PATH ANALYSIS: A MULTIVARIATE TECHNIQUE

Margaret A. Chmielewski, Texas A&M University

I. Introduction

Path analysis is a method for quantifying the causal relationships between a set of exogenous and endogenous variables. A variable is said to be endogenous if it has a direct cause and exogenous otherwise. The first step in path analysis is to define the causal relationships by a causal diagram. Straight line arrows are drawn from each variable to its direct effect. If no variable is both a cause and effect of another variable then the system is said to be recursive. Otherwise it is non-recursive. Unexplained correlation between exogenous variables is indicated by curved double headed arrows. Each direct effect has a residual, U. In Figure 1, X_1 and X_2 are exogenous and X_3, X_4 and X_5 are endogenous variables. The residuals U_3, U_4 and U_5 are treated as exogenous variables and can possibly be correlated.

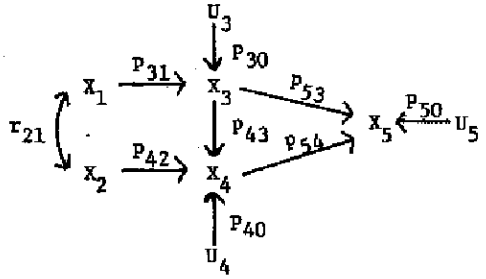


Figure 1

Path coefficients were introduced by Wright (1921, 1934, 1954) as a method for relating the correlation coefficients among a system of variables to the functional relationships among the variables. Wright was interested in genetics and considered path diagrams which related the genetic constitution of parents to the genetic constitution of their offspring. For recursive systems he gave rules for decomposing the correlation between two variables into direct, indirect and spurious effects.

Simon (1954) used a three variable system to determine whether the correlation between two variables was spurious or genuine. He discussed the following aspects of correlation. Suppose that we are interested in whether the significant correlation, r_{xy}, between two variables x and y means that there is a causal relation between x and y. To check this out, introduce a third variable, z, compute $r_{xy \cdot z}$, and compare this partial correlation with r_{xy}. If $r_{xy \cdot z}$ is approximately zero while r_{xy} is not then we can conclude that either (i) z is an intervening variable, i.e. the causal effect of x on y (or y on x) operates through z or (ii) the correlation between x and y is spurious, i.e. it results from the causal effect of z on both x and y. Thus the assumed relationships, rather than the statistical evidence, determine whether the interpretation is (i) or (ii). These assumptions are incorporated in the causal network and indicated by the arrows drawn between the variables. Thus (i) is represented by $x \to z \to y$ and (ii) is represented by $z \overset{x}{\underset{y}{\nwarrow\!\nearrow}}$.

Path analysis showed a resurgence in the 1960's with most of the papers, which were application oriented, appearing in sociological journals. Without a theoretical base some misconceptions, which are discussed later, were fostered. A theoretical framework for path analysis is given by Kang and Seneta (1980). It is this framework which helps to clear up the confusion found in the social science literature. Today, path analysis is still used in the social sciences, e.g. psychology, education, sociology, as well as in the biological sciences. Included in the bibliography are a few references pertaining to these applications.

II. The Causal Network

The variables in the causal network form a multivariate system. Thus, multivariate techniques will be used to define the path coefficients. For the present we will assume that sample sizes are large enough so that the estimates are representative of the "true values".

Let $\underline{X}' = [X_1, X_2, \ldots, X_m]$ with $E[\underline{X}] = \underline{0}$ and $\mathrm{Var}[\underline{X}] = \Sigma$. If we wish to predict one of them, say X_1, in terms of a linear combination of the others, $\beta_2 X_2 + \ldots + \beta_m X_m = \underline{\beta}'\underline{X}^*$, then that value of $\underline{\beta}$ which minimizes $U = E[X_1 - \underline{\beta}'\underline{X}^*]^2$ is given by $\underline{\beta} = \Sigma_2^{-1}\underline{\sigma}$ where $\Sigma_2 = \mathrm{Var}[\underline{X}^*]$ and $\underline{\sigma} = \mathrm{Cov}[X_1, \underline{X}^*]$. For $\underline{X}^* = \underline{x}^*$ the best linear predictor of X_1 is $\underline{\beta}'\underline{x}^* = \underline{\sigma}'\Sigma_2^{-1}\underline{x}^*$ and the $\underline{\beta}$ are called the partial regression coefficients. Also, the residual, $U = X_1 - \underline{\sigma}'\Sigma_2^{-1}\underline{X}^*$, is uncorrelated with $X_2, \ldots, X_m$. For further details the reader can consult Kshirsagar (1972).

In path analysis, the variables are usually standardized to a mean of zero and variance of one. Thus $\Sigma = R$ where R is a correlation matrix and the β_i are now denoted by p_i and called path coefficients.

A causal network consists of several subsystems. For each subsystem the path coefficients are found. For example, the causal network in Figure 1 can be broken down into three subsystems (Figure 2). The path coefficients are now denoted by p_{ji} where j denotes the dependent variable in a particular subsystem. To find the p_{ji} you need a table of pairwise correlations.

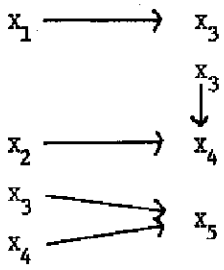


Figure 2

For each subsystem, Σ_2, $\underline{\sigma}$ and $\underline{X}^*$ will be defined with respect to the variables within the subsystem. The path coefficients can also be found using the raw data. For the subsystems in Figure 2 we could find the path coefficients by fitting the following models:

$$X_3 = p_{31}X_1 + p_{30}U_3$$

$$X_4 = p_{42}X_2 + p_{43}X_3 + p_{40}U_4$$

$$X_5 = p_{54}X_4 + p_{53}X_3 + p_{50}U_5.$$

Here, $p_{j0} = (1 - R^2)^{\frac{1}{2}}$ where R^2 is the square of the multiple correlation coefficient for that particular model. The relationship

$$p_{j0}^2 + R^2 = 1$$

or

$$p_{j0}^2 + \underline{\sigma}'\underline{p} = 1$$

where $\underline{p}$ is the vector of path coefficients for the particular subsystem is called the equation of complete determination.

Once the path coefficients are found the usual procedure for recursive causal systems, is to decompose the total pairwise correlation into direct effects, indirect effects and spurious effects. As an example of these effects, consider Figure 1. Here, the pathway from X_1 to X_3 represents a direct effect. The pathway from X_1 to X_3 to X_4 represents an indirect effect between X_1 and X_4 (through X_3). Finally, any relationship between X_2 and X_3 represents a spurious effect (due to the fact that they both cause X_4). Note that these effects are defined in terms of the hypothesized underlying causal network.

III. Clarification of Some Misconceptions

One aspect of the social science literature is the discussion of overidentified and underidentified systems. Given a causal network we know we can find the path coefficients by $\underline{\beta} = \Sigma^{-1}\underline{\sigma}$ where Σ and $\underline{\sigma}$ are defined appropriately. We also know that an equivalent method is to fit several regression models. The computational equivalence of the two methods is what causes difficulty. Since the regression method is used almost exclusively the tendency has been to impose multiple linear regression assumptions on the problem. For example, consider the following two regression models, each representing a subsystem in a causal network.

$$X_2 = p_{21}X_1 + p_{20}U_2$$

$$X_3 = p_{31}X_1 + p_{32}X_2 + p_{30}U_3.$$

The assumptions imposed [e.g. Asher (1976)] on these two equations are:

$$\text{(i) } \mathrm{Corr}(X_1, U_2) = \mathrm{Corr}(X_1, U_3) = \mathrm{Corr}(X_2, U_3) = 0$$

and

$$\text{(ii) } \mathrm{Corr}(U_2, U_3) = 0.$$

But (i) is not an assumption, it is a fact and (ii) may not necessarily be true. The reason for assumption (ii) is taken from simultaneous equation theory. In order for the two equations to be solved simultaneously one restriction is needed. This is because there are two equations in three unknowns (p_{21}, p_{31}, p_{32}) and thus the set of equations is underidentified. A similar reasoning leads to overidentified models. For example, consider the following three equations which define a causal network.

$$X_2 = p_{21}X_1 + p_{20}U_2$$

$$X_3 = p_{31}X_3 + p_{32}X_2 + p_{30}U_3$$

$$X_4 = p_{42}X_2 + p_{43}X_3 + p_{40}U_4.$$

There are three equations in five unknowns. Imposing the assumptions $\mathrm{Corr}(U_2, U_3) = \mathrm{Corr}(U_2, U_4) = \mathrm{Corr}(U_3, U_4) = 0$ gives three equations in $5 - 3 = 2$ unknowns and the system is overidentified. However, the problem of overidentified and underidentified systems vanishes when the assumption of uncorrelated residuals is removed and the best linear predictor theory is used.

IV. Internal Consistency

In path analysis the causal network does not usually include all possible arrows from each of the variables to all the other variables. If this were the case then we would only need to look at the partial correlation coefficients, i.e. each variable adjusted for all the others. Thus we need to consider whether a direct effect should be included between two variables X_1 and

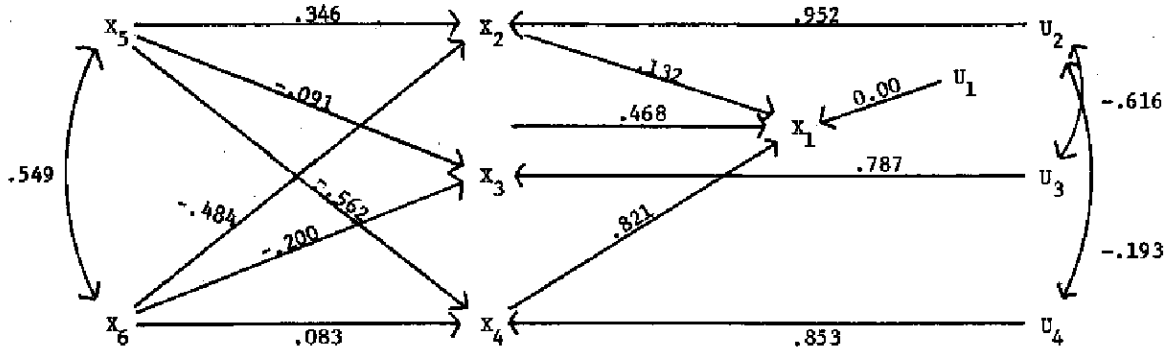
X_j when presently X_i is only an indirect effect of X_j. If $p_{ji} \neq 0$, i.e. the direct effect should not be included, then the system is said to be internally consistent. For example, consider $X_1 \rightarrow X_3 \rightarrow X_2$. Should a direct effect from $X_1 \rightarrow X_2$ also be included? We know that $p_{21} = 0$ if and only if $\rho_{12.3} = 0$. Thus, in general, we need only check partial correlation coefficients to see if the system is internally consistent. An equivalent method, which does not involve any extra computations, is given by Kang and Seneta (1980). Let X_i be an indirect cause of X_j and let $X_{j(1)}, \ldots, X_{j(m)}$ be direct causes of X_j. Then Kang and Seneta prove that $\rho(i, j \cdot j(1), \ldots, j(m)) = 0$ for all such X_i and X_j is equivalent to $\text{Corr}(U_K, U_\ell) = 0$, $K \neq \ell$, whenever X_K is a cause (direct or indirect) of X_ℓ for all such X_K, X_ℓ. Thus internal consistency can easily be checked by examining the pairwise correlations between the exogenous variables. Internal inconsistency will result in the addition of arrows to the path diagram and the recalculation of certain path coefficients.

V. *An Example*

The following example is from Duncan (1966). The data are from a population study conducted in Chicago in 1940. $X_1 \equiv$ population density (in logs), $X_2 \equiv$ persons per dwelling unit (in logs), $X_3 \equiv$ dwelling units per structure (in logs), $X_4 \equiv$ structures per acre (in logs), $X_5 \equiv$ distance from center and $X_6 \equiv$ recency of growth. The correlation matrix is as follows:

	X_2	X_3	X_4	X_5	X_6
X_1	-.419	.636	.923	-.663	-.390
X_2		-.625	-.315	.296	.099
X_3			.305	-.594	-.466
X_4				-.517	-.226
X_5					.549

The completed path diagram is in Figure 3.



$$X_1 = p_{12}X_2 + p_{13}X_3 + p_{14}X_4 + p_{10}U_1$$
$$X_2 = p_{25}X_5 + p_{26}X_6 + p_{20}U_2$$
$$X_3 = p_{35}X_5 + p_{36}X_6 + p_{30}U_3$$
$$X_4 = p_{45}X_5 + p_{46}X_6 + p_{40}U_4$$

Figure 3

Notice that correlations between residuals are included with the rule of thumb that any correlation less than .1 in absolute value is deemed negligible. The system is also internally consistent since the only nonzero correlations involve variables where neither is a cause of the other.

VI. *Decomposition of Correlation*

Recall that earlier we mentioned that the correlation between two variables could be decomposed into direct, indirect and spurious effects. The actual decomposition is expressed in terms of the sum of all simple and compound paths. A simple path comprises a direct effect

from one variable to another while a compound path comprises an effect from one variable to another variable through one or more variables. The compound path is then the product of all the simple paths of which it is composed. The rules for decomposition, which involve tracing an effect back to all its causes, are given by Wright (1921) and are as follows:

(1) no path may pass through the same variable more than once;

(2) no path may go backward on (against the direction of) an arrow after the path has gone forward on a different arrow;

(3) no path may pass through a double-headed curved arrow more than once in any single path.

To illustrate consider the following example from Asher (1976) and Miller and Stokes (1966). Here $X_1 \equiv$ constituency's attitudes, $X_2 \equiv$ representative's perceptions of constituency's attitudes, $X_3 \equiv$ representative's attitudes, and $X_4 \equiv$ representative's roll call behavior. The completed path diagram is given in Figure 4. Using Wright's rules the correlations are decomposed as follows with D indicating a direct effect, I an indirect effect and S a spurious effect.

$$\rho_{13} = p_{31} + p_{21}p_{32} \qquad (D + I)$$

$$\rho_{23} = p_{32} + p_{21}p_{31} \qquad (D + S)$$

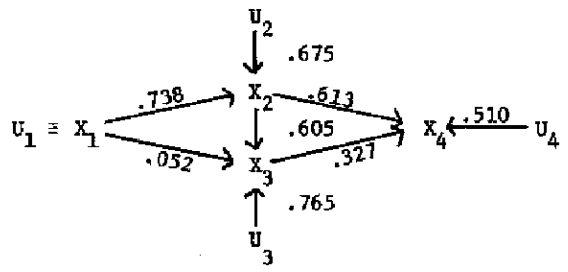
$$\rho_{12} = p_{21} \qquad (D)$$

$$\rho_{14} = p_{21}p_{42} + p_{31}p_{43} + p_{21}p_{32}p_{43} \qquad (I + I + I)$$

$$\rho_{24} = p_{42} + p_{32}p_{43} + p_{21}p_{31}p_{43} \qquad (D + I + S)$$

$$\rho_{34} = p_{43} + p_{32}p_{42} + p_{31}p_{21}p_{42} \qquad (D + S + S)$$

Using the decomposition it is seen that (i) most of the indirect effect of constituency's attitudes on representative's behavior (X_1 on X_4) is through the representative's perception of constituency attitudes ($p_{21}p_{42} = .452$) and (ii) the representative's perception of constituency attitudes has a larger direct effect on representative's behavior than the attitudes do (.613 vs. .327).



$$\text{Corr}[X_1, U_4] = 0.064$$

$$\text{Corr}[U_2, U_3] = 0$$

$$\text{Corr}[U_2, U_4] = -0.070$$

$$\text{Corr}[U_3, U_4] = 0.004$$

Figure 4

VII. Conclusions

In this last section we briefly discuss standardized vs. unstandardized coefficients, the Koopman's-Hood approach, a matrix formulation of path analysis, non-recursive models, and removal of the large sample criteria.

Recall, that the first step in path analysis was to standardize the variables. However, over the years there has been some argument as to whether the unstandardized coefficients should be used. The usual recommendation is to use unstandardized regression coefficients when comparing causal structures across different populations because of the possibility of different variances of the variables across the populations [e.g. Kim and Mueller (1976)]. Tukey (1964) believes that unstandardized coefficients should be used solely. The reasoning behind this decision goes back to regression and correlation. That is, regression coefficients are used for defining functional relationships while correlation coefficients are mainly descriptive measures. Path coefficients are similar to correlation coefficients except that they do not have the nice property of being between plus or minus one. However, path coefficients can easily be changed into either regression coefficients or partial correlation coefficients by the following relationships:

$$p_{12} = \beta_{12\cdot 34\ldots n}\frac{\sigma_2}{\sigma_1} = \rho_{12\cdot 34\ldots n}\frac{\sigma_{1\cdot 34\ldots n}}{\sigma_{2\cdot 34\ldots n}}$$

where

$$\sigma^2_{1\cdot 23\ldots n} = \sigma_1^2(1 - \rho^2_{12})(1 - \rho^2_{13\cdot 2})$$
$$(1 - \rho^2_{14\cdot 23})\ldots(1 - \rho^2_{1n\cdot 23\ldots n-1}).$$

In section III we briefly mentioned certain misconceptions. The assumption which led to the system of equations possibly not being identifiable was the one of uncorrelated residuals. In the best linear predictor theory this assumption is unnecessary. Suppose, however, one wishes to invoke this assumption. Then the best linear predictor approach is no longer applicable. Rather the Koopman's-Hood approach is now used to find the path coefficients. Using Koopman's-Hood means that we can no longer define the residuals as previously defined. This essentially is the confusion which exists in the social sci-

ence literature. That is, the Koopman's-Hood assumptions are used but the path coefficients, including the residuals, are found using the best linear predictor theory. For further discussion see Kang and Seneta (1980).

Kang and Seneta (1980) also give a matrix formulation of path analysis which can be very useful for programming purposes. They also give an example of a non-recursive causal network noting that the best linear predictor theory is still appropriate. Thus, there is no need for such procedures like two-stage least squares which is advocated in the social science literature as a method of estimation for non-recursive models [e.g. Asher (1976)].

Finally, consider the relaxation of the large sample assumption. If it is reasonable to assume that the variables have a multivariate Gaussian distribution then normal theory results can be used to make inferences about multiple and partial correlation coefficients.

Bibliography

Alwin, D. F. and Hauser, R. M. (1975). The Decomposition of Effects in Path Analysis. *American Sociological Review 40*, 37-47.

Asher, H. B. (1976). *Causal Modeling*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-003. Beverly Hills and London: Sage Publications.

Duncan, O. D. (1966). Path Analysis: Sociological Examples. *American Journal Sociology 72*, 1-16.

Hortynski, J. A. (1979). Correlation and path analysis in strawberry seedlings. *Genetica Polonica 20*, 549-566.

Kang, K. M. and Seneta, E. (1980). Path Analysis: An Exposition. In *Developments in Statistics* (P. R. Krishnaiah, ed), Vol. 3, Chapter 4, pp. 217-246. Academic Press, New York.

Kim, J. and Mueller, C. W. (1976). Standardized and unstandardized coefficients in causal analysis: An expository note. *Sociological Methods and Research 4*, 423-438. Sage Publications, Inc.

Kshirsagar, A. M. (1972). *Multivariate Analysis*. Marcel Dekker, Inc., New York.

Land, K. C. (1969). Principles of path analysis. In *Sociological Methodology 1969*, E. F. Borgatta (ed). Jossey-Bass, San Francisco.

Miller, W. E. and Stokes, D. E. (1966). Constituency influences in congress. In *Elections and the Political Order* (A. Campbell et al., eds.), pp. 351-372. John Wiley and Sons, New York.

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

Tukey, J. W. (1964). Causation, regression and path analysis. In *Statistics and Mathematics in Biology*, O. Kempthorne (ed.), Hafner Publication Co., Inc., New York.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research 20*, 557-585.

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

Wright, S. (1954). The interpretation of multivariate systems. In *Statistics and Mathematics in Biology* O. Kempthorne et al., eds), Chapter 2, pp. 11-33. Iowa State College Press, Ames, Iowa.

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

Wright, S. (1960). The treatment of reciprocal interaction, with or without lag, in path analysis. *Biometrics 16*, 423-445.