtained simply by using:

$$r_{22} = 1 = \frac{1}{n} \Sigma x_2^2 = \frac{1}{n} \Sigma x_2 (p_{21} x_1 + p_{2u} x_u)$$
$$= p^2_{21} + p^2_{2u}$$

Hence

$$p^2_{2u} = (1 - p^2_{21})$$

i.e.,

$$p_{2u} = \overline{(1 - p^2_{21})} \qquad (2.4)$$

and

$$r_{00} = 1 = \frac{1}{n} \Sigma y^2 = \frac{1}{n} \Sigma y (p_{01} x_1 + p_{02} x_2 + p_{0v} x_v)$$
$$= p^2_{01} + p^2_{02} + 2p_{01} p_{02} p_{21} + p^2_{0v}$$

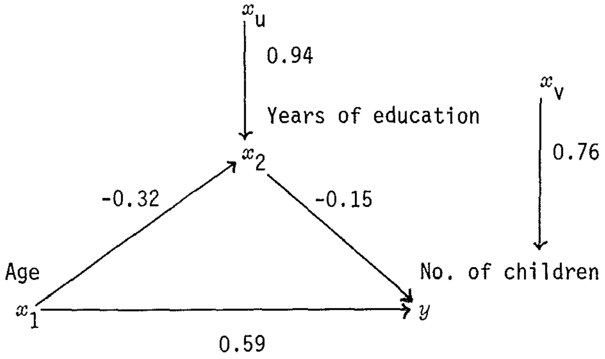
i.e.,

$$p_{ov} = \overline{1 - p^2_{01} - p^2_{02} - 2p_{01} p_{02} p_{21}} \qquad (2.5)$$

In this case

$$p_{2u} = 0.94$$
$$p_{0v} = 0.76$$

Inserting the values on the diagram, we obtain



2.2 REGRESSION EQUATIONS

This path model amounts to a sequence of conventional regression analyses and the solutions of the simultaneous equations (1.3) are simply the standardized regression coefficients - the *beta coefficients*. Thus the path p_{21} may be obtained by regressing x_2 on x_1, and the paths p_{01} and p_{02} may be obtained by regressing y on x_1 and x_2, using ordinary least squares. This is a useful result since it places the method of path analysis in the framework of standard statistical analysis and provides estimates for the standard errors of the coefficients obtained.

As a statistical technique, therefore, path analysis adds nothing to conventional regression analysis when applied recursively to a system of equations. But it does make the rationale for the system of regression equations explicit. And it presents "a method of measuring the direct influence along each separate path in such a system and thus of finding the degree to which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations" (Wright 1921).

A conventional path coefficient gives the expected effect of a change of one standard deviation in the explanatory variable (holding other variables constant); this expected change is expressed in terms of the standard deviation of the predicted variable. In this example,we wish to apportion the explanation of the dependent variable between the two explanatory variables.

The total effect of age may be expressed by the correlation between age and number of children i.e., $r_{01} = 0.64$. From equation (2.2) we see that this can be expressed as the sum of two components - p_{01}, the direct effect of age, and $p_{02}\, r_{12}$ $(=p_{01}\, p_{21})$ the indirect effect of age acting through years of education. Numerically this is

$$0.64 = 0.59 + (-0.32)(-0.15).$$

Thus the direct effect is +0.59.

The total effect of years of education is not however given by the correlation of years of education with number of children. A part of this correlation is due to the effect of the causally prior variable, age, on years of education. Thus,in terms of the model specified the total effect of years of education is the direct effect $p_{02} = -0.15$. We return to this problem in a more complex model later.

2.3 GENERALIZATION OF THE MODEL

There are five general characteristics of the simultaneous equation models we consider here. First, the models consist of a set of equations each of which possesses a disturbance term which summarizes the influence of unmeasured or unknown variables on the structure of interest. The models are thus not exact or deterministic but stochastic. Second, most applications are concerned with variables measured in cross-sectional surveys and are, therefore, static rather than dynamic models. Third, the models generally rule out two-way causation and are thus recursive. Fourth, the models are assumed linear in the variables and the disturbances. And finally, the disturbances are assumed to be contemporaneously independent.

We are concerned with linear additive asymetric relationships among a set of variables which are measured on an interval scale. In the qualitative diagram every included variable is represented either as completely determined by certain others or as an ultimate (exogenous) factor. (In the example above, there is only one exogenous variable: age). In a structural equation model, each equation represents a causal link rather than a mere empirical association. This is in contrast to a regression model where each equation represents the conditional mean of the dependent variable in that equation as a function of the explanatory variables. The most important special feature of the structural model is the simultaneity of the equations, i.e., the estimation of the parameters of a single equation is carried out in the context of the other equations in the system. The optimum properties of ordinary least squares (OLS) regression apply only to a single equation at a time. We must also take into account the fact that each equation is embedded in a set of equations which constitutes our recursive model with independent disturbances. Thus we need an estimation method which is optimal with regard to the joint estimation of the parameters of the system which make up the model. The simple example above shows that solving the set of simultaneous equations (1.3) is equivalent to equation-by-equation least squares regression. This result holds for all linear recursive models with independent disturbances (for proof, see for example Land (1973)).

The initial assumption for path analysis must be the specification of the causal (or temporal) ordering between the variables of the model. The data themselves cannot give us any assistance either for this or for the selection of the variables to be included in the model. The validity of these assumptions cannot be evaluated from the data; external criteria or substantive theory must provide the basis for this stage. Regardless of our ordering, the method of analysis will work and will provide results. No indication of error will emerge nor will the results be inconsistent. However, the diagrammatic representation of the model makes the assumptions explicit and provides a framework for the critical evaluation of the results.

The next important assumption involved is that the relationships are linear. Although this may not hold exactly in practice the linear regression of y on x may always be interpreted as the best linear approximation to the relationship when the latter is non-linear.

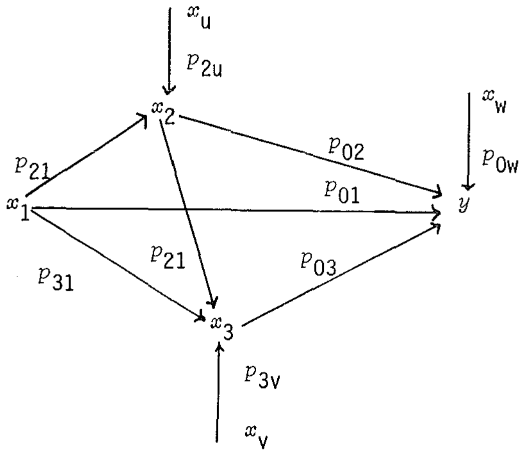
We also assume that each equation is additive. In other words we assume that a unit change in x_1, say, has the same effect on x_2 whatever the value of x_2. And we assume also that a unit change in x_1 has the same effect on x_2 whatever the values of the other variables. These assumptions may not be realistic, but fortunately, although non-linear and interaction effects are not included in the simple model, they can be included in it. An examination of the residuals in the equations can provide us with evidence as to whether such modifications of the model are necessary.

The error terms are assumed uncorrelated with all prior variables and hence with each other. They need not, however, be regarded as representing real variables but simply as an expression of the lack of information in the model and hence can reasonably be defined as being independent of the explanatory variables in the same equation.

The explanatory variables were described above as being measured on an interval scale. There is one important exception to this constraint. Binary variables (dichotomies) can be included and treated as interval level variables if the two categories are assigned numerical scores. We shall use 0 and 1 but the scores assigned to the two categories will not affect the standardized coefficients. As predictors, binary variables can be invaluable. Through them we can also incorporate polytomies (nominal scale variables) although there may be some difficulties with interpretation.

3. THE SATURATED ADDITIVE MODEL

A recursive model in which each variable is assumed to be dependent on all causally prior variables may be described as saturated. An example used with the Fiji Fertility Survey data is given below



EXAMPLE 2

where y : number of children
x_1: age in years
x_2: education in years
x_3: age at marriage

Empirical evidence suggests that x_1, x_2 and x_3 are all related to fertility. The link between x_1 and x_2 expresses the fact that the younger the age cohort the higher the proportion educated. The link from x_1 to x_3 will hold if age at marriage has changed over time. The link between x_2 and x_3 is based on the (testable) assumption that education delays entry into marriage either directly or by changing the alternatives available to the woman.

An additional factor in this example is that since the data come from a cross-sectional survey of ever-married women, there is an inbuilt positive correlation between age and age at marriage. This could be removed by censoring the date and considering only those women of 25 and over, say, who were married before 25. However, such modifications are not considered here since the examples merely provide illustrations of many of the technical and interpretative problems which will also arise in any serious attempt at model-building.

The causal ordering implied in this case is that age is causally prior to education, which is causally prior to age at marriage, which in turn is causally prior to number of children. This model in fact simply includes the variable x_3 (age at marriage) in the causal sequence between years of education and number of children in the simple model of the first example. The model can be written as the following set of equations:

$$x_2 = p_{21} x_1 + p_{2u} x_u \tag{3.1}$$

$$x_3 = p_{31} x_1 + p_{32} x_2 + p_{3v} x_v \tag{3.2}$$

$$y = p_{01} x_1 + p_{02} x_2 + p_{03} x_3 + p_{0w} x_w \tag{3.3}$$

Using the first method of Example 1, we can obtain equations for each of the six correlation coefficients in terms of the paths in the model. The equations are:

$$r_{21} = p_{21} \tag{3.4}$$

$$r_{31} = p_{31} + p_{32} r_{21} \tag{3.5}$$

$$r_{32} = p_{31} r_{12} + p_{32} \tag{3.6}$$

$$r_{01} = p_{01} + p_{02} r_{21} + p_{03} r_{31} \tag{3.7}$$

$$r_{02} = p_{01} r_{12} + p_{02} + p_{03} r_{32} \tag{3.8}$$

$$r_{03} = p_{01} r_{13} + p_{02} r_{23} + p_{03} \tag{3.9}$$

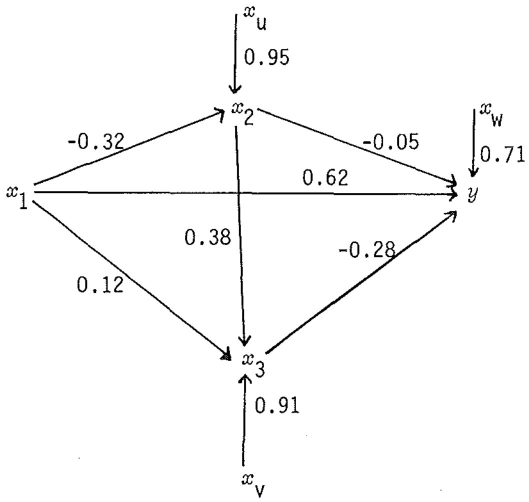
These equations are all of the same general form given by

$$r_{ij} = \sum_{q} p_{iq} r_{qj} \tag{3.10}$$

which is the basic theorem of path analysis, where q runs over all variables from which paths lead directly to x_i.

Equation (3.4) provides a solution for the value of p_{21}. Equations (3.5) and (3.6) provide the solution for the two unknowns p_{31} and p_{32}. Equations

(3.7), (3.8) and (3.9) provide the solutions for p_{01}, p_{02} and p_{03}. The algebra becomes cumbersome, however, even for this model and it is easier to obtain a solution by running the three regressions indicated by equations (3.1), (3.2) and (3.3). The standardized regression coefficients are the values of the path coefficients. First, we run the regression of x_2 (education) on x_1 (age); this gives p_{21}. Next, we run the regression of x_3 on x_1 and x_2, which gives the path coefficients p_{31} and p_{32}. Finally, we run the regression of y on x_1, x_2 and x_3, which gives p_{01}, p_{02} and p_{03}. The residual path coefficients p_{2u}, p_{3v} and p_{0w} are the square roots of the residual variances in the three regressions. When these runs were carried out on the Fiji data, the numerical values indicated on the diagram below were obtained:



The predictive model is represented by equation (3.3) and is

$$y = 0.62\, x_1 - 0.05\, x_2 - 0.28\, x_3$$

This simply represents the direct effects of the three explanatory variables and would be obtained by a normal regression analysis. The principal advantage of the structural model is that it enables us to proceed further in our analysis of the mechanism involved.

The total effect of age can be represented by the correlation between age and number of children and is equal to 0.64. However, from (3.7)

$$r_{01} = p_{01} + p_{02}\, r_{21} + p_{03}\, r_{31}.$$

Expanding further by substituting for r_{21} and r_{31} from (3.4) and (3.5) gives

$$r_{01} = p_{01} + p_{02}\, p_{21} + p_{03}\, p_{31} + p_{03}\, p_{32}\, p_{21}$$

This is the decomposition of the overall correlation of age and number of children and each of the terms above can be interpreted.

p_{01} is the <u>direct effect</u> of age; = +0.62

$p_{02}\, p_{21}$ is the <u>indirect effect</u> of age, working through its relationship with education; (-0.32)(-0.05) = + 0.02

$p_{03}\, p_{31}$ is the <u>indirect effect</u> of age, working through its relationship with age at marriage; (0.12)(-0.28) = -0.03

$p_{03}\, p_{32}\, p_{21}$ is the <u>indirect effect</u> of age, working through education, in turn working through age at marriage; (-0.32)(0.38)(-0.28) = +0.03

The four effects add up to the total effect r_{01} = 0.64.

As indicated in Example 1, the total effect of x_2 (years of education) is not equal to r_{02} (which is -0.34) but is equal to the sum of the direct and indirect paths from x_2 to y.

p_{02} is the <u>direct effect</u> of education; = -0.05

$p_{03}\, p_{32}$ is the <u>indirect effect</u> of education, working through age at marriage; (0.38)(-0.28) = -0.10

The total effect of education is $p_{02} + p_{03}\, p_{32}$ = -0.15

The total effect of x_3 (age at marriage) is the direct effect p_{03} since there are no variables between x_3 and y in the model; = -0.28.

Contrasting these results with the results obtained in Example 1 shows the effect of introducing a new variable into the structural model. The variable x_3 has been introduced explicitly into the system between x_2 and y. The total effect for both x_1 and x_2 remain unchanged since the new variable is causally posterior to both. However, the allocation of this total effect between direct and indirect effects changes for both x_1 and x_2. The table below gives the details:

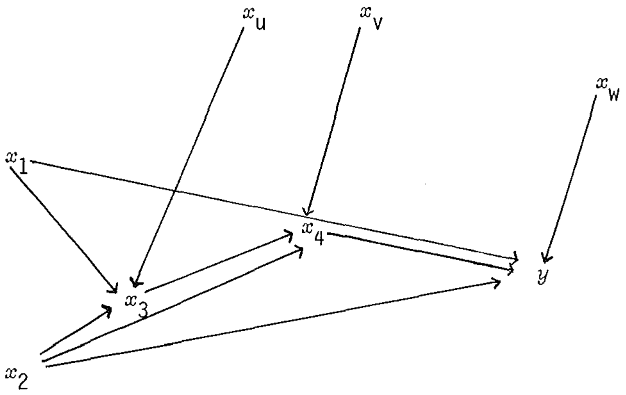
DECOMPOSITION OF TOTAL EFFECT FOR AGE (x_1) AND EDUCATION (x_2)

Variable	Type of effect	(x_3) excluded from the model	x_3 included in the model
Age (x_1)	Direct effect	+0.59	+0.62
	Indirect effects		
	(i) through x_3	not applicable	-0.03
	(ii) through x_2 and x_3 jointly	not applicable	+0.03
	(iii) through x_2 directly	+0.05	+0.02
Education (x_2)	Direct effect	-0.15	-0.05
	Indirect effect		
	through x_3	not applicable	-0.10

Thus omitting a variable from the model does not invalidate the results; it simply reduces the amount of information we obtain from the data. Introducing the variable x_3 into the model does not reduce the explanatory power of education; it does however provide an explanation of part of the mechanism through which education influences fertility.

4. UNSATURATED MODELS

A model in which some of the variables are not dependent on all causally prior variables may be described as unsaturated. Example 3 below represents such a system. The data again come from the Fiji study



EXAMPLE 3

where y : number of children

x_1: age in years

x_2: race

x_3: education in years

x_4: desired family size.

Three points distinguish this example from the others. First, there are two exogenous variables (ultimate factors) x_1 and x_2 which while prior to all the other variables are not ordered with respect to each other. These are not connected in the diagram since they are uncorrelated. Second, the variable x_2 is a binary variable (two categories Fijian and Indian). Third, two paths are omitted from the diagram - the paths p_{41} and p_{03}.

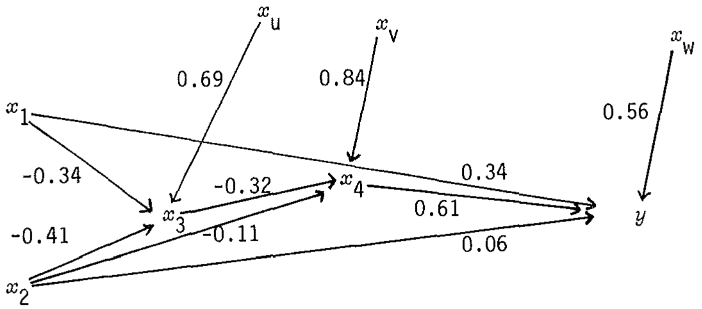
The reason for trying an unsaturated model is that there may be some theoretical basis for suggesting that some of the paths take zero values. It is the omission of these paths from the model which leads to the description of the model as *unsaturated*. The model can be written as:

$$x_3 = p_{31}\, x_1 + p_{32}\, x_2 + p_{3u}\, x_u \qquad (4.1)$$

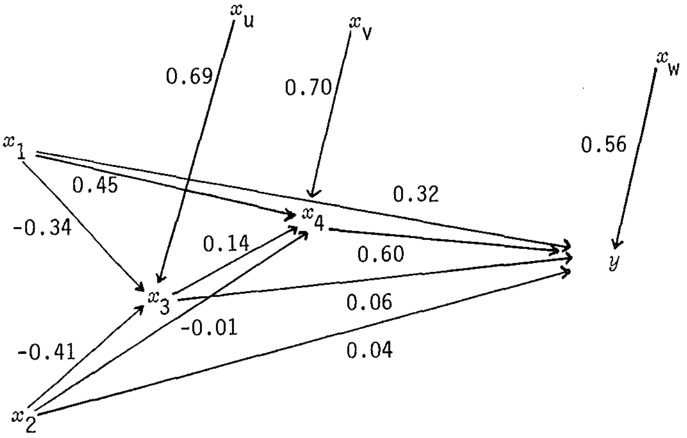
$$x_4 = \qquad p_{42}\, x_2 + p_{43}\, x_3 + p_{4v}\, x_v \qquad (4.2)$$

$$y = p_{01}\, x_1 + p_{02}\, x_2 + p_{04}\, x_4 + p_{0w}\, x_w \qquad (4.3)$$

The paths may be estimated directly by regression as before, which gives the numerical values inserted below:



However, the alternative method of estimation indicates that there may be some problems here. There are nine simultaneous equations available to estimate the seven paths in the model. Thus without further constraints the model is overidentified. The regression approach enables us to test these constraints. In essence we proceed by estimating all the coefficients in the fully saturated model and testing for significance the coefficients which we wish to remove. The result for the saturated model is:

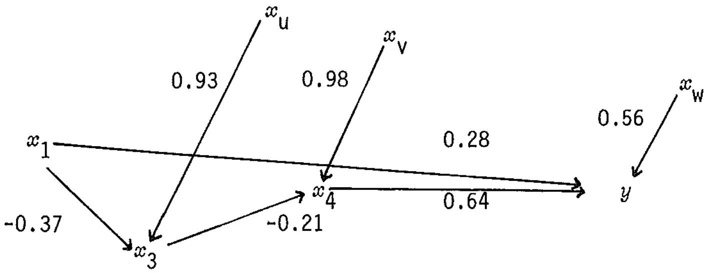


Since the modifications to the model involve only one coefficient in each of two equations, each of these coefficients can be tested using a t-test with $(n-p-1)$ degrees of freedom where p is the number of predictors in the equation. In fact both the coefficients are significant and should not be excluded from the model. If we wish to test more than one coefficient in any equation we must use an F test with appropriate degrees of freedom to test the full equation against the equation omitting the variables we wish to constrain. Since we are dealing with sample estimates of the population coefficients it is appropriate, even when we are dealing with a saturated model, to test whether the values obtained are simply due to sampling error.

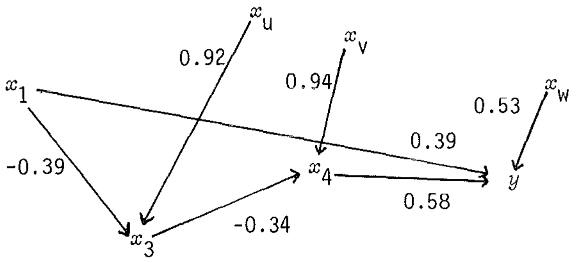
5. THE USE OF BINARY VARIABLES

The introduction of a binary variable - race (x_2) - into the structural model does not raise any special difficulties. Formally, a binary variable can be treated quite properly as an interval level variable. The assumption of additivity can be examined very easily in this case if we construct separate models for each of the two races and estimate the coefficients directly. Returning to the unsaturated model above, the results for the separate models are:

FIJIAN (RACE = 0)



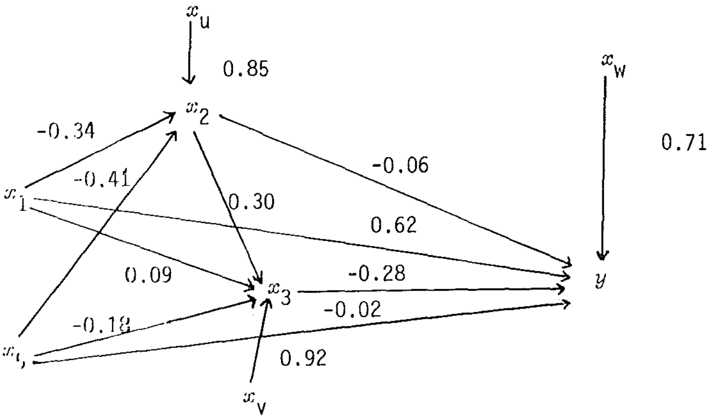
INDIAN (RACE = 1)



The pattern of effects is the same for the two models although the numerical values of the coefficient are not equal. It is possible to test the null hypothesis of equality for the coefficients using a *t*-test of an *F*-test. In this case, by inspection, the implications of the two sets of coefficients seem to be the same. The test results confirm this.

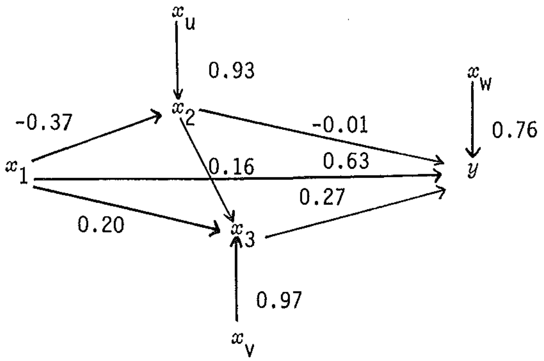
A further example illustrates the need for care in setting up the model. If race is included as an additional variable (x_5) in the model of Example 2, we obtain the result below:

x_1 = age in years
x_2 = education in years
x_3 = age at marriage
x_5 = race

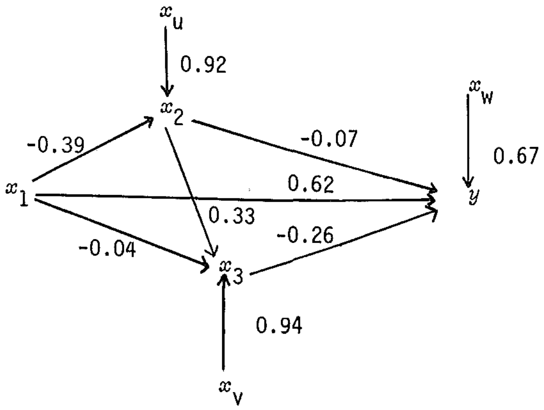


This model gives the additive effect of race as an exogenous variable in this structural model. However, in this case separate analyses for the two races give substantially different results. The subsample sizes are: Fijian, 2045; Indian, 2688.

FIJIAN (RACE = 0)



INDIAN (RACE = 1)



There are two important differences between the two cases. First, the difference in the relationship between age and age at marriage in the two races. For the Fijians, age and age at marriage are positively and strongly correlated; for the Indians, the correlation is negative. Also the direct effect of age at marriage on fertility is negative for each race and equal to within sampling error (-0.27 for Fijians; -0.26 for Indians). Thus the indirect effect of age on fertility through age at marriage is negative for the Fijians and positive, though small, for the Indians. Second, the effect of education is different in strength for the two cases. Both the direct and indirect effects of education are considerably larger for the Indians.

For this model, there is interaction between race and the other explanatory variables. Thus the two separate models provide a much more valuable representation than the pooled model. This does not invalidate the model from which race is omitted. That model (Example 2) provides an average or summary description of the way in which the other explanatory variables operate.

We can easily incorporate such interaction effects in the model by constructing new variables which represent the interaction. We must also include the binary variable itself in the equations. If we construct an interactive term for every predictor, the result will be equivalent to running two separate regressions. Thus, this technique is valuable only if we can assume that some of the predictors are stable for the whole population.

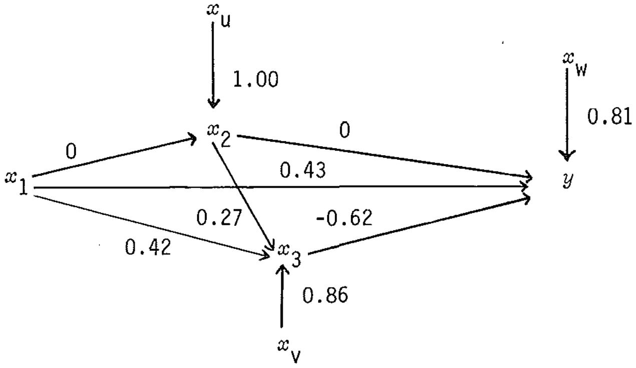
6. DISCUSSION

6.1 USE OF AGE AS AN EXPLANATORY VARIABLE

The association between age and number of children is the strongest association in the data and age is exogenous with respect to all the explanatory variables in the model. Conventionally in the analysis of fertility, the data are divided into age cohorts and the analysis is carried out independently for each cohort. Inspection of the data for the Fiji Fertility Survey shows that the effect of age is linear but does not easily provide evidence on interaction between age and the other explanatory variables. It seems likely, however, that education and age at marriage may have different effects for different age groups. This may be investigated by estimating the path coefficients for the structural model for each cohort separately and comparing the results.

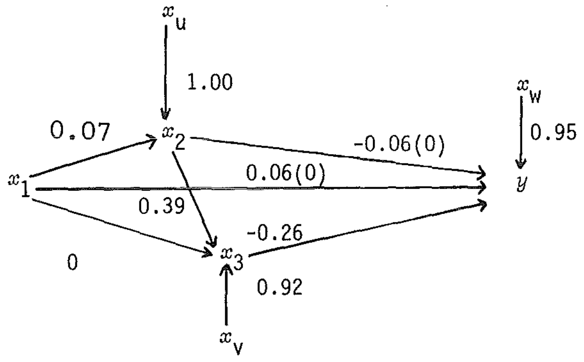
The two diagrams below present the results for the youngest and oldest age groups:

AGE GROUP 15-19 (n = 224)



Age still shows a strong direct effect for this age group and omitting it from the model would considerably reduce the explanatory power of the model. The direct effect of years of education almost disappears and the explanation is given by the indirect path through age at marriage. In the older age group, the direct effect of age almost disappears and the rest of the effect on y is mainly through education and age at marriage jointly. The results are intuitively reasonable and support the view that even within age cohorts, age should be included as an explicit variable in the structural model. No information is lost by this inclusion and considerable gains are possible both in terms of overall explanatory power and the decomposition of effects.

AGE GROUP 45-49 (n = 442)



6.2 STANDARDIZED VERSUS UNSTANDARDIZED COEFFICIENTS

A conventional path coefficient gives the expected effect on the predicted variable of a change of one standard deviation in the predictor. The assumption which underlies this is that the effect of a variable is relative to the distribution of the variables in the population. The unstandardized coefficient (the regression coefficient on the raw data) gives the effect in terms of a unit change in the predictor. Both coefficients give useful information. In the latter case, we consider the effect, say, of an additional *year* of education; in the former we consider the effect of a unit increase in years of education, defined in terms of the distribution of education in the population. The two approaches are compatible and represent different modes of interpretation, both of which can be useful in identifying the structural parameters of the model. Indeed, a mixture of standardized and unstandardized variables can be used in the same model.

6.3 CONCLUSION

A note of caution is appropriate here. The estimation of the coefficients of the structural model provides us with predictive equations for the variables in the model. It may seem in order to apply the results to the formulation of policy. For example, it is clear from the analysis that education has a negative effect on fertility. However this does not mean that an overall increase in level of education will reduce fertility. The equation we have produced is an equation for individuals within the present system. If we change the population distribution of the variable (and therefore the system), the result may not hold. Although for a particular individual an increase in education may lead to lower fertility than would otherwise be the case, this does not imply that a change in the overall level of education in the population will affect fertility. The non-experimental nature of the data precludes such inferences. Similarly, although age at marriage is negatively related to fertility, a change in the average age of marriage in the population may have no effect on fertility. The context-dependence of the results is important to bear in mind and is related in part to the omission of intermediate variables from the model. Education (or age at marriage) may be a useful proxy for prediction in the system due to a relationship with some of these intermediate variables. But both may simply represent cultural or social differences in the population which we are not measuring directly. If the population distributions of the variables are changed, they may lose their usefulness as predictors and a new structural model may be required.

The caution above is not a criticism of path analysis - it is a statement of the inherent constraints on the analysis of cross-sectional social science data. Path analysis models can be very helpful in disentangling a complex set of relationships and, used with care, can add considerably to our knowledge of the mechanisms at work in the population. Not the least of the advantages is the fact that the use of such models forces the researcher to be explicit about his theorizing and permits criticism and evaluation of the assumptions built into the models.

REFERENCES

O.D. Duncan, "Path Analysis: Sociological Examples", *American Journal of Sociology*, 72, pp. 1-16, 1966.

O.D. Duncan, *Introduction to Structural Equation Models*, (New York: Academic Press, 1975).

T.C. Koopmans and O. Reiersol, "The identification of structural characteristics", *Annals of Mathematical Statistics*, 21, pp. 165-181, 1950.

K.C. Land, "Identification, parameter estimation and hypothesis testing in recursive sociological models", *Structural Equation Models in the Social Sciences*, eds. A.S. Goldberger and O.D. Duncan, (New York: Seminar Press, 1973) Ch.2.

K.I. Macdonald, "Path Analysis", *Analysis of Survey Data*, eds. C.A. O'Muircheartaigh and C.D. Payne, (London: Wiley, 1977), Vol.2, Ch.3.

S. Wright, "Correlation and Causation", *Journal of Agricultural Research*, 20, pp. 557-585, 1921.

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

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