DOCUMENT RESUME

ED 060 132 24 TM 001 324

AUTHOR Pelz, Donald C.; Faith, Ray E.
TITLE Causal Connections in Educational Panel Data. Final Report.
INSTITUTION Michigan Univ., Ann Arbor. Survey Research Center.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-9-0459
PUB DATE Aug 71
GRANT OEG-5-9-239459-0076
NOTE 15p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Correlation; *Critical Path Method; Data Analysis; Educational Research; Evaluation Techniques; Individual Characteristics; Mathematical Applications; *Mathematical Models; *Multiple Regression Analysis; *Predictive Measurement; Predictor Variables; Testing; Time

ABSTRACT

This final report summarizes past research and suggests new approaches to the problem of estimating long-term individual constants using path analysis. The general objective of the research was to detect and measure the likelihood that one variable, x, measured at time, t, has a causal influence on another variable, y, measured at a subsequent time t + k (k being the measurement interval). Work on this problem continues under a grant from the National Science Foundation. (CK)

FINAL REPORT

Project No. 9-0459

Grant No. OEG-5-9-239459-0076

CAUSAL CONNECTIONS IN EDUCATIONAL PANEL DATA

Donald C. Pelz and Ray E. Faith

Survey Research Center
Institute for Social Research, P.O. Box 1248
The University of Michigan
Ann Arbor, Michigan 48106

August, 1971

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

FINAL REPORT

Project No. 9-0459

Grant No. OEG-5-9-239459-0076

CAUSAL CONNECTIONS IN EDUCATIONAL PANEL DATA

Donald C. Pelz and Ray E. Faith

Survey Research Center
Institute for Social Research, P.O. Box 1248
The University of Michigan
Ann Arbor, Michigan 48106

August, 1971

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

CONTENTS

FIGURES & TABLES

Overview

The present report is a final one only in the sense of being the last written under the present grant from the U.S. Office of Education. Work on the same problem continues under a grant from the National Science Foundation, with the same general objective: given a set of panel data with the same variables remeasured at several intervals on a given set of individuals (such as measures of educational motivation and performance), it it possible (a) to detect and (b) to measure the likelihood that one variable x measured at time t has a causal influence on another variable y measured at a subsequent time t + k (k being the measurement interval)?

While some progress has been made, it is not sufficient as yet, in our view, to warrant a comprehensive technical treatment (recapitulation of objectives, methods, hypotheses, conclusions). Therefore the following section will summarize what has been presented in interim reports, to which the reader is referred for further detail.

Additional work is then reported, using techniques of path analysis, on the problem of predicting autocorrelations and lagged cross-correlations in a two-variable system in which causal influence is exerted in either direction over several causal intervals.

A simple method is described and illustrated for recovering the underlying path coefficients (including causal coefficients) from regression analysis of the auto- and cross-correlations.

A final section suggests a new approach, using path analysis, to the problem of estimating long-term individual constants. Future steps are indicated.

Summary of work from September 1969 through December 1970

We began by obtaining two data banks with repeated measurements: height, weight, and grip of 100 boys and 100 girls measured at 6-month intervals between ages 6 and 9; and 3,000 students measured by the Educational Testing Service in grades 5, 7, 9, and 11 with complete data for four years on test batteries SCAT and STEP (aptitude and performance measures) and for three years on BEQ (questionnaire items on interest and behavior).

Previous work with two-variable simulated time series (Pelz, Magliveras, and Lew, 1968) suggested that the appearance of causal connections between two variables would be obscured by the tendency for each individual to remain relatively stable on each variable. Such stable long-term trends may be conceptualized as individual constants around which short-term disturbances occur. The extremely high correlations among successive height and weight scores suggested the presence of such individual constants, rising of course with time with

each individual retaining the same relative position.

Several months of effort were spent in devising a means for estimating such a constant for each individual so that it could be removed. The first progress reports under this grant (see list of references, Pelz 1969-70) describe these efforts, particularly #3 for April-June 1970, which deals in depth with an attempt to separate empirical height and weight data into stable long-term trends and relatively unstable but autocorrelated short-term disturbances. A relatively complex computerized method is given for making such a separation, subject to some restrictive assumptions placed on the underlying model as well as on the number of time periods at which measurements have been taken.

The method appears conceptually sound. However, it assumes that the variable is causally independent; we have not yet devised a procedure relevant for a dependent variable. Furthermore, when residual scores were obtained by subtracting the estimated individual constants at each time period, cross-correlations among the residuals did not lend themselves to a simple interpretation of causal influence among height, weight, and grip.

Accordingly the last six months of 1970 were spent in a different approach, as described in progress reports #4 and #5, in which we moved away from consideration of empirical data and explored mathematical models of hypothetical causal structures. This work resulted in a paper (Pelz and Faith, 1970) presented at the American Statistical Association meetings in Detroit, December 27, 1970. We coupled the methods of path analysis with those of matrix theory to give a more compact form of the two-variable unidirectional causal scheme, and thereby greatly simplified the derivation of the correlational properties of such a hypothetical model. A detailed exposition of this is to be found in the technical appendix of the ASA paper.

Reciprocal causation and distributed lags

After completion of the ASA paper, our attention turned to somewhat more complex situations--specifically those in which the causal influences between two variables x and y were reciprocal (i.e. $x \longrightarrow y$ and $y \longrightarrow x$) and in which these causal influences were exerte not over one single time interval but rather were distributed over several time periods. We found that, with slight modifications, the methods employed in studying the simple model described above were applicable for these more complex situations. The basic structure of the problem turned out to be very much the same, and although it is more complex, the problem is of no greater depth.

We define the model as follows. Corresponding to the recursive relations (3) and (4) in the appendix of Pelz and Faith (1970) are:

$$(1) \qquad x_t = p_{xx}x_{t-1} + \sum_{g=1}^{G} p_{xy_g}y_{t-g} + p_{xu}u_t \qquad \text{and}$$

$$(2) \qquad y_t = p_{yy}y_{t-1} + \sum_{g=1}^{G} p_{yx_g}x_{t-g} + p_{yv}v_t \quad .$$

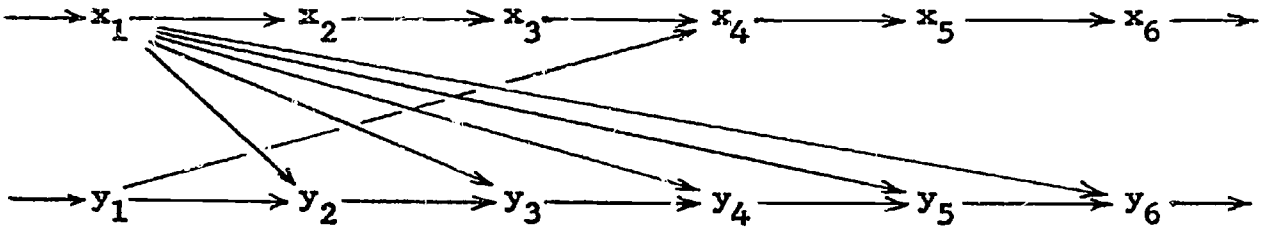
To get this model started requires the inclusion of the correlated inputs x_s and y_s, where s takes on the values -1, -2, ..., -G. As before, this model is stationary with respect to translation in the time domain if the correlations between the inputs satisfy certain conditions, which in this case take the form of a system of linear equations. Solving this system permits one to determine all of the correlations of the more complex model, since all of these are determined by the input correlations.

As in the simpler case, it is possible to generate a theoretically impossible model, if the path coefficients are too large. A simple test was derived, similar to that used in the proof of (15b) in the appendix, which will notify one of this situation if it arises.

Using the above procedure we have in fact been successful in generating theoretical correlations for the two-way distributed-lag model (see Pelz, Magliveras, and Lew, 1968) using the same parameters.

Illustration with two examples

Given below are two sample test runs of our procedure. In each of them the influence is exerted in both directions. In the first example the two variables are identical except that the influence of x on y is distributed over 5 lags (i.e. intervals of 1, 2, 3, 4, and 5 time units) whereas the influence of y on x is concentrated at one lag (3 time units), the total amount of influence being the same in each case. A partial diagram of the path model (many of the paths being omitted for simplicity) is:



The parameters of the first model are:

$$p_{xx} = p_{yy} = .7$$

$$p_{xy_i} = \begin{cases} .25 \text{ for } i = 3 \\ 0. \text{ for } i \neq 3 \end{cases}$$

$$p_{yx_i} = .05 \text{ for } i = 1,2,3,4,5; \quad \sum = .25$$

The latter notation may be read: "influence on y exerted by x over time lags of 1, 2, ..." The auto-and cross-correlograms resulting from this model are platted in Figure 1.

Note that in the left half of the cross-correlogram, the effect of concentrating the causal influence y⟶x at a single lag is to make the peak higher and sharper; in the right half, the effect of distributing the causal influence x⟶y over 5 lags is to make the peak lower and wider.

Figure 1 here

In the second example, the autoregressive path coefficient p_{yy} is considerably higher than the autoregressive coefficient p_{xx} for the other variable. Each variable now influences the other over five separate causal intervals, and the total influence of y on x is twice as great as the total influence of x on y. The parameters of this model are set as follows:

$$p_{xx} = .5 \qquad p_{yy} = .94$$

$$p_{xy_i} = .04 \text{ for } i = 1,2,3,4,5; \quad \sum = .20$$

$$p_{yx_i} = .02 \text{ for } i = 1,2,3,4,5; \quad \sum = .10$$

The resulting auto- and cross-correlations are plotted in Figure 2.

Figure 2 here

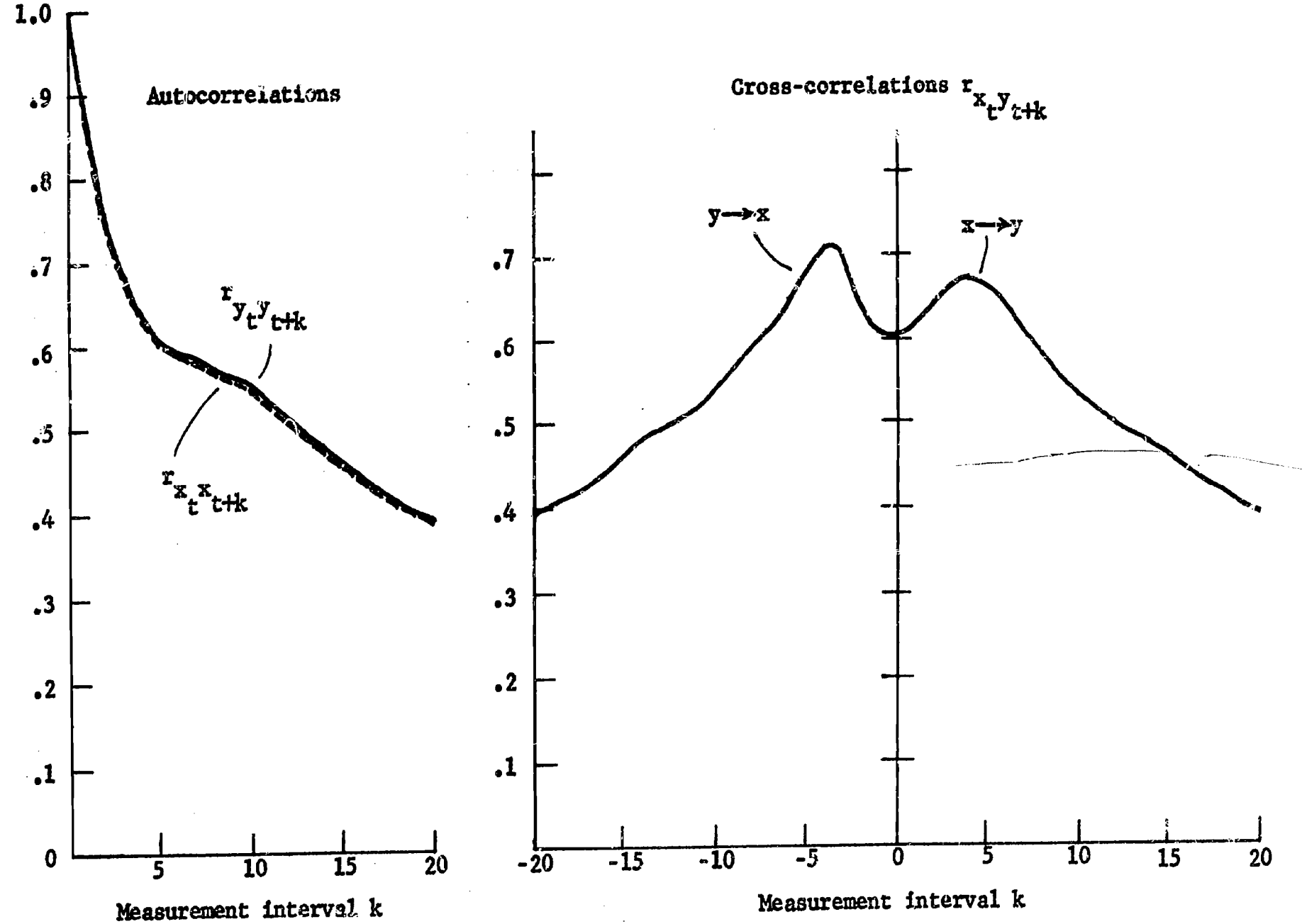


Figure 1. Auto- and cross-correlograms generated by "example 1" in text. Variables x and y each influence the other by the same amount, but the x→y influence is distributed over 5 lags, whereas the y→x influence is concentrated at one lag.

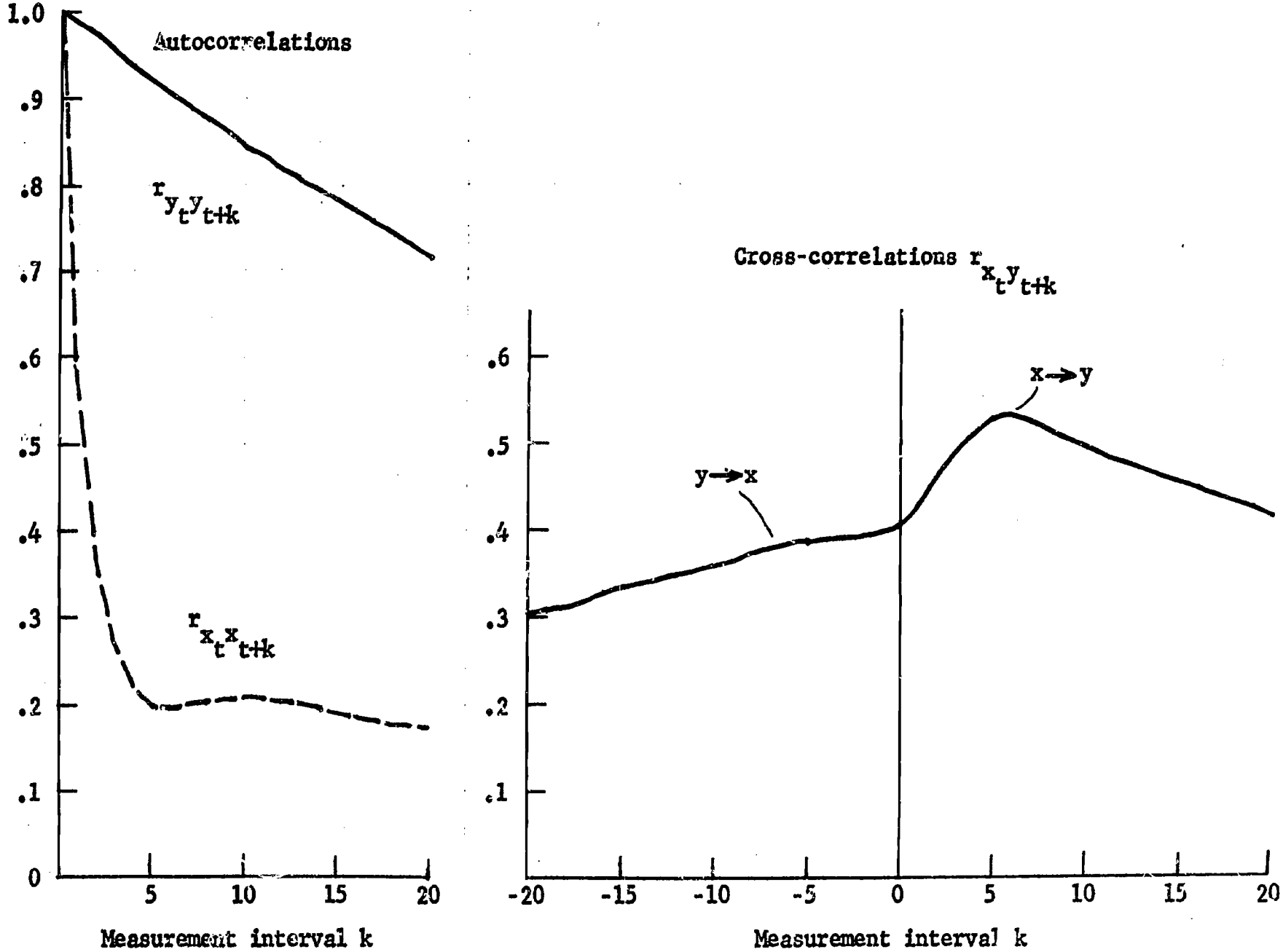


Figure 2. Auto- and cross-correlations generated by "example 2". The y→x influence is twice as strong as the x→y, but the correlogram for the latter is higher because of the high autoregressive stability of dependent variable y.

Since the causal influence y⟶x is twice that of x⟶y, one would normally expect the cross-correlogram to be higher in the left half than in the right half. As shown in Figure 2, this is not the case. The reason for this apparent paradox lies in the fact that the y variable, because of its extremely high autoregressive coefficient, has a high "memory" of past influences from x, which therefore accumulate over time and increase the cross-correlations. Variable x on the other hand, with a relatively small autoregressive coefficient, soon "forgets" past influences from y.

The "recovery" process

The apparent paradox in Figure 2 now raises an important question: is it possible to analyze the auto- and cross-correlograms in such a way as to recover the true values of the various path coefficients, showing that the causal effect of y⟶x is in fact stronger than x⟶y, despite the visual evidence to the contrary? It turns out that this objective can be easily accomplished by applying linear multiple regression to the correlational matrices, to express each of the variables as a function of previous variables. The regression (beta) coefficients for each predictor are then theoretically equivalent to the path coefficients between that predictor and the particular dependent variable. In the absence of extraneous factors such as measurement error or long-term stability in equations (1) and (2), the variables x and y are defined at each point in time by regression equations in which the predictors are the same variables measured at previous points in time.

As shown in Table 1, the resulting regression coefficients were very close to the values of the path coefficients used to specify the model.

Table 1. True path coefficients compared with regression coefficients based on theoretically derived correlations, for two examples.

	Example 1		Example 2	
Parameter	True value	Regression coefficient	True value	Regression coefficient
p_{xx}	.70	.70	.50	.50
p_{yy}	.70	.70	.94	.92
p_{xy_1}	.00	.00	.04	.048
p_{xy_2}	.00	.00	.04	-.009
p_{xy_3}	.25	.25	.04	.803
p_{xy_4}	.00	.00	.04	.051
p_{xy_5}	.00	.00	.04	.018
			$\sum$ = .20	$\sum$ = .209
p_{yx_1}	.05	.055	.02	.018
p_{yx_2}	.05	.045	.02	.023
p_{yx_3}	.05	.050	.02	.019
p_{yx_4}	.05	.053	.02	.019
p_{yx_5}	.05	.048	.02	.021
	$\sum$ = .25	$\sum$ = .251	$\sum$ = .10	$\sum$ = .100

In example 1 all of the path coefficients were estimated with rather high accuracy. There were slight errors in estimating the various p_{yx_i}, but the sum of these estimates (regression coefficients) was very close to the sum of the true values. Such errors arise from computational inaccuracies (such as rounding).

In example 2 some inaccuracies appeared, chiefly in estimating p_{yy} and p_{xy_i} (but note that the sum of the latter estimates was reasonably close to the sum of the true values). Hence it may be difficult to get accurate estimates of causal coefficients for a highly autocorrelated variable functioning as an independent variable. As mentioned previously, this situation gives rise to an oddly-shaped cross-correlogram.

Unresolved problem--estimating long-term tendencies

The chief unresolved problem for this project has been the one indicated in section 1 of the progress report for April-June, 1970--that is, the separation of variables into long-term and short-term components. In that report is explained how such a separation can be accomplished for each variable in isolation. Once this is accomplished, the short-term components of the variables are compared to determine causal inferences. Unfortunately such a two-stage procedure has the disadvantage that the two stages make different assumptions about the nature of the variables. The first stage treats each of them as simple autoregressive panels, each causally independent, whereas the second regards them as interacting with each other.

One solution is to find some way to estimate all parameters simultaneously. Work in this direction is not complete, although some progress has been made by formulating the problem in terms of path analysis, thereby reducing its complexity considerably. For example, a much simpler and more direct solution has been discovered for the problem of separating a single variable into short and long term components, by employing path analysis. Consider the following path model diagrammed in Figure 3, and defined by the equations:

$$(3) \quad x_t = p_{xx}x_{t-1} + p_{xu}u_t \qquad t = 1,2,\ldots$$

$$(4) \quad X_t = p_{Xxt}x_t + p_{Xz_t}z + p_{Xe_t}e_t \qquad t = 0,1,2,\ldots$$

Figure 3 here

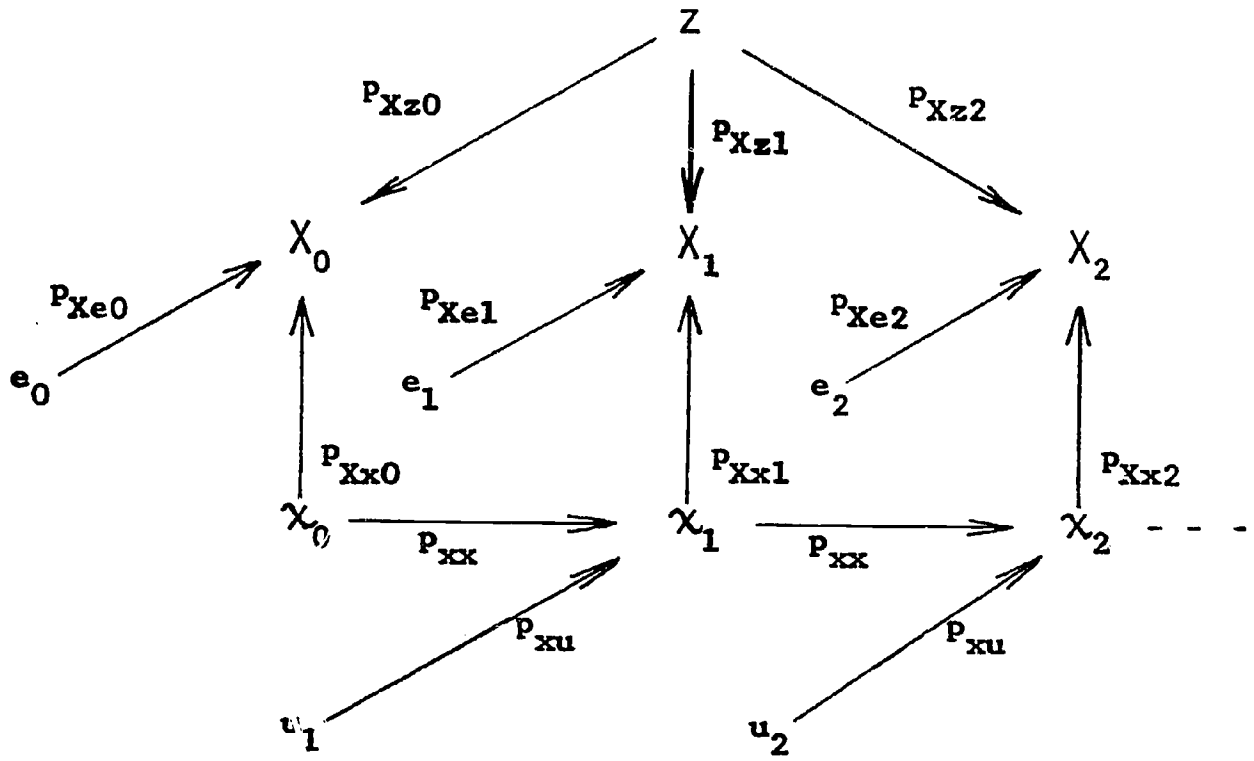


Figure 3. Path model of an autocorrelated variable with measurement error and long-term individual constants.

In Figure 3 the terms x_t represent true scores, X_t measured scores, e_t measurement error, and z represents the long-term constant for each individual. The terms x_0, e_0, z, u_t, and e_t $t = 1, 2, \ldots$ are uncorrelated inputs to the system, and the variables u_t are the disturbances in the short-term component x_t. The theoretical autocorrelations for this model are given by the relation

$$(5) \qquad r(X_s, X_t) = p_{xx}^{|t-s|} p_{Xx_s} p_{Xx_t} + p_{Xz_s} p_{Xz_t} ,$$

where s, t = any pair of times.

Since the correlations are non-linear functions of the path parameters, the problem of recovering these is a non-trivial one. In fact, if p_{xx} is one, it is impossible to solve uniquely for all the parameters, as may be seen by observing equation (5).

Even when this is not the case it is not known yet under precisely what conditions a unique solution is obtainable, and more work is required in this area. So far we are able to provide an estimate of the parameters if there are at least six measurements over time of the variable X_t. The uniqueness of this solution is presently open to doubt, as is its sensitivity to sampling errors in the empirical correlations.

The problem is simplified somewhat if some additional constraint is placed on the model. For example, if it is assumed that the influence of measurement error is constant over time--i.e. that $p_{Xes} = p_{Xet} = p_{Xe}$ for all s and t--then it appears that four measurements are sufficient to solve for the path parameters. Again, the existence and uniqueness of this solution requires further study.

Even if equations (3) and (4) can be solved for the path coefficients, however, the most interesting questions, those concerning the causal relations between the variable x and other variables, are left unanswered, since this model describes only one variable, independent of the outside causal factors. Although the solution to the more complex problem is not at hand, it would appear that the most direct approach to it would be by means of path analysis, because of the degree to which this simplifies the statement of most linear causal models.

Future steps

Work is continuing in the further generalization of the linear causal model described above and in the ASA paper (Pelz and Faith, 1970). In particular it appears possible to compute the correlations that will arise in a model consisting of an arbitrary number of variables which causally influence each other over an arbitrary number of causal intervals. The theoretical concepts involved are the same as those employed for the two-variable model with distributed lags and reciprocal causation. Only the mechanics of the computation are more elaborate, because of the great increase in the number of equations to be solved for as the number of variables in the model increases.

References

Pelz, D.C. with S. Magliveras and R.A. Lew. 1968. "Correlational properties of simulated panel data with causal connections between two variables." Interim Report #1, Causal Analysis Project, Survey Research Center, University of Michigan, Ann Arbor.

Pelz, D.C. and R.E. Faith. 1970. "Some effects of causal connections in simulated time-series data." Interim Report #2, Causal Analysis Project, Survey Research Center, University of Michigan, Ann Arbor. Published without technical appendix in Proceedings of the Social Statistics Section, Annual Meeting of the American Statistical Association, Dec. 27-30, pp. 32-39.

Pelz, D.C. and R.A. Lew. 1970. "Heise's model applied." In E.F. Borgatta and G.W. Bohrnstedt (eds.), Sociological Methodology--1970. San Francisco: Jossey-Bass.

Pelz, D.C. 1969-1970. "Causal connections in educational panel data: progress reports." Project No. 9-0459, Bureau of Research, U.S. Office of Education. #1 September-December 1969; #2 January-March 1970; #3 April-June 1970; #4 July-September 1970; #5 October-December 1970.