DOCUMENT RESUME

ED 173 203 SO 011 761

AUTHOR	Hannan, Michael T.
TITLE	Final Report for Dynamic Models for Causal Analysis of Panel Data. Methodological Overview. Part II, Chapter 1.
INSTITUTION	Center for Advanced Study in the Behavioral Sciences, Stanford, Calif.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.; National Science Foundation, Washington, D.C.
PUB DATE	Dec 77
GRANT	BNS-76-22943; NIE-G-76-0082
NOTE	29p.; For related documents, see SO 011 759-772
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Analysis; Measurement Techniques; *Models; *Research Design; *Research Methodology; *Research Problems; *Social Change; Social Science Research; Sociometric Techniques; *Validity

ABSTRACT

This technical document, part of a series of chapters described in SO 011 759, describes a basic model of panel analysis used in a study of the causes of institutional and structural change in nations. Panel analysis is defined as a record of state occupancy of a sample of units at two or more points in time; for example, voters disclose voting intentions in a sequence of surveys leading up to an election. The author first defends use of the model by comparing it with a more complex alternative. Three arguments for using the basic model are presented: that the basic model approximates the true causal structure, that the model relates levels of X and Y to changes in X and Y, and that the basic model is a reduced-form of the proper model which allows for causal effects over the study period. The next section identifies special methodological issues, outlines strategies for addressing them, and cites relevant technical literatures containing more detailed treatment. Issues include autocorrelation of disturbances (variables omitted from the study such as material infrastructure), technology, cultural organization, national history, the assumption of constant error variance, measurement errors, and functional forms of relationships. (Author/KC)

Final Report for

DYNAMIC MODELS FOR CAUSAL ANALYSIS OF PANEL DATA. METHODOLOGICAL OVERVIEW*

Michael T. Hannan

Center for Advanced Study in the Behavioral Sciences

December 1977

*This chapter is being published in National Development and the World System: Educational, Economic, and Political Change, 1950-1970, coedited with John W. Meyer. Preparation of this chapter was supported by National Institute of Education Grant NIE-G-76-0082 and by National Science Foundation Grant BNS76-22943 to the Center for Advanced Study in the Behavioral Sciences. I wish to thank Robert M. Hauser, John W. Meyer, Aage Sørensen, and Nancy Brandon Tuma for comments on an earlier draft.

Perspectives on Panel Analysis

Though sociologists have employed panel designs for some time, there is little agreement about the formulation and estimation of panel models. In part this lack of agreement reflects the varied purposes that sociologists bring to panel analysis. Our motivations for choosing this design were sketched in the previous chapter. The purpose of this chapter is to outline the class of models and estimators we use and our reasons for using them. To place our research in a broader methodological context, I contrast our procedures with certain widely used alternatives.

As we discussed in the previous chapter, our goal is to study the effects of various institutional sectors on each other. The literature on national development suggests that effects may run in all directions, that is, that feedback effects occur more often than not. For simplicity I treat the case where there are only two variables of concern, X and Y. We might think of school enrollments and national production, or national production and economic dependence, etc. Suppose further that the variables are measured at only two points in time, t_0 and t. Then the operative question is how to model the causal relations among the four variables, X_0, X_t, Y_0, and Y_t.

Suppose we had the ability to experiment. We would hold X_0 constant (through either randomization or nonvariation), introduce variation in Y_0 and observe the consequences for X_t. In another experiment we would similarly vary X_0 and observe consequences for Y_t. In each case we would examine the impacts of variation in one lagged variable on changes in the other variable. The nonexperimental model that we and most other social scientists use is a simple analogue to this pair of experiments. We write a two equation system in which variation in each outcome variable at time t is a linear function of its initial level and of the initial level of the other variable:

$$Y_t = \alpha_0 + \alpha_1 Y_0 + \alpha_2 X_0 + u \qquad (1)$$

$$X_t = \beta_0 + \beta_1 X_0 + \beta_2 Y_0 + v \qquad (2)$$

For shorthand, I refer to the model in (1-2) as the basic model.

Each equation states that the outcome depends on the pair of causally predetermined variables and on a disturbance that we assume is uncorrelated with the right hand side variables (but see below). Our interest focuses on α_2 and β_2. These parameters indicate the degree to which variations in one variable affect another, holding constant initial levels. If $\alpha_2 = 0$, we conclude that X does not affect Y over time; and if $\beta_2 = 0$, we conclude that Y does not affect X over time.

But this is just the usual structural equation or causal modeling perspective. The pair of equations, (1) and (2), embody assumptions both about the nature of the process and about the behavior of disturbing factors (summarized in u and v). Quality of inferences about the causal relations obviously depend on the accuracy of both types of assumption. There is no magic in panel analysis. Nor is there a special theory of statistical inference for panel analyses. Once the model is formalized in a set of structural equations, the methods used are identical to those used for structural equation models generally.

This perspective differs from that which informs much social scientific work on panel analysis. It is more common to propose specialized methods for panel inference. Most of this work follows the tradition begun by Lazarsfeld (1955, 1972), which poses the methodological problem as follows. Given that X and Y are correlated, construct measures of effect that permit one to conclude that X causes Y <u>or</u> Y causes X (but not both). Lazarsfeld and his students constructed indexes (for sixteenfold tables) that seek to answer this question.[1] This approach has been generalized to quantitative analysis in the form of cross-lagged correlation (see Campbell and Stanley 1963; Pelz and Andrews 1964; Kenny 1973).

Undoubtedly there are some situations in which X causes Y but not vice versa. But, in general, "does X cause Y <u>or</u> Y cause X?" is not the right question. As we argued in the previous chapter, for many macrosociological questions it is reasonable to begin with the assumption that X causes Y <u>and</u> Y causes X. This issue is not merely rhetorical. Attempting to choose either X <u>or</u> Y as the causal variable leads to testing procedures that are far from optimal when both variables have effects.[2] Thus we eschew this approach to panel analysis.

So we work squarely in the structural equation tradition. We presume familiarity with the logic and procedures of this approach at the level of Duncan (1975) or Namboodiri, Blalock, and Carter (1975). In several cases, we use more advanced methods. They are noted below and discussed in the chapters in which they arise. The remainder of this chapter is devoted to a consideration of alternative specifications of the relations among X_0, Y_0, X_t, and Y_t, and to various methodological issues that arise in our work. The purpose of this discussion is to place our work in a broader methodological context, to note limitations on our findings, and to suggest ways in which future researchers can improve on our work.

Why Not Use Change Scores?

In discussing our work with other sociologists we are often asked why we do not relate changes in X to changes in Y and vice versa. Addressing this question leads to consideration of issues that clarify the interpretation of estimates of the parameters of the model in (1-2). It will facilitate exposition to consider separately use of change scores as dependent variables and as causal variables. To keep the algebra simple we will consider only the Y-equation.

The first model we consider relates changes in Y over the period of observation to the initial levels of X and a disturbance, w:

$$\Delta Y_t = Y_t - Y_0 = \alpha'_0 + \alpha'_2 + w$$

But the model in (3) is just the special case of (1) where $\alpha_1 = 1$. So two things can happen. The restriction that $\alpha_1 = 1$ may be correct in which case estimation of (1) will give approximately that result. Then it does not matter whether one estimates (1) or (3). On the other hand, the restriction may be incorrect. In that case estimation of the change score model usually gives biased estimates of the effect of X_0 on Y_t. To see this note that by subtracting Y_0 from each side of (3) we obtain

$$Y_t - Y_0 = \alpha_0 + \alpha_1 Y_0 + \alpha_2 X_0 + u - Y_0$$

or

$$\Delta Y_t = \alpha_0 + \alpha_2 X_0 + q \qquad (4)$$

where $q = (\alpha_1 - 1) Y_0 + u$

Equation (4) has exactly the same form as the change score model except that the disturbance, q, contains Y_0. If X and Y are causally related over time, it will almost always be the case that they will be correlated cross-sectionally. If so, the disturbance, q, is correlated with X_0 and least squares estimates of (4) will be biased. The bias is zero only when the constraint is true, i.e. $\alpha_1 = 1$. The model in (1) avoids such bias.

What about changes in X over the period of observation? As long as the two-equation model in (1-2) is appropriate, both X and Y will change over the study period. In an experiment, we can study one process at a time. Lacking such controls, we cannot presume that X, say, is fixed at X_0 over the period we observe. To do so leads to an obvious inconsistency when we shift attention to the X_t equation. That is, if equation (2) models change in X over the period (t_0, t), we cannot maintain that X is constant in the equation for Y_t.

From the perspective of causal analysis the real issue is whether changes in X over the study period affect Y_t (holding constant Y_0 and X_0). This important question has far reaching implications as we shall see below. We can place the methodological problem in clear focus by writing a model that incorporates the effects of changes in each variable:

$$Y_t = \gamma_0 + \gamma_1 Y_0 + \gamma_2 X_0 + \gamma_3 (X_t - X_0) + u \qquad (5)$$

$$X_t = \delta_1 + \delta_1 X_0 + \delta_2 Y_0 + \delta_3 (Y_t - Y_0) + v \qquad (6)$$

For the moment we presume that the model in (5-6) is the true model. Note that the original model in (1-2) is a special case of this model with $\gamma_3 = 0$ and $\delta_3 = 0$.

It turns out to be more convenient to rewrite the model as follows:

$$Y_t = \gamma_0 + \gamma_1 Y_0 + \gamma_2' X_0 + \gamma_3 X_t + u' \qquad (7)$$

$$X_t = \delta_0 + \delta_1 X_0 + \delta_2' Y_0 + \delta_3 Y_t + v' \qquad (8)$$

where

$$\gamma_2' = \gamma_2 - \gamma_3 ; \delta_2' = \delta_2 - \delta_3 .$$

This new model has an obvious advantage over the basic model; it permits causal effects over lag periods shorter than (t_0, t). But it introduces a new methodological problem. While the original model was recursive, the model in (7-8) contains simultaneous causation. Both X_t and Y_t appear as both independent and dependent variables in a pair of equations. So we have lost an important type of simplicity. More important, the model in (7-8) is not identifiable (to use the language of econometrics). Its parameters cannot be uniquely estimated from any data set.

The literature on simultaneous equations estimation commonly advises two procedures for repairing underidentified models (see Johnston 1971: Ch. 12 for example). One can use theory or prior research to place constraints on the parameters of the model. Unfortunately we cannot avail ourselves of this strategy as we lack such theory and evidence. The second alternative involves the use of instrumental variables. The model in (7-8) can be identified by finding a pair of variables that are uncorrelated with u' and v' such that one of them affects Y_t but not X_t directly and the other affects X_t but not Y_t. Inserting these variables into estimators will resolve the indeterminacy of the model. But again we find ourselves blocked by the lack of prior theory. As the literature argues that almost everything affects everything else, it is hard to argue authoritatively a priori that some variables behave as instruments (that is they affect only some of the variables in the model).

In the long run, the resolution of the problem under discussion will likely involve these sorts of strategies. Our assessment was that the field was not sufficiently well developed for us to rely on theory and prior research for the strong assumptions that must be used to motivate the simultaneous equations methods. Thus we decided to settle for simpler methods.

Our situation is as follows. We propose to use the model in (1-2) to draw inferences about institutional and structural change in nations. However, we realize that this model is not completely general and that strong arguments can be made for the superiority of the model in (7-8). So we must defend the choice of the basic model over the more complex alternative. I just argued that the more complex model cannot be uniquely estimated as it stands. But that is not a justification for the simpler model. Are there any general arguments that favor the model in (1-2)? I will advance three.

One line of argument claims that changes during the study period do not have effects, that is, the basic model is correct. This amounts to arguing that all the causal effects have a lag at least as long as twenty years. But, it is easy to show that if the lags are longer than twenty years we will not observe systematic effects with this model. Therefore, this position requires that one know exactly the structure of the lags in the effects. This sort of reasoning motivated Heise's (1969) treatment of panel analysis.

This position has some appeal. Some social scientists argue that twenty years is too short a period over which to observe inter-institutional effects. Readers who take this view will not be troubled by our neglect of effects with lags shorter than twenty years. However, we observe that some institutions, notably political structures, change literally over night. Such changes can have strong and almost immediate effects on other institutions. So we cannot always rest assured that twenty years is a conservatively short period that rules out effects of shorter lag.

We find another aspect of this position disturbing. It presumes that the process operates over discrete time intervals. In formal terms the model implied is a difference equation with a twenty year lag. I prefer the view that adjustments to institutional changes are continuous in time and not specialized to any particular causal lag. What happens if we model the process in this fashion, that is as a continuous time process? Our general argument presumes that changes in X and Y depend on the levels of both variables. The simplest continuous time model that is consistent with this argument is the linear system:

$$\frac{dY(t)}{dt} = a_0 + a_1Y(t) + a_2X(t) \quad (9)$$

$$\frac{dX(t)}{dt} = b_0 + b_1X(t) + b_2Y(t) \quad (10)$$

But the differential equations in (9-10) cannot be estimated directly. Instead we solve the system subject to appropriate initial conditions (see Coleman 1968). The solution to this system has exactly the same form as our basic model, (1-2). That is, it is a two-equation model in which X_t depends linearly on X_0 and Y_0, and Y_t depends linearly on Y_0 and X_0. The parameters of the panel model can thus be considered as nonlinear functions of the parameters of the differential equation model (see Coleman 1968; Dorian and Hannan 1976; Hannan and Freeman 1978).

This relationship may surprise readers trained in classical panel analysis traditions. It shows that the basic model is less restrictive than conventional treatments imply. Despite the fact that it ignores effects of ΔX_t and ΔY_t, it turns out to be an implication of a model in which X and Y adjust continuously to levels of X and Y during the entire period. For details, see Kaufman (1976).

The linear system in (9-10) can be shown to be a reasonably good approximation to a variety of more complex change models. This seems a good argument in favor of the model in (1-2). However, we cannot assert that the basic model is the correct model for panel analysis. Other more complex differential equation models will give rise to different panel specifications. Nonetheless, it is comforting to learn that the basic model has more general value than is commonly acknowledged in discussions of panel analysis.

Both these arguments have a "best case" flavor. That is, they depend on nature's working in the interests of valid inference. But, what happens if the more complex model is correct? More concretely, what are the consequences of estimating the basic model when ΔX_t and ΔY_t have causal effects?[3] If the qualitative implications from the basic model hold over plausible ranges of the parameters of the model with change effects (7-8), this argues strongly in favor of the basic model. If, on the other hand, results from the basic model change radically with alterations in parameters of (7 - 8), we must proceed cautiously. In particular, we must identify those ranges of parameter values over which inferences from the basic model have some validity and restrict our attention to empirical contexts that meet the conditions.

The first two arguments treat the basic model as a structural model in its own right. But in the present discussion, it must be considered as the <u>reduced-form</u> of the model in (7-8). That is, we reconceptualize the basic model as an algebraic rearrangement of (7-8). It is obtained by solving (7-8) for Y_t and X_t:

$$Y_t = C^{-1}\left[(\gamma_0 + \gamma_1\delta_0) + (\gamma_1 + \gamma_3\delta_2')Y_0 + (\gamma_2' + \gamma_3\delta_1)X_0 + (\gamma_3 v' + u')\right] \quad (11)$$

$$X_t = C^{-1}\left[(\delta_0 + \delta_1\gamma_0) + (\delta_1 + \delta_3\gamma_2')X_0 + (\delta_2' + \delta_3\gamma_1)Y_0 + (\delta_3 u' + v')\right] \quad (12)$$

Where

$$C = 1 - \gamma_3\delta_3.$$

Equations (11) and (12) provide a new interpretation of the coefficients of the basic model in terms of the parameters of the change model.

Recall that the change effects model is not identified. Thus one cannot use reduced-form coefficients to obtain unique estimates of the α's and β's. But as long as u' and v' are well behaved (i.e., uncorrelated with X_0 and Y_0) we can always obtain good estimates of reduced form coefficients. That is why we investigate the usefulness of the reduced-form for informing us about causal relations holding under the more complex model.

In using the basic model, we report estimates of α_2 as the effect of X_0 on Y_t and β_2 as the effect of Y_0 on X_t. We confine attention to the former as treatment of the two cases is completely parallel. According to (11),

$$\alpha_2 = C^{-1}[\gamma_2' + \gamma_3\delta_1] \quad (13)$$

Thus we investigate whether the quantity in (13) is a reasonable measure of the effect of X on Y over time.

The quantity in parentheses in (13) is familiar to structural equation analysts. It is the sum of a direct effect and an indirect effect. The direct effect, γ_2', is the so-called cross-lag effect. The indirect effect, $\gamma_3\delta_1$, is the effect of X_0 on Y_t via X_t. Thus the reduced-form effects summarize both direct and indirect effects. The total effect of X_0 (the sum of the direct and indirect effect) is multiplied by C^{-1}. This multiplier adjusts the total effect for the cycle of causation implied by the model. According to the model, any increment in X_t produces a change in Y_t, which in turn affects X_t, etc. So effects of X_0 are propagated over an infinite cycle of causal loops joining X_t and Y_t. The multiplier is the appropriate rescaling of the effects of X_0. Thus to understand the reduced-form coefficients we must consider both the total effects and the multiplier.

Since the model is underidentified, we cannot conduct an exhaustive analysis. The reader can always choose some combination of γ's and δ's that will cause trouble. However, it is informative to consider a series of special cases.

Case 1: X has no effect on Y_t: $\gamma_2 = 0$, $\gamma_3 = 0$. Clearly $\alpha_2 = 0$ for this case. This result, though obvious, is important. It tells us that when X has no effect, we will not mistakenly identify effects of Y or Δ Y on X_t as effects of X on Y.

Case 2: X_0 affects Y_t but ΔX_t does not: $\gamma_2 > 0$, $\gamma_3 = 0$. Again the result is simple: $\alpha_2 = \gamma_2$, the cross-lag effect of X_0 in the structural form. This holds even when ΔY has an effect on X_t.

Now we consider more troublesome cases. Throughout we assume that the autoregression, δ_1, is positive, since this is the case in our empirical work. Without loss of generality, we assume that the effect of X_0, γ_2, is positive as well. In the general case, the multiplier, C^{-1}, can take on any value. Until further notice we assume that C^{-1} is also positive.

Case 3: X_0 and ΔX_t have the same sign effects: $\gamma_2 > 0$, $\gamma_3 > 0$.

In this case the reduced-form contains both a direct and an indirect effect. Since for the moment we assume a positive multiplier, the sign of the reduced-form effect is:

$$\gamma_2' + \gamma_3\delta_1 = \gamma_2 + \gamma_3(\delta_1 - 1). \qquad (14)$$

If both $\gamma_2 > 0$ and $\gamma_3 > 0$, the quantity in (14) is positive when

$$\delta_1 > 1 - \frac{\gamma_2}{\gamma_3}. \qquad (15)$$

We will consider three cases. When the lagged effects and change effects are equal, i.e., $\gamma_2 = \gamma_3$, the requirements for the quantity in (15) to be positive is $\delta_1 \gtrless 0$. This amounts to requiring negative feedback in the process generating the X_t. I mentioned earlier that we do not find this problematic in our research.

Similarly, when $Y_2 > Y_3$, the criterion is that δ_1 exceed some negative quantity, $(1 - K)$ where K is the ratio of Y_2 to Y_3. So neither is this case problematic.

Finally, there is the case in which the change term effect exceeds the lagged effect: $Y_3 > Y_2$. The criterion for the reduced-form effect to be positive, that δ_1 exceed $1 - K$, is now potentially problematic. Since $Y_3 > Y_2$, $0 \leq K \leq 1$. Thus δ_1 must exceed some positive quantity. For example, if Y_3 is twice Y_2, then we require that $\delta_1 > \frac{1}{2}$. So we will run into trouble when Y_3 is much larger than Y_2 and the autoregression of X is small. Such cases do not seem likely in our research as the autoregressions are rarely small. Nonetheless, we must recognize that the effect of X_0 in the reduced form will have opposite sign from the effects of S_0 and ΔX in the structural form (but see below).

So in all but exceptional circumstances, we expect that the sign of α_2 agree with those of the relevant structural parameters. However, quantitative estimates of the effect of X_0 will be wrong for two reasons. Using the reduced-form gives X_0 credit for the causal effects of ΔX_t. This is a mistake as far as inferences about X_0 are concerned. But it is not wrong regarding the effect of X considered globally. We attach no special significance to the 1950 levels of variables. We have chosen the time points out of design convenience. So it does not upset me in this stage of inquiry to confuse effects of ΔX_t with those of X_0.[4] The second reason why the quantitative estimates of α_2 differ from the structural parameters involves the multiplier. We delay discussion of this issue.

Case 4: The system is close to equilibrium: X_0 and ΔX_t have the same effect as do Y_0 and ΔY_t: $\gamma_2 = \gamma_3$; $\delta_2 = \delta_3$. When systems are close to equilibrium, the effects of (small) changes are close to the effects of initial levels. In this special case,

$$\alpha_1 = \gamma_1 C^{*-1}$$

$$\alpha_2 = \gamma_2 \delta_1 C^{*-1}$$

$$\beta_1 = \delta_1 C^{*-1}$$

$$\beta_2 = \delta_2 \gamma_1 C^{*-1}$$

where $C^* = 1 - \gamma_3\delta_3 = 1 - \gamma_2\delta_2$. Notice that

$$\gamma_2 = \alpha_2/\beta_1 \text{ and } \delta_2 = \beta_2/\alpha_1.$$

If, as is often the case in our research, the reduced-form autoregressions β_1 and α_1 are close to unity, the reduced-form cross-effects are close to the structural-form cross-effects.

Case 5: X_0 and X_t have opposite sign effects: $\gamma_2 > 0$, $\gamma_3 < 0$. Here we find real problems, as well we should. If lagged effects are positive and change effects are negative, any attempt to come up with an overall effect of X will mislead. The relevant expression is again (15). The autoregression coefficient in the X_t equation, δ_1, again plays a central role in determining the sign of the reduced-form effect. If $\delta_1 < 1$, then (15) is positive and the reduced-form effect has the same sign as the structural effect of X_0. When $\delta_1 = 1$, the situation is even better; it is the same as Case 2. However, when δ_1 exceeds unity, expression (15) is the sum of a positive and a negative quantity and (15) can be either positive or negative depending on the relative size of γ_2 and γ_3 (holding δ_1 constant). The qualitative result is similar to that in Case 3. When γ_3 is much larger than γ_2, we can obtain the wrong sign effect (here, we also need $\delta_1 > 1$).

To this point we have seen that inferences from the basic model have at least the proper sign as long as the effect of ΔX_t is not much larger than the effect of X_0. This conclusion is reinforced when we focus on the multiplier. Recall that $C^{-1} = 1/[1 - \gamma_3\delta_3]$. Thus it is positive when $\gamma_3\delta_3$ is less than unity. If $\gamma_3\delta_3 > 1$, the system will grow explosively. Any exogenous impact on X_t, say, will be amplified through each cycle of $X_t \rightarrow Y_t \rightarrow X_t$. The system is unstable and cannot long maintain this causal structure.

This is not to say that it can never happen in nature. Many of the structural properties we study grow exponentially during the period. However, in many analyses we adopt log transforms to linearize such growth. (See substantive chapters for more details, especially Ch. 6.) In light of such transformations, it does not seem likely that $\gamma_3\delta_3$ exceeds unity for the cases we consider. Nonetheless, we recognize that violations of this condition wreak havoc with inferences from the reduced-form. It is difficult to establish precisely the implications of explosive growth for reduced-form estimates since it is not obvious that other parameters, e.g., δ_1, would remain unchanged under large increases in γ_3 and δ_3. All I can do is caution the reader that the value of even the qualitative inferences from our basic model may be incorrect if this condition is not met.

This third perspective on the basic model, then, leads to more cautious conclusions. In general, it tells that we will not go far wrong when the effects of ΔX_t and ΔY_t are small relative to those of X_0 and Y_0 and the growth is stable (in the transformed variables).

So the class of models we use can be justified from any of three perspectives: (1) that the basic model in (1-2) approximates the true causal structure, operating with a certain lag; (2) that (1-2) is the solution of a differential equation system relating levels of X and Y to changes in X and Y; or (3) that the basic model is the reduced-form of the proper model which allows for causal effects over the study period. At the time we began our research, the sociological literature emphasized the first of these perspectives. However, none of us believed that the lag periods we used correspond with any fundamental feature of the processes under study. Thus we are forced to either of the other positions. I favor the second; other members of our research group incline more towards the third. We agree, however, that empirical analysies of the sort we report are necessary preliminaries to adequate specification of the dynamics of national development.

Issues in Estimation

Our research context poses some special methodological issues. These issues hold with equal force when the basic model is viewed as a structural model or whether it is considered a reduced-form for a more complex model. In this section I will identify the issues, outline our strategies for addressing them, and cite relevant technical literatures that contain more detailed treatments.

So far I have assumed that the disturbances (the total effect of all omitted variables) were well behaved. In particular I assumed that they were independent of the variables appearing on the right hand side of the basic model. If this is so, ordinary least squares (OLS) estimators are consistent (see Johnston 1972: Ch. 9). If, in addition, the disturbances have constant variance, OLS is also asymptotically efficient. The first two complications we consider involve likely failures of the independence and constant variance assumptions.

a. Autocorrelation: Most readers are undoubtedly sensitive to the possibility that the disturbances are not independent from period to period. The problem is endemic to panel studies, and the studies we report are no exception. Consider the composition of the disturbances. In our research, the set of omitted variables includes: material infrastructure (e.g., national systems of transport, communication, etc.), technology, cultural organization, national history, etc. Each of these features of social organization has two properties: (1) they are stable, that is, they do not shift radically from period to period; and (2) they affect many of the processes we study. As long as these causal effects are not included in the model, they are forced into the disturbance terms. And, given the pair of properties identified above, the disturbances will be correlated from period to period. Nations with unobserved characteristics that generate exceptionally high levels of some outcome in one period will tend to have exceptionally high levels in the next period. The more enduring are the unobserved causal forces, the stronger will be the autocorrelation of the disturbances, the dependence from period to period.

The impact of autocorrelation of disturbances on OLS estimators is well known (see, for example, Johnston 1972: Chs. 8, 10; Hannan and Young 1977). We must consider two effects. First, autocorrelation implies some non-independence of observations; the analyst using such data has less information about the process than would be given by the same number of independent observations. But, OLS assumes that the observations are independent; and calculations of confidence intervals and tests of significance reflect this assumption. Thus use of OLS with autocorrelated disturbances biases standard errors towards zero and thus gives inflated levels of statistical significance.

The second effect applies particularly to models, like the basic model, that include lagged dependent variables. It is easy to show (by writing the basic model for the period from -t to t_0 and making appropriate substitutions into the basic model) that autocorrelation of disturbances implies that the disturbances at time 0 will be correlated with both X_0 and Y_0. In this case OLS estimators are biased and inconsistent.

We are concerned with the properties of OLS estimators in small samples. We have some relevant evidence from Monte Carlo simulations. In regressions of a variable on only the lagged dependent variable, autocorrelation (of the type termed "first-order autoregressive") produces an upward bias in the slope associated with the lagged dependent variable (Malinvaud, 1961). Addition of another regressor reduces this upward bias. Nerlove (1971) and Hannan and Young (1977) obtained similar results using a different error structure (a variance components models with random but constant individual-specific effects, see below). The slope of the lagged dependent variable is biased upward while the slope of an exogenous variable is biased downward. The latter effect reflects the negative correlation of estimators for positively correlated regressors. The autocorrelation biases pushes the slope of the lagged dependent variable upwards and thereby pushes the slope of a correlated regressor downwards.

The results of the simulation studies do not apply directly to our model. Here both X_0 and Y_0 are correlated with the disturbance in each equation while in the simulation studies the additional regressor was not correlated with the disturbance. Nonetheless, the results seem hold approximately. Take equation (1) for example. The enduring portion of the disturbance affects both Y_0 and Y_1 directly but X_0 only indirectly through effects on Y_{-t} and effects of Y_{-t} on X_0. These indirect effects will usually be much smaller than the direct effects. Consequently the correlation of Y_0 with the disturbance should be considerably higher than that of X_0 with the same disturbance. Then the coefficient of Y_0 will be more affected by autocorrelation bias. It is not clear what will be the effect of autocorrelation on the slope of X_0. This depends on the magnitudes of the various effects and the strength of autocorrelation. I cannot assert that the slope of X_0 will be biased towards zero as in the simulation studies. Nonetheless it seems highly unlikely that the bias in this slope would be positive. In the cases we analyze I expect this bias to be relatively small and unsystematic.

All but two of our empirical studies use OLS estimators. The results just cited imply that their estimates of slopes of the lagged dependent variables will be biased upwards. The estimates of the cross-effects on which we base our inferences may be biased in complex ways. But it seems unlikely that any such biases will systematically distort our inferences across the multiple analyses that we report.

In two cases, Chapters 4 and 7, we correct for the most pervasive form of autocorrelation bias. In Chapter 4 we pool several waves of panel observations and estimate models that include disturbance terms that are specific to each nation. The nation-specific factors summarize the most enduring features of social organization and are assumed to be constant over the study period. The model should also control for unobserved causal factors that change slowly relative to the processes under study. Use of a pooled model enable us to identify the causal structure and to form generalized least squares estimators of causal effects. These estimators have been shown to have good small sample properties (Hannan and Young, 1977).

Chapter 7 also treats the factors producing autocorrelation as unobserved latent variables. Use of multiple indicators permits identification of a model with latent variables producing autocorrelation for each indicator of the dependent variable. The model is estimated by maximum likelihood using Jöreskog's (1969, 1973) procedures for analyzing linear structural equation models.

b. Heterosceasticity: The assumption of constant error variance is also problematic for some of the outcomes we study. Some of our outcomes measure the scale of a system, e.g., GNP, school enrollments. It is unlikely that Guyana, say, has the same error variance in either outcome as the Soviet Union or the U.S. A small percentage deviation for the latter is larger than the initial level for the former. In general period to period fluctuations for giant nations will be much larger than those for small nations. In other words, we expect that the error variance for such scale variables to be increasing functions of the scale.

Suppose the error variance of u_t in equation (1) is proportional to, say, the square of X_0, i.e.

$$\text{Var}\ (u_t) = \theta X_t \sigma_u^2$$

Then, it is straightforward to correct for heteroscedasticity by forming weighted least squares estimators (WLS). The procedure is as follows. Divide the entire structural equation by X_0 (this transforms the disturbance to u_t/X_0 which has variance $\theta^2\sigma_u^2$ -- constant across values of X_0). Apply OLS to the transformed equation but treat the coefficient of X_0^{-1} as the constant, the constant as the coefficient of X_0, and the coefficient of Y_0/X_0 as the coefficient of Y_0. In other words, we continue to focus on the original structural equation and introduce the transformation only to repair a problem in the distribution of the disturbances.

This procedure has a superficial similarity to use of structural equation models for ratios of components to scale factors, e.g., enrollment ratios, GNP per capita. Much published research and some of our research deals with relations among such ratios. Alternatively we treat the relevant scale factors, e.g. population size, as causal variables in their own right. Using the first strategy we might regress per capita GNP on lagged per capita GNP and some political variable. In the second we would regress GNP on lagged GNP, the political variable, and population size (perhaps at both time 0 and time t). Since the latter strategy is used much less widely than the former, the matter deserves some comment.

Ideally our theories would arbitrate between these formuations. Since our theories lack such precision, we must use other criteria for choosing between them. Use of ratio variables is commonly defended as a method for adjusting or controlling for scale. But the second formulation does this as well, though the nature of the adjustment differs. Are there any other grounds for choosing between the two? Some sociologists have recently argued that the use of ratio variables may complicate inference (Freeman and Kronenfeld 1973; Schuessler 1973; Fuggit and Lieberson 1974). They have noted particularly the difficulty that arises when one regresses, say, the ratio of expatriated profits to GNP on GNP (or GNP per capita). Since the same variable, GNP, appears in the numerator on one side of the equation and as a denominator on the other, the relationship between the two ratios is constrained. So, for example, a hypothesis that two variables be negatively related, does not face much likelihood of being rejected

A more general problem has not been addressed in the sociological literature. Consider a model in which 1970 GNP/Pop depends on 1950 GNP/Pop and 1950 primary enrollment ratio (number of primary students divided by size of the relevant age group). The coefficient for the cross-effect of the enrollment ratio on per capita GNP summarizes a number of possible effects: the effect ot size of 1950 enrolled population on 1970 GNP, a (nonlinear) effect of 1950 enrollments on 1970 population, a (nonlinear) effect of 1950 school aged population on 1970 total population and on 1950 GNP. I would not attribute the same substantive significance to each of these effects. Therefore, I am unwilling to summarize all these effects in a single term. I prefer to model the process in terms of primitive variables for this reason.

My argument hinges on my substantive orientation. Other members of our research group are more comfortable formulating arguments in terms of ratio variables. So we do not speak with one voice on this issue.

c. Measurement Error: Much of the work of our research project concerned measurement. We were constrained by the availability of data gathered by international organizations and, in several instances, by research groups. However, we devoted considerable effort to comparing estimates of the same quantity from numerous sources, checking the internal consistency of various indices (both over time on the same index and across indices of the same construct at a point in time), etc. As a result of this work we believe that most of the data we use are of high quality.

However, the sources of error in reports of population size, GNP, and school enrollments are well documented. Although we have eliminated many gross errors (and during the period of our research the U.N. published revised and corrected figures for the entire span we cover) we have certainly not eliminated measurement error. And, even random measurement error biases OLS estimates of structural parameters. Elsewhere (Hannan, Rubinson, and Warren 1974) we discussed at length the likely patterns of random and nonrandom errors in these data. We proposed and estimated crudely some models with unobservable variables measured with error by multiple indicators. Use of these models enables one to adjust estminates of structural parameters for both random and specified nonrandom errors in variables. Chapter 7 contains a more refined analysis of similar models using maximum likelihood procedures.

We have not, however, used models with latent variables widely in our research. They demand the use of multiple indicators. For several of the variables we study, particularly political structure and economic dependence, measures are available for only small numbers of nations on any one indicator. Using several indicators at once reduces the effective sample size below the levels at which a multivariate analysis can be trusted at all. This constraint is currently losing force as gaps in the data are being filled. As Chapter 7, the most recently completed of the empirical papers, documents, sufficient data on political organization are now available for the sensible use of unobservable variables models with multiple indicators. We expect the use of such models to become increasingly important in future research on these issues. In the meanwhile, we take consolation in the findings in Chapter 7 and in Hannan, Rubinson, and Warren (1974) that many of the variables we analyze are measured quite reliably by sociological standards.

d. Functional Form of Relationships: Structural equation models are sometimes criticized for limiting attention to linear-additive relationships. Such a charge has no basis in logic or fact. A structural equation model can have any functional form. Some of the models we estimate are linear in both variables and parameters, as in equations (1-2). Others are linear only in the parameters, as in regressions of logarithms of variables on lagged variables and other variables. The reader can establish that such relations are nonlinear by taking antilogs of the equation. Finally some of the models we estimate are nonlinear in the parameters. For instance, in Chapters 4, 6, and 7 we estimate relationships separately for rich and poor nations. The functional forms used are defended in the empirical chapters. This is not the place for any extended discussion. Rather my point here is to emphasize that the structural equation perspective we adopt does not restrict us to any particular substantive hypothesis, e.g., linear-additive relations.

Summary

Most of our empirical work involves variations on a simple model. We treat each outcome as a function (sometimes linear) of a set of lagged variables, including the lagged dependent variable. This type of model entered the social science literature as an analogue to an experiment. But the model is not completely general and in no sense does it render causal inference unproblematic. We take the position that the model must be treated as a substantive model of the process and that the quality of inferences from the model depends on the quality of the substantive representation. I advanced three justifications for the use of the simple model in our research. In the course of stating these arguments, I indicated possible difficulties with using this model to study national social change.

In large measure, these difficulties involve the strength of the effects of changes in variables during the study period. If the effects of changes are different in sign from the effects of initial levels, the basic model will tend to misstate causal effects. For reasons that I outlined, I do not know and cannot determine empirically whether this is a problem in our research. In effect, we have proceeded under the assumption that the effects of initial levels of variables represent reasonably well the effects of all levels that occur during the period. This, then, is the major substantive assumption underlying our work.

We reject the view that special estimation procedures must be designed for panel studies. Instead, we treat the model as a structural equation model and use structural equation procedures. Our work was conducted during a period in which sociological methodology underwent profound changes. In the early years of our research, the sociological literature did not offer useful guidelines on most estimation issues. However, rapid progress has been made and sociologists have begun to utilize a wider variety of estimators. The empirical chapters completed in different epochs of the project reflect some of these differences. In our earlier research we relied heavily on ordinary least squares estimators. As our research progressed, we altered procedures to handle certain complications, particularly autocorrelation, heteroscedasticity, and measurement error. In most cases the more complex procedures had never been applied to studies of national social change. Our experience is that more refined analyses generally give the same qualitative picture as did our earlier analyses (with certain exceptions -- see Chapter 7), but with a sharper focus. ~~I believe~~ that it is important that future research on national social change extend the use of these and related procedures.

Footnotes

[1]Lazarsfeld's work actually led to two traditions. His treatment of sixteen-fold tables (the full cross-classification of two dichotomous variables measured at two points in time) involved the line of thought discussed in the text. At the same time, his treatment of the so-called eight-fold table (the cross-classification of later measures on one variable on the earlier values of both variables) was much closer to the position we take.

[2]For example, it is easy to develop simple situations in which such measures fail. For example, suppose X_0 affects Y_t but Y_0 does not affect X_t. Suppose further that X is highly autocorrelated but Y_0 is not and that X_0 and Y_0 are reasonably highly correlated. Then the cross-lag correlation procedure will lead to the inference that Y causes X (the correlation of Y_0 with X_t will be larger than the correlation of X_0 with Y_t). It is equally easy to come up with additional complications that have the same consequence for the cross-correlation test. Kenny (1973) has stated a set of conditions under which a generalization of cross-lag correlation procedure yields valid inferences. These restrictions are exceedingly restrictive and appear to hold only for systems in equilibrium (among other things the contemporaneous correlations among variables must be constant over time). The structural-equation approach is suited to a wide range of situations beyond the scope of cross-lag correlation analysis.

[3]For reviews of this perspective on panel analysis, see Duncan (1968-69, 1972, 1975), Goldberger (197), Hannan and Young (1977), and Heise (1969).