Conference on Applied Statistics in Agriculture   1990 - 2nd Annual Conference Proceedings

PATH ANALYSIS IN AGRICULTURAL RESEARCH

K. Bondari

Recommended Citation

PATH ANALYSIS IN AGRICULTURAL RESEARCH

K. Bondari

Department of Statistical & Computer Services, Coastal Plain Experiment Station, University of Georgia, Tifton, GA 31793

ABSTRACT

Path analysis introduced by Wright in 1921 as "correlation and causation" has been extensively used in agriculture, sociology, and epidemiology, among many other fields. This study will review path diagrams, algorithms, and the relationship to standardized and multivariate regression analyses. Basic assumptions underlying path analysis (e.g., cause and effect relationship, linearity of regression, complete additivity) will also be discussed. Several research examples will be presented to better acquaint statisticians involved in agricultural research with the methodology and application of path analysis suitable for agricultural data. The method of path coefficient is simple, easy to use, and if "tracing rules" in a path diagram are learned, the method of path coefficient could be an important research tool.

Keywords: Path coefficient, Path diagram, Causal relations, Tracing rules, Inbreeding, Multiple regression.

1. INTRODUCTION

The method of path analysis was originated and developed by Sewall Wright in the 1920s. He published a series of papers describing the theory and application of the path method (Wright 1921a, 1923, 1934, 1954, 1960a, 1960b, 1968). Some of these publications were in response to criticisms concerning the philosophical basis of "causal relation" and reliability of the results (Niles 1922, 1923) and the use of correlation vs. regression or covariance (Tukey, 1954; Turner and Stevens, 1959). Lush (1947), Kempthorne (1957), Li (1975), Duncan (1975) Borgatta, Bohrnstedt (1969), Kang (1971), Seneta (1972), Kang and Seneta (1980), and Emigh (1984), among many other authors, have described path analysis in theory and application.

Wright (1960a) described the method of path analysis as a technique to deal with a system of interrelated variables. The original idea (Wright, 1921a) was to combine the knowledge of causal relationships among variables with the degree of relationship measured by correlation coefficient to determine the direct influence along each separate path and the degree to which variation of a given effect is determined by each particular cause. Correlation does not generally imply causation, but Wright (1960a) reemphasizes his account of Wright (1921a) that the path coefficient method is by no means restricted to relations that can be described as ones of cause and effect. Furthermore, Wright (1960a) defends his

assumption of linear relations among variables on the ground that non-linear relations may sometimes be transformed systematically throughout a path diagram into linear ones. In nearly 70 years after its introduction, concerns about the validity of the method and assumptions made still remain.

Although it is difficult to give a simple definition of path analysis, the path method was originally introduced by Wright (1921a, 1921b) to study the interrelationship and persistence of variously linked heterozygotic attributes after successive brother-sister mating in an inbreeding experiment. Animal breeders and plant breeders have extensively used this method. Application, however, is not limited to agriculture. Several authors have discussed the use of path analysis in epidemiology (e.g., Goldsmith and Berglund 1974; Goldsmith 1977; Rao et al. 1976; Cloninger et al. 1983; Fulker 1988) and in sociology (e.g., Blalock 1964; Duncan 1966; Land 1969; Heise 1969).

The objectives of this paper were to: (1) review the statistical basis of path analysis in relation to regression and correlation coefficients and (2) demonstrate the application in agricultural research.

2.   METHODOLOGY

Multiple Regression

Suppose in a set of n+1 random variables, we wish to estimate one (call it y) by a linear combination of the others (x_1, x_2, ..., x_n). Our multiple regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e \qquad (1)$$

The residual random variable e has zero expectation and is uncorrelated pairwise with each of x_1, x_2, ..., x_n. Equation (1) can simply be rewritten as:

$$y = (\beta_0 + \beta_1 \overline{x}_1 + \beta_2 \overline{x}_2 + \ldots + \beta_n \overline{x}_n) + \beta_1 (x_1 - \overline{x}_1) + \beta_2 (x_2 - \overline{x}_2) + \ldots + \beta_n (x_n - \overline{x}_n) + e$$

Or simply

$$y - \overline{y} = \beta_1 (x_1 - \overline{x}_1) + \beta_2 (x_2 - \overline{x}_2) + \ldots + \beta_n (x_n - \overline{x}_n) + e \qquad (2)$$

We can fit model (2) by least squares and get exactly the same estimated parameters and fitted values that we would have obtained from fitting model (1). If we divide Equation (2) by σ_y throughout we have in fact standardized the dependent variable y. We can make similar transformation on the variables $x_1, x_2, \ldots x_n$ and rewrite as follows:

$$\frac{y - \overline{y}}{\sigma_y} = \frac{\beta_1 \sigma_{x_1}}{\sigma_y} \frac{(x_1 - \overline{x}_1)}{\sigma_{x_1}} + \frac{\beta_2 \sigma_{x_2}}{\sigma_y} \frac{(x_2 - \overline{x}_2)}{\sigma_{x_2}} + \ldots + \frac{e \sigma_e}{\sigma_y \sigma_e} \qquad (3)$$

Note that the random variable e is also standardized in (3).

We can rewrite Equation (3) as:

$$Y = p_1 X_1 + p_2 X_2 + \ldots\ldots\ldots + p_n X_n + p_e E \qquad (4)$$

where,

$$Y = (y - \overline{y})/\sigma_y \quad, \quad X_i = (x_i - \overline{x})/\sigma_{x_i} \quad, i = 1, \ldots, n, \quad E = e/\sigma_e.$$

The p_i's, i=1, ..., n, in Equation (4) are standardized partial regression coefficients (path coefficients in Wright's terminology) to be estimated from the transformed (standardized) data as:

$$p_i = \beta_i \sigma_{x_i}/\sigma_Y \qquad i = 1, \ldots, n \qquad (5)$$

and could directly be computed by the **STB** option of the **PROC REG** in SAS (1985). Comparing (4) to (1) indicates that standardization has the effect of making $\beta_0 = 0$ and p_i's, i=1, ..., n, invariant with regards to changes in the units in which the quantities are measured. The coefficients β_1, β_2, β_3, etc. in (5) are actually:

$$\beta_{y1.2,3,\ldots n} \quad, \quad \beta_{y2.1,3,\ldots n} \text{ and } \beta_{y3.1,2,4,\ldots n} \quad, \qquad \text{respectively.} \qquad \text{For}$$

example $\beta_{y1.2,3,\ldots n}$ represents how y would vary per unit change in X_1 if X_2, X_3, . . ., X_n were all held fixed which gives only a partial picture of what happens to y in Equation (1) as X_1 changes.

The path diagram corresponding to Equation (4) is shown in Figure 1. This diagram shows the linear relations between Y and X_i, i=1,2,...,n, and Y and E, represented by directed arrows pointing from X or E to Y. Double-headed curved arrows show correlations between X's and the fact that these correlations are symmetric. X's and E are assumed uncorrelated and thus there are no paths connecting them in Figure 1. The normal equations in terms of path coefficients (a) and the X'X matrix in terms of correlations between x's (b) are shown below
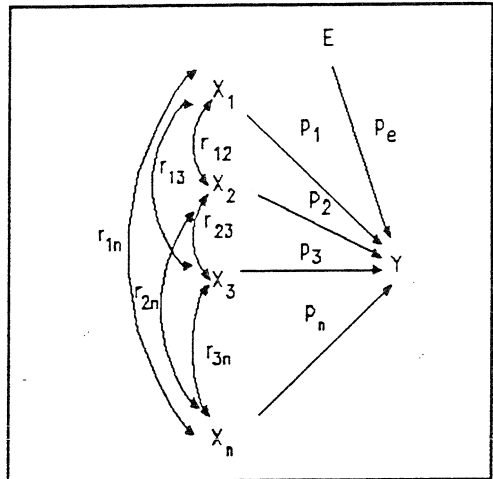

Figure 1. Path diagram for n correlated X's.

where, r_{ij} represent correlations between X_i and X_j (e.g., r_{12} indicates correlation between X_1 and X_2) and r_{yi}, i=1,2,...,n (total correlation in Wright's terminology) indicate correlations between Y and X_i.

These normal equations indicate that the correlations of Y with any X_i are made up of (1) direct path from X_i to Y and (2) indirect paths through variables X_js (correlated with X_is) to Y. For instance, the direct effect of Y on X_1 is shown by the path from X_1 to Y or p_1. The indirect effect of Y on X_1

$$\begin{vmatrix} p_1 & + p_2 r_{12} & + p_3 r_{13} & + . . & + p_n r_{1n} & = r_{Y1} \\ p_1 r_{21} & + p_2 & + p_3 r_{23} & + . . & + p_n r_{2n} & = r_{Y2} \\ p_1 r_{31} & + p_2 r_{32} & + p_3 & + . . & + p_n r_{3n} & = r_{Y3} \\ . & + . & + . & + . . + . & = . \\ . & + . & + . & + . . + . & = . \\ p_1 r_{n1} & + p_2 r_{n2} & + p_3 r_{n3} & + . . & + p_n & = r_{Yn} \end{vmatrix}$$

$$\begin{bmatrix} 1 & r_{12} & r_{13} & . & . & r_{1n} \\ r_{21} & 1 & r_{23} & . & . & r_{2n} \\ r_{31} & r_{32} & 1 & . & . & r_{3n} \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{n1} & r_{n2} & r_{n3} & . & . & 1 \end{bmatrix}$$

(a) Normal Equations for Figure 1. (b) X'X matrix for Figure 1.

via X_2 is $p_2 r_{12}$ and via X_n is $p_n r_{1n}$ as shown in the first line of equations (a). This aspect of path analysis which partitions the total correlation into causal components can be very useful in describing relationships among variables. If X's are not correlated:

$$p_1^2 + p_2^2 + p_3^2 + . . . + p_n^2 + p_e^2 = 1 \qquad (6)$$

The square of the path coefficients (p_i^2) are called coefficients of determination by Wright and were used to measure the portion of the variability in Y that is directly attributable to X_i, i=1, . . . , n.

Tracing rule

Only a few simple rules will be given here for tracing the paths of a diagram. Wright (1960a; 1968) or Li (1975) should be consulted for more details. In reading a path diagram, one should follow the direction of the arrows going backward first and then going forward (e.g., $Y \leftarrow X_1 \rightarrow Y = p_1^2$ or $Y \leftarrow X_1 \rightarrow X_2 \rightarrow Y = p_1 r_{12} p_2$, in Figure 1). Tracing should proceed without going back after going forward along an arrow and without passing through any variable twice in the same path (e.g., $Y \leftarrow X_1 \rightarrow X_2 \leftarrow X_1 \rightarrow Y$ is not permitted in Figure 1). A double-headed curved arrow ($\leftrightarrow$) used to indicate correlation between two variables may be used in either direction, but any specific path cannot involve more than one correlation step. For example, $Y \leftarrow X_1 \leftrightarrow X_2 \leftrightarrow X_3 \rightarrow Y$ in Figure 1 is not permitted, but $Y \rightarrow X_1 \rightarrow X_3 \rightarrow Y = p_1 r_{13} p_3$ is valid.

Two or More Dependent Variables

Extension to more than one standardized dependent variable is not difficult. For two dependent variables Y_1 and Y_2:

$$Y_1 = p_1 X_1 + p_2 X_2 + p_3 X_3 + . . . + p_n X_n \quad , \quad Y_2 = \acute{p}_1 X_1 + \acute{p}_2 X_2 + \acute{p}_3 X_3 + . . . + \acute{p}_n X_n$$

$$r_{X_1 X_2} = p_1 \acute{p}_1 + p_2 \acute{p}_2 + p_3 \acute{p}_3 + . . . + p_n \acute{p}_n + \sum_{i \neq j} p_i \acute{p}_i r_{ij} = \sum_{i,j} p_i \acute{p}_i r_{ij}$$

$r_{ij} = 1$ if i=j. A simple case involving $Y_1 = X_1 + E_1$ and $Y_2 = X_2 + E_2$ is

shown in Figure 2. More details are given in Kempthorne (1957). This concept could easily be extended to the case where variables $Z_1...Z_t$ result in $X_1...X_n$ which in turn result in $Y_1...Y_n$.

$$r_y = p_1\, r_x\, p_2 + e_1\, r_e\, e_2$$

Figure 2. Path diagram involving two independent variables.

Causal Relations

According to Kempthorne (1957), a path coefficient is always a standard partial regression coefficient. A standard partial regression coefficient of Y on X_i is a path coefficient if: (a) X_i is a cause more or less remote in the chains of causation leading to Y, (b) other variables included in the prediction equation are also causes of Y, and (c) all relevant variables are included. In statistical terms, cause and effect are similar to independent and dependent variables. Assuming causal relationships, the p coefficients in (4) could be defined as Wright's path coefficient. In applying predictor regression to path analysis, it is customary to deal with the relationships among the standardized variables all equal in mean (zero) and all equal in variance (unity).

Assume feed intake (F) and outside temperature (T) in dairy cows as independent variables and milk yield (M) as dependent. In conventional stepwise multiple regression, F and T are assumed to have a symmetrical relationship. It is also assumed that F and T have an asymmetrical relation with M. The three variables produce six possible bivariate asymmetrical path relationships shown below:

$(1) F \to M$ $(2) F \to T$ $(3) T \to M$ $(4) M \to F$ $(5) T \to F$ $(6) M \to T$

Of these relationships, paths numbers (2) and (6) above are unlikely. Assuming that the sequence of events provide natural association among variables, if F preceded M, then # (4) above is not a causal path. The remaining three relationships (Figure 3) have a polarity based on the assumption that a relationship between each set of two variables exists in only one of two possible directions.
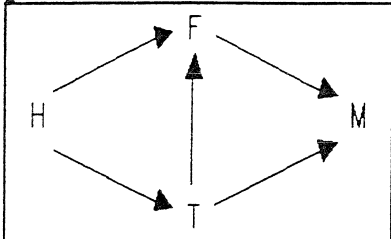

Figure 3. Path diagram relating milk production (M) to temperature (T) and feed intake (F).

If we further assume that F and T changes are influenced by changes in humidity (H), then Figure 4 will represent the path diagram. Further extension to multivariate could be made by assuming that changes in F and T change body weight (W) which in turn influences milk production (Figure 5). E is a random
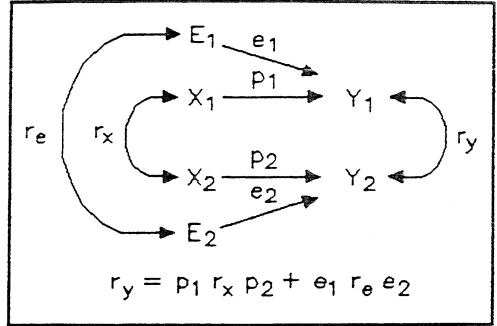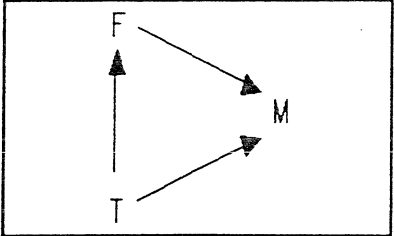

Figure 4. Path diagram relating milk production (M) to temperature (T), feed intake (F), and humidity (H).

variable representing unaccounted for variables. More detailed information is given in Kempthorne (1957), Wright (1968), and Li (1975).

Simple linear regression of Y on X

In case of the simple linear regression using model (7),

$$Y=\beta_0 + \beta_1 x + e \qquad (7)$$

Figure 5. Path diagram relating milk production (M) to temperature (T), feed intake (F), humidity (H), and body weight (W).

the residual random variable e is assumed to have zero expectation and uncorrelated with x. After standardization, Model (8) will replace Model (7) as:

$$Y = p_1 X + p_e E \qquad (8) \qquad \text{where,}$$

$$Y=(y - \overline{y})/\sigma_y \ , \quad X=(x - \overline{x})/\sigma_x \ , \quad E=e/\sigma_e.$$

The p_1 (p_{YX}) in Equation (9) is standardized partial regression coefficient (path coefficient) to be estimated from the transformed (standardized) data as:

$$p_1=\beta_1 \sigma_x/\sigma_y \qquad (9)$$

Where, $\sigma^2_Y=p_1^2\sigma_X^2+p_e^2\sigma_E^2=1$ or $p_1^2+p_e^2=1$ which divide the total

standardized variation of the Y variable into square of the paths (Wright's coefficient of determination) connecting X and E to Y. The Normal Equations for untransformed (concrete) and standardized variables are:

Concrete

Standardized

$$\begin{bmatrix} b_0 n & + b_1\sum x & = \sum y \\ b_0\sum x & + b_1\sum x^2 & = \sum xy \end{bmatrix} \qquad \begin{bmatrix} 0 + & p_1\sum X & = \sum Y \\ 0 + & p_1\sum X^2 & = \sum XY \end{bmatrix}$$

In standardized form $b_0=0$ and p_1 replaces b_1. We have:

$$\sum X=0, \sum Y=0 \ , \ p_1 = \sum XY/\sum X^2 = \sum XY = Cov(X,Y) = r_{XY} = r_{xy} \ ,$$

$$p_e = \sqrt{1 - p_1^2}$$

In this case, regression, correlation, and covariance are all measures of the same thing. Path diagram and square of the path coefficients (Wright's coefficients of determination) are shown in Figure 6.

When X and E are correlated as in P = G + E where, G and E represent genetic and environmental effects, respectively and P is the phenotypic (observed) representation, the standardized model can be written as:
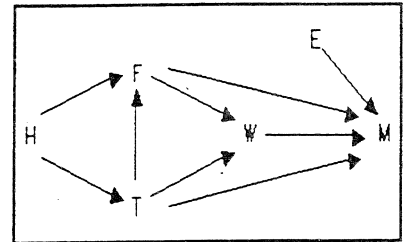
$P = hG + eE$, $\sigma_p^2 = h^2 + 2hre + e^2 = 1$

where h and e represent paths from G to P and from E to P, respectively. The additional term, 2hre, indicates indirect contributions due to correlation between G and E. This correlation is shown by a doubleheaded curved arrow in Figure 7.

Figure 6. Path diagram assuming X and E are not correlated.

Regression of y on x_1 and x_2

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \qquad (10)$$

The residual random variable e has zero expectation and is assumed uncorrelated pairwise with each of x_1, and x_2.

The standardized model is shown in Equation (11).

$$Y = p_1 X_1 + p_2 X_2 + p_e E \qquad (11)$$

where,

$Y = (y - \overline{y})/\sigma_y$, $X_i = (x_i - \overline{x})/\sigma_{x_i}$, i=1,2,

Figure 7. Path diagram assuming E and G are correlated.

and $E = e/\sigma_e$. The p's are as in Equation (5) and are standardized partial regression coefficients (path coefficients) to be estimated from the transformed (standardized) data. The Normal Equations are shown below:

$$\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} r_{1Y} \\ r_{2Y} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} p_1 & + p_2 r_{12} & = r_{1Y} \\ p_1 r_{21} & + p_2 & = r_{2Y} \end{bmatrix}$$

The solution to these equations results in:

$$p_1 = \frac{(r_{1Y} - r_{12} r_{2Y})}{1 - r_{12}^2} \quad \text{and} \quad p_2 = \frac{r_{2Y} - r_{12} r_{1Y}}{1 - r_{12}^2} .$$

The correlation coefficient r_{12} between X_1 and X_2 shows the degree of dependence in the normal equations and if equal to ± 1, then solutions for p_1 and p_2 can not be derived. When a large number of p's are involved, then normal equations should be solved by a computer. The path diagram and coefficients of determination for the case when X_1 and X_2 are correlated are shown in Figure 8. When variables are not correlated:

Figure 8. Path diagram and Wright's coefficient of determination for two correlated independent variables.

$r_{12} = 0$ and $p_1^2 + p_2^2 + p_e^2 = 1$.

Computations for determining p_1 and p_2 are simple if SAS (1985) is used using the following statements:
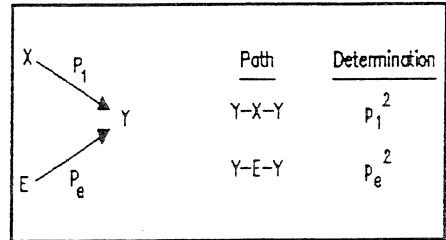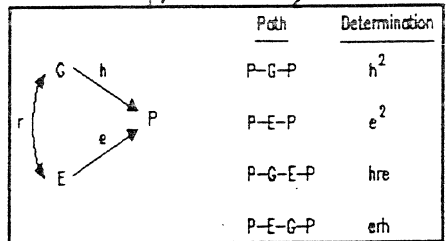
PROC REG; MODEL Y=X_1 X_2/STB;

The **STB** option of the **REG** Procedure will directly compute the standardized regression coefficients (path coefficients). The path from E to Y can be computed from: $p_e = \sqrt{1 - (p_1^2 + p_2^2)}$.

3. APPLICATION in AGRICULTURAL RESEARCH

Inbreeding

Wright's most successful use of path analysis method was in deducing genetic consequences of a continued inbreeding system such as brother-sister mating. Details of this application have been given in several publications including Wright (1921a, 1921b, 1934, 1968), Tukey (1954), Kempthorne (1957), and Li (1975) and will not be repeated here. However, an inbreeding example will be presented to transfer some idea of the usefulness of the path analysis in agricultural research. An alternative method based on properties of the mixed-model methods (Henderson, 1975, 1976) will also be presented to determine relationships among pedigree members.

As an example consider the pedigree in Figure 9 to illustrate construction of a relationship matrix (R). This matrix is ordered by generation number (0-4) so that parents precede their progeny. The assumption made here is that individuals of the base generation (A, B, and C in generation 0) are unrelated and not inbred. The identity of one of the parents of G is not known but is assumed to be unrelated to other individuals in generations 0 to 2.

Individuals F, and H in Figure 9 are inbred. The base generation individuals are noninbred unrelated and with unspecified parents. The pedigree in Figure 9 includes 8 individuals and thus R will be an 8X8 matrix of relationships. Henderson (1975, 1976) and Kennedy and Sorensen (1988) have discussed the method of constructing "Relationship" matrix which will be followed here considering the pedigree

Figure 9. Path diagram of individuals A-H.

in Figure 9. A lower triangular matrix T (8x8) and a diagonal matrix D (also 8x8) can be constructed with the following properties: (1) Diagonal elements of T are 1 and upper off-

	Parents			A-B	B-C	D-E	F-	F-G
	A	B	C	D	E	F	G	H
T=	1	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0
	0	0	1	0	0	0	0	0
	.5	.5	0	1	0	0	0	0
	0	.5	.5	0	1	0	0	0
	.25	.5	.25	.5	.5	1	0	0
	.125	.25	.125	.25	.25	.5	1	0
	.1875	.375	.1875	.375	.375	.75	.5	1

diagonal elements are all equal to 0. (2) The lower off-diagonal elements are equal to the average of the coefficients of parents (P1 and P2) if both parents are known, or half of the coefficient of the known parent if only one parent is known. The coefficient is 0, otherwise.

For instance, the element $T_{6,1}$ is half of the coefficients for parents of F, namely .5 for D and 0 for E which equals .25. $T_{7,1}$ is half of this value (.25) since only one parent is known (.125). This method of computation is well adapted to computerization. However, the path coefficient method could also be used to compute any of these coefficients. For example, $T_{6,2}$ indicates the relationship between F and B in Figure 9. There are two paths connecting F and B; B-D-F and B-E-F and the coefficient of BF is the sum of probabilities over these two paths or $(1/2)^2 + (1/2)^2 = .5$. These pathways represent the direct transfer of genes from B to F and each equals 1/2 since parents transfer a sample half of their genes to their offspring.

The diagonal elements of matrix D are equal to $(2 - F_{P1} - F_{P2})/4$ if both parents (P1 and P2) of an individual are known, equal to $(3 - F_{P1})/4$ or $(3 - F_{P2})/4$ if only one parent is known, or equal to 1 if neither parent is known. F indicates inbreeding coefficient of the parent. Both upper and lower off-diagonal elements of this matrix are equal to 0. Now the diagonal elements of the D matrix can be computed as {1, 1, 1, .5, .5, .5, .71875, .46875}. The first 3 diagonal elements of D matrix ($D_{1,1}$, $D_{2,2}$, and $D_{3,3}$) are equal to 1 since neither parents of A, B, or C are identified. $D_{4,4}$, $D_{5,5}$, and $D_{6,6}$ are .5 because both parents of D, E, and F are identified but none is inbred. The computation of $D_{7,7}$ element is based on only one parent identified for G which is inbred (F). Both parents of F descended from a common ancestor (B) and are connected by two paths B-D and B-E or 1/2 each. Thus the inbreeding coefficient of the individual F in Figure 1 is $(1/2)^{2+1} = .125$. $D_{7,7} = (3 - .125)/4 = .71875$. Since both parents of H are identified but only one (F) is inbred, $D_{8,8}$ is computed as: $(2 - .125 + 0)/4 = .46875$.

R is a symmetric matrix shown below and is readily computed from **R=TDT'** where **T** is the triangle and **D** the diagonal matrix. All diagonal elements of this matrix except $D_{6,6}$ and $D_{8,8}$ are equal to 1. $D_{6,6}$ indicates coefficient of inbreeding of .125 for individual F in Figure 9. The corresponding coefficient for individual H is computed to be .28125.

Path analysis has been used extensively in agricultural research, but only a few examples will be cited here. Wright applied path coefficient method to investigate the cause of negative regression of guinea-pig birth weight on litter size (Wright, 1921a, 1968). This example has been further

$$R = \begin{bmatrix} 1 & 0 & 0 & .5 & 0 & .25 & .125 & .1875 \\ 0 & 1 & 0 & .5 & .5 & .5 & .25 & .375 \\ 0 & 0 & 1 & 0 & .5 & .25 & .125 & .1875 \\ .5 & .5 & 0 & 1 & .25 & .625 & .3125 & .46875 \\ 0 & .5 & .5 & .25 & 1 & .625 & .3125 & .46875 \\ .25 & .5 & .25 & .625 & .625 & 1.125 & .5625 & .84375 \\ .125 & .25 & .125 & .3125 & .3125 & .5625 & 1 & .78125 \\ .1875 & .375 & .1875 & .46875 & .46875 & .84375 & .78125 & 1.28125 \end{bmatrix}$$

discussed by several authors including Niles (1922), Tukey (1954), and Kempthorne (1957). Dewey and Lu (1959) applied path analysis to investigate direct and indirect contributions of seed size, spikelets per spike, fertility, and plant size (independent variables) of crested wheatgrass to the variability in seed yield (dependent variable). A similar approach was used by Bhatt and Reddy (1981) to study direct and indirect effects of seven attributes on bean yield/plant. Other recent examples are analyses of carcass composition (Bennett, 1989), milk yield (Erb et al., 1985; Curtis et al., 1985), and blueberry yield (Siefker and Hancock, 1986). These examples show wide applications of path analysis in various fields of agricultural research.

4. DISCUSSION

Path analysis involves three components: (1) a path diagram representing a set of linear equations showing variables and their relationships, (2) a set of tracing rules to compute correlations among variables using a simplified algebra, and (3) algorithm to estimate parameters from the data and to test the adequacy of the model. In general, the algebra required for path analysis is very similar to those of regression, correlation, and ANOVA, except that variables are standardized in path analyses. One advantage over the conventional statistical method is that the algebra and tracing rules have been simplified so that even people with very little statistical training could perform path analyses.

In the method of path analysis causal relationships between variables are based on observation, experience, and common sense rather than experimental evidence. Interpretation of observed events may not be as clear-cut as those of the designed experiment and if a large number of variables are involved (which usually is the case) understanding causal relations among them could become very difficult. Of course one could use the stepwise regression to reduce the number of variables. Additionally, two scientists given the same data on a set of variables, could very well propose two different path diagrams for analysis. With several possibilities to analyze the same data, it would be difficult to judge which is right and which is not. In this

case, one's intention should be to select a system of causal relations compatible with the observed data and to change the relations if necessary to compare the differences between systems.

The method itself has come under the criticism by several scientists. Use of the "coefficient of determination" when variables are correlated has been criticized. However, when sources are independent, there is little difficulty in this respect. Partitioning the correlation between variables is an important accomplishment of the path analysis, but undesirable in ANOVA. Some reports favor the use of regression over correlation in path analysis. In this regard, if prediction is the purpose of the analysis, regression should apply. The path coefficients are closely related to correlation coefficients which are standardized covariances and may be further decomposed into causal components in a path analysis. However, path coefficients are not restricted to values between -1 and +1 as are correlations and they are not symmetric.

Concerns have been expressed on the assumption of linearity and the validity of path analysis under non-linear conditions. Wright's response has been to use transformation if non-linearity is suspected. Even in the case of non-linearity, after transforming variables, the path method would be a good approximation.

5. ACKNOWLEDGEMENTS

Technical assistance of T. Kim Giddens of USDA-ARS in graphical presentation of this manuscript is appreciated.

6. REFERENCES

Bennett, G. L. 1989. Path analysis and robust prediction of lamb carcass composition. *Anim. Prod.* 48:139-148.

Bhatt, D. and Reddy, T. P. 1981. Correlations and path analysis in castor (*Ricinus communis*). *Can. J. Genet. Cytol.* 23: 525-531.

Blalock, H. M. 1964. Causal Inferences in Non-Experimental Research. *University of North Carolina Press*, Chapel Hill.

Borgatta, E. F. and Bohrnstedt, G. W. (Eds.). 1969. Sociological Methodology. *Jossey-Bass*, San Francisco.

Cloninger, C. R., Rao, D. C., Rice, J., Reich, T., and Morton, N. E. 1983. A defense of path analysis in genetic epidemiology. *Am. J. Hum. Genet.* 35: 733-756.

Curtis, C. R., Erb, H. N., Sniffen, C. J., Smith, R. D., and Kronfeld, D. S. 1985. Path analysis of dry period nutrition, postpartum metabolic and reproductive disorders, and mastitis in Holstein cows. *J. Dairy Sci.* 68:2347-2360.

Dewey, D. R. and Lu, K. H. 1959. A correlation and path-coefficient analysis of components of crested wheatgrass seed production. *Agron. J.* 51: 515-518.

Duncan, O. D. 1966. Path analysis: Sociological examples. *Amer. J. Soc.* 72: 1-16.

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

Emigh, T. H. 1984. Path analysis, correlation, and the analysis of variance. In: *Experimental Design, Statistical Models, and Genetic Statistics*. Eds. Hinkelmann, K., pp.213-230, Marcel Dekker, Inc. New York.

Erb, H. N., Smith, R. D., Oltenacu, P. A., Guard, C. L., Hillman, R. B., Powers, P. A., Smith, M. C., and White, M. E. 1985. Path model of reproductive disorders and performance, milk fever, mastitis, milk yield, and culling in Holstein cows. *J. Dairy Sci.* 68: 3337-3349.

Fulker, D. W. 1988. Genetic and cultural transmission in human behavior. In: *Proc. 2nd. Int. Conf. Quant. Genet.* Eds. Weir, B. S., Eisen, E. J., Goodman, M. M., and Namkoong, G., pp.318-340, Sinauer Associates, Inc., Sunderland, MA.

Goldsmith, J. R. 1977. Paths of association in epidemiological analysis: application to health effects of environmental exposures. *Int. J. Epidemiol.* 6: 391-399.

Goldsmith, J. R. and Berglund, K. 1974. Epidemiological approach to multiple factor interactions in pulmonary disease: The potential usefulness of path analysis. *Annals N. Y. Acad. Sci.* 221: 361-375.

Heise, D. R. 1969. Problems in path analysis and causal inference. In: *Sociological Methodology.* Eds., Borgatta, E. F. and Bohrnstedt, G. W., pp. 38-73. Jossey-Bass, San Francisco.

Henderson, C. R. 1975. Rapid method for computing the inverse of a relationship matrix. *J. Dairy Sci.* 58: 1727-1730.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32: 69-84.

Kang, K. M. 1971. A review of S. Wright's method of path analysis. Honours Essay, *Dept. Statist., S.G.S., Australian National University*, Canberra.

Kang, K. M. and Seneta, E. 1980. Path analysis: an exposition. In: *Developments in Statistics,* VOL. 3, Krishnaiah, P. R., Ed. , PP. 217-246, Academic Press.

Kennedy, B. W. and Sorensen, D. A. 1988. Properties of mixed-model methods for prediction of genetic merit. In: *Proc. 2nd. Int. Conf. Quant. Genet.* Eds. Weir, B. S., Eisen, E. J., Goodman, M. M., and Namkoong, G., pp. 91-103, Sinauer Associates, Inc., Sunderland, MA.

Kempthorne, O. 1957. An Introduction to Genetic Statistics. *John Wiley & Sons, Inc.*, New York.

Land, K. C. 1969. Principles of path analysis, In: *Sociological Methodology.* Eds., Borgatta, E. F. and Bohrnstedt, G. W., pp. 3-37. Jossey-Bass, San Francisco.

Li, C. C. 1975. Path Analysis - A Primer. *The Boxwood Press*, Pacific Grove, CA.

Lush, J. L. 1947. Family merit and individual merit as bases for selection: I. *Amer. Naturalist* 81: 241-261.

Niles, H. E. 1922. Correlation, causation and Wright's theory of "path coefficients". *Genetics* 7: 258-273.

Niles, H. E. 1923. The method of path coefficients - an answer to Wright. *Genetics* 8: 256-260.

Rao, D. C., Morton, N. E., and Yee, S. 1976. Resolution of cultural and biological inheritance by path analysis. *Am. J. Hum. Genet.* 28: 228-242.

SAS Institute, Inc. 1985. SAS User's Guide: Statistics, Version 5 Edition, *SAS Institute Inc.* Cary, NC.

Seneta E. 1972. Population projection variances and path analysis. *J. Amer. Statist. Assoc. 67: 617-619.*

Siefker, J. H. and Hancock, J. F. 1986. Yield component interactions in cultivars of the highbush blueberry. *J. Amer. Soc. Hort. Sci.* 111: 606-608.

Tukey, J. W. 1954. Causation, regression, and path analysis. In: *Statistics and Mathematics in Biology.* Eds., Kempthorne, O., Bancroft, T. A., Gowen, J. W., and Lush, J. L. pp. 35-66, Iowa State University Press, Ames.

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

Wright, S. 1921a. Correlation and causation. *J. Agric. Res.* 20:557-585.

Wright, S. 1921b. Systems of mating. II. The effect of inbreeding on the genetic composition of a population. *Genetics* 6: 124-143.

Wright, S. 1923. The theory of path coefficients - a reply to Niles's criticism. *Genetics* 8: 239-255.

Wright, S. 1934. The method of path coefficients. *Ann. Math. Stat.* 5:161-215.

Wright, S. 1954. The interpretation of multivariate systems. In: *Statistics and Mathematics in Biology,* Eds., Kempthorne, O., Bancroft, T. A., Gowen, J. W., and Lush, J. L., pp. 11-33. Iowa State University Press, Ames.

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

Wright, S. 1960b. The treatment of reciprocal interaction, with or without lag, by path analysis. *Biometrics* 16: 423-445.

Wright, S. 1968. Evolution and the Genetics of Populations. Vol. 1, *University of Chicago Press*, Chicago.