SOCIAL SCIENCE RESEARCH 14, 80–101 (1985)

Panel Data and Models of Change: A Comparison of First Difference and Conventional Two-Wave Models

JEFFREY K. LIKER, SUE AUGUSTYNIAK, AND GREG J. DUNCAN

The University of Michigan

The method of first differences as an approach to modeling change is described and it is compared to more conventional two-wave panel models. Substantial advantages are found to the first-difference approach, especially if there are unmeasured, unchanging predictor variables in the model. It is also argued that there are substantial problems in the interpretation of results from the conventional two-wave models. Some of the analytic results are illustrated with a number of applications to the area of stressful life events. © 1985 Academic Press, Inc.

Recent theoretical advances emphasize the dynamic, everchanging aspects of humans and their social environments (Elder, 1979; Brim and Kagan, 1980). Static conceptions are criticized for failing to view aging as a developmental process. The complex interaction of changing humans in changing environments is not thought to be captured adequately by simple relationships among variables at a point in time. Hence, many advocate longitudinal analysis of panel data collected at multiple time points as a way to capture these complex, often reciprocal influences over time.

Panel data on individuals, families, and their life conditions are becoming increasingly available and open new possibilities for analysis. But they also present new challenges and problems. The greatest potential contribution of longitudinal data is that they permit empirical analyses of dynamic aspects of behavior. Such analyses can range from the simple description of change or allowance for some lagged response to the estimation of truly dynamic structural models that explain short-run or life cycle behavior of individuals. Unfortunately we are still far from reaching this ultimate goal of obtaining reliable parameter estimates of truly dynamic models for a number of reasons.

Authors are listed in random order to denote their equal contributions. Helpful comments from James Morgan, Willard Rodgers, Ronald Schoenberg, and an anonymous referee are gratefully acknowledged. Requests for reprints should be sent to Jeffrey K. Liker, Department of Industrial & Operations Engineering, University of Michigan, IOE Bldg., 1205 Beal Ave., Ann Arbor, MI 48109-2117.

0049-089X/85 $3.00

A first prerequisite is that social science theory must advance to the point where we can specify all of the necessary components of dynamic, reciprocal processes (Bohrnstedt, 1975). For example, general statements of the complexity and interplay of human–environment interactions that are prevalent in discussions of human behavior fall short of specifying all relevant variables, the functional form, and the timing and duration of influences. Second, complex dynamic models make great demands on measurement quality and frequency. In the absence of very specific theories on the form and process of influences, time series with short intervals over long time spans are needed to determine empirically time lags and complex patterns of causation. Moreover, the "signal" must far exceed the "noise." For many applications, social scientists are fortunate to have two waves of data of somewhat dubious measurement quality, making it difficult to refine their models.

Given these theoretical and data limitations, major advances from the use of longitudinal data for the estimation of truly dynamic models are apt to be slow in coming. Given this, it is important not to overlook the more modest benefits of longitudinal data for the estimation of what are essentially static models. After all, a tremendous amount of professional time and effort have been devoted to the formulation and estimation of static models. In some cases, panel data from several points in time can be used to obtain better parameter estimates of these models through the method of first differences. The basic properties of such models are well known (e.g., Wonnacott and Wonnacott, 1970), but their implications have not been well developed for the kinds of social science models that have been estimated with cross-sectional or short-term panel data. In the case of some models, dynamic structure can be captured in a cross-sectional formulation and the method of first differences is likely to produce better estimates of the dynamic parameters.

In this paper, we detail the method of first differences focusing on its strengths and limitations compared to cross-sectional analysis and conventional two-wave models. The strengths include differencing out unmeasured and unchanging causes of the outcome measure that may be associated with measured independent variables, eliminating measurement error biases under certain conditions, and adequately representing dynamic processes under certain rather restrictive conditions.

Taking first differences, that is, the difference between equations representing processes at two points in time, results in a differenced equation which includes change scores. Several researchers argue that the use of change scores for modeling purposes is problematic (Bohrnstedt, 1969, 1975; Cronbach and Furby, 1970). Their reasons include inadequacy of change scores for representing continuous processes, the unreliability of measurement, and the presumed biases that result if the correlation between the initial level of a dependent variable and its change is not explicitly modeled. We argue that the first of these arguments is probably correct for many social phenomena, but its full implications cannot be understood except in the context of a specified theoretical model. By contrast, the statements on measurement error in the dependent variable and the necessity of accounting for the correlation between initial level and change (sometimes called "regression to the mean") do not usually apply to equations derived from first differences. We demonstrate why this is the case and argue that many published applications of two-wave panel models control for lagged values of endogenous variables without adequate theoretical justification. This discussion suggests that there are serious problems of interpretation with the commonly estimated cross-lagged and other two-wave panel models.

The paper begins with a discussion of a general model derived by first differences. The effects of measurement error are considered in the context of this discussion. In the next section a model of first differences is compared to conventional two-wave panel models. Regression to the mean, state dependence, heterogeneity, and dynamic models of behavior are all examined as possible justification for the inclusion of initial level of the dependent variable on the right-hand side of a change equation. A final section considers an example from the literature on stressful life events that is well represented by the first-difference model in its dynamic aspects and an example that is poorly represented by such a model.

THE GENERAL MODEL OF FIRST DIFFERENCES

Most estimates from survey data are of cross-sectional models. A very simple form for such models is:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_{it} + \varepsilon_{it} \tag{1}$$

where Y_{it} is the value of the dependent variable for the i^{th} individual at time t, and X_{it} and Z_{it} are values of the two independent variables for the i^{th} individual at time t. In most survey applications, t corresponds to the time of the interview. Although the model as specified by Eq. (1) is static in the sense that all of the measures are taken at a single point in time, Schoenberg (1977) has shown that such models are generally consistent with dynamic processes in which the state at time "t" is a function of the initial conditions of the system. Such models are termed "nonergodic" and are contrasted with ergodic models in which the past can be forgotten in the sense that the current state does not depend upon the initial conditions. We assume initially that model (1) is correctly specified and later consider variations of this model.

Suppose that a panel survey design gives us measures of Y, X, *and* Z for the same individual at a subsequent point in time, say $t + 1$, and further suppose that β_1 and β_2 (the effects of X and Z on Y, respectively) do not change between t and $t + 1$. Can those panel data give us better estimates of the crucial parameters in Eq. (1)?

The answer to this question depends upon the measurement properties of the X and Z variables and upon the nature of the error term in Eq. (1). Under certain circumstances, a simple difference equation such as Eq. (2) below, in which changes in Y are regressed on changes in the X and Z variables, give better estimates of β_1 and β_2 than would come from estimation of Eq. (1).

$$\Delta Y = \Delta\beta_0 + \beta_1 \Delta X + \beta_2 \Delta Z + \Delta\varepsilon \tag{2}$$

where "Δ" represents simple change between t and $t + 1$ and the individual subscript "i" has been (now and hereafter) suppressed for notational convenience. Those circumstances include the following:

1. *Instances where the Z variables are unmeasured and unchanging.* Most social science theories reserve a place for the effects of unchanging "taste," background, or personality factors. Typically these factors are not measured. If they are correlated with observed X variables and with the Y variables, then estimation of Eq. (1) with cross-sectional data may well be biased by the omission of these Z variables. (This is demonstrated more formally below.) A change equation like (2) solves this omitted variable problem because changes in these unchanging Z variables are, by definition, zero and Eq. (2) reduces to a regression of ΔY on ΔX. The parameter β_1 in this change equation corresponds to the same parameter in level Eq. (1), but the estimation problems caused by the Z variable have been eliminated. Note, however, that this method of obtaining cleaner parameter estimates of observed X variables will only work in instances where the values of X change for a substantial portion of individuals over time.[1] Less obvious but also crucial is the assumption that the error terms in the process that generates changes in X are independent of the error term in (2).

2. *Instances where observed X variables are measured with errors that persist over time.* Measurement error in the X variable is likely to bias estimates of β_1 in (1) toward zero. To see this, suppose that the "true" value of X_t is related to the observed value of X_t by

$$X_t^* = X_t + u_t \tag{3}$$

where X_t^* is the observed amount of X at time t, X_t is the true amount, and $u_t \sim N(0,\sigma_u^2)$.

For the cross-sectional Eq. (1), it can be shown that as the sample size becomes very large, the OLS estimate of β_1 will not, in general, collapse on the value β_1. Instead, this probability limit (plim) will be

$$\text{plim}\,(\hat{\beta}_1) = \frac{\beta_1 + \beta_2 b_{ZX}}{1 + [\text{Var}(u_t)/\text{Var}(X_t)]} \tag{4}$$

[1] Hausman and Taylor (1980) develop a method for estimating parameters of unchanging X variables in first-difference equations.

where b_{ZX} is the regression coefficient of X when Z is the dependent variable.

There are two sources of inconsistency affecting the estimate of β_1 from cross-sectional Eq. (1), the first one due to measurement error and the other due to the correlation between the omitted Z variable and X. In general, the larger the measurement error (Var(u_t)) and the greater the correlation between X and Z, the greater the inconsistency of the OLS estimate of β_1 from Eq. (1).

But now consider the change Eq. (2). If measurement error in the X variable is correlated over time according to

$$u_t = \rho u_{t-1} + v_t, \tag{5}$$

and if the Z variable is unchanging, then the probability limit of the OLS estimate of parameter β_1 from change Eq. (2) is

$$\text{plim}(\hat{\beta}_1) = \frac{\beta_1}{1 + [(1-\rho)^2\text{Var}(u_t) + \text{Var}(v_t)]/\text{Var}(\Delta X)} \tag{6}$$

where $\text{Var}(\Delta X) = \text{Var}(X_t) + \text{Var}(X_{t-1}) - 2\text{Cov}(X_t, X_{t-1})$. If there is no autocorrelation in the measurement error in the X variable (i.e., $\rho = 0$) and no biasing effect of unmeasured Z variables, then the cross-sectional form is generally preferred to the change form since the variance of ΔX will generally be smaller than the variance of X. On the other hand, a remarkable result from (6) is that perfect autocorrelation between the measurement errors of X at the two points in time (i.e., $\rho = 1$ and $v_t = 0$) will cause no estimation problems for the method of first differences; the OLS-estimate of β_1 from (2) is consistent. Although perfect autocorrelation will be rare in survey data, substantial positive autocorrelation is likely to arise when respondents persistently over- or understate responses or persistently misinterpret questions.

3. *Instances where the panel data give more reliable measurement of changes in X between t and t + 1 than in the level of X at time t.* Suppose that Y is earned income and X is work experience. Cross-sectional surveys obtain measures of X retrospectively, a procedure filled with possible memory error. But a panel that provides annual measures of work hours over the period between t and $t + 1$ will provide relatively more reliable measurement of changes in X between t and $t + 1$. (Corcoran, Duncan, and Ponza, 1983).

In sum, the case for the superiority of change Eq. (2) over level Eq. (1) is obvious from a comparison of (6) and (4). The advantages of first-difference equations over cross-sectional formulations are greatest when (1) powerful, unmeasured Z variables bias the cross-sectional estimates, (2) errors in the X variable over time are highly autocorrelated, and (3) ΔX is measured more reliably than X. The advantages of first differences are weakened (and cross-sectional formulations may be preferred) when these conditions are not true and when the X variables are highly correlated over time.

Extensions of model (1) to include lagged values of the exogenous X variables result in the need to have more than two waves of data to estimate the first-difference form, but the basic advantages of differencing remain. Suppose that Y is thought to be a function of not only the current level of X but also its past value, i.e.,

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 Z_t + \varepsilon_t \qquad (7)$$

A differenced form of Eq. (7) is

$$\Delta Y = \Delta\beta_0 + \beta_1 \Delta X_t + \beta_2 \Delta X_{t-1} + \beta_3 \Delta Z + \Delta\varepsilon \qquad (8)$$

where $\Delta X_t = X_t - X_{t-1}$ and $\Delta X_{t-1} = X_{t-1} - X_{t-2}$. Estimation of this model would require three waves of data, since three observations on X (but only two observations on Y) are called for. All of the advantages of the first-differenced form derived above, especially the benefits of removing the biasing effects of unmeasured and unchanging Z variables, apply to this instance of a lagged X variable.

Extensions of model (1) to include lagged values of Y on the right-hand side of the equation raise a number of theoretical and statistical issues. If Y_t is *caused by* past levels of Y, then Eq. (1) is clearly misspecified and so is change Eq. (2). Differencing obviously does not solve fundamental problems of model misspecification. Although many change models that have appeared in the sociological literature do have lagged values of the dependent variables on the right-hand side, their specification is not usually well grounded in theory. Nor is it the case that there are good statistical reasons for the conventional treatments of lagged endogeneous variables. These issues are discussed in the context of two-wave panel models in the next section.

FIRST-DIFFERENCE EQUATIONS AS COMPARED TO CONVENTIONAL TWO-WAVE, TWO-VARIABLE MODELS

Algebraic Properties

First difference models such as (2) bear a superficial resemblance to the general class of two-wave, two-variable (2W2V) models explained by Duncan (1969) and developed at greater length by other authors (e.g., Kessler and Greenberg, 1981). Although certain restrictions can be placed on the 2W2V models to make them identical to (2), these restrictions are typically not imposed. As a result there is a fundamental difference in how the two approaches treat the initial level (i.e., lagged values) of the dependent variable in the change model. These issues are explored in this section. A primary focus is on the possible justification for the conventional treatment of the initial level of the dependent variable on the right-hand side of the equation. We show that it is extraordinarily difficult to justify its inclusion, and, as a result, parameter estimates from conventional 2W2V models have dubious validity.

The general two-wave linear panel model can be expressed in equation form as follows:

$$X_2 = \alpha_0 + \gamma_1 X_1 + \gamma_2 Y_1 + \gamma_3 Y_2 + v \tag{9}$$

$$Y_2 = \alpha_1 + \beta_2 X_1 + \beta_1 Y_1 + \beta_3 X_2 + u. \tag{10}$$

The most general form allows for correlations between X_1 and Y_1, for possible two-way causation between the time 2 measures of X and Y, and for the possible correlation between the error terms in each of the two equations. The general model is underidentified and unless some restrictions are placed on this model (e.g., constraining some of the parameters to equal zero or to equal one another), there is no hope in estimating its parameters (Duncan, 1969).

We begin by discussing the way model specification is typically approached by sociologists and then discuss how the two-wave, two-variable model relates to a "first difference" approach to using two waves of panel data.

1. Conventional sociological treatment. In the jargon of sociologists' use of two-wave models, the parameters representing the effects of Y_2 on $X_2(\gamma_3)$ and X_2 on $Y_2(\beta_3)$ are called "contemporary" effects and are thought somehow to occur instantaneously. The parameters γ_2 and β_2 are called "lagged" effects. A number of approaches can be used to place constraints on (9) and (10) so the equation system is identified. One common approach is to assume there are no contemporary effects—all effects are lagged by some specified amount. In this case γ_3 and β_3 are costrained to equal zero and the equation system is identified.

These constraints produce what is usually referred to as a cross-lagged panel model. How can we substantively interpret the coefficients of this model? Taking Eq. (10) as an example and incorporating the constraints that there are no contemporaneous effects, no constant term, and arranging terms in a more intuitive way, one obtains

$$Y_2 = \beta_1 Y_1 + \beta_2 X_1 + u. \tag{11}$$

The effect of Y_1 on $Y_2(\beta_1)$ is typically called a "stability coefficient." That is, controlling for X_1 and any other exogenous variables included in the equation, β_1 represents the extent to which the dependent variable remains stable over time. In social psychology, the dependent variable of interest is often a personal trait such as psychological distress (see example below) and β_1 would be interpreted as the extent to which this characteristic is a stable trait of individuals. The problem with this interpretation is that "stability" is not a theoretical justification for including Y_1 as an explanatory variable predicting Y_2 and the interpretation of β_1

and β_2 as estimates of parameters of a causal model is problematic. All that has been said is that individuals high on distress at time 1 tend to be high at time 2 and individuals starting out low tend to remain low. One must probe further into the causes of the observed stability.

One possible explanation for stability is that some unmeasured variables such as childhood experiences and genetic structure account for psychological distress at times 1 and 2. This is referred to in the econometric literature as "heterogeneity" and, as discussed below, including Y_1 on the right-hand side as a control for heterogeneity can lead to serious biases in the estimates of parameters β_1 and β_2.

When Y_1 is included as a predictor of Y_2 to control for the "stability" of Y, how can we interpret the effects of X_1? Sociologists using this model would like to think that β_2 is an unbiased estimate of the true causal effect of X_1 on Y_2, though it is often acknowledged that this assumption depends on whether the proper lag between time 1 and 2 has been specified. Even if this time span represents the correct lag structure and if Y_1 is related to Y_2 because of "heterogeneity," we show below that β_2 will probably be a biased estimate of the causal effect of X_1 on Y_2.

Another way of deriving Eq. (11) is with the approach used by Bohrnstedt (1969). He begins with the assumption that one is interested in the effect of X_1 on the difference between Y_1 and Y_2 as

$$Y_2 - Y_1 = \beta_2 X_1 + u. \tag{12}$$

Bohrnstedt then notes that Y_1 is correlated to the change score ΔY and argues this represents "regression to the mean" and creates bias. He then argues that Y_1 should be included as a predictor variable on the right-hand side of (12) to control for this bias. Minor algebraic manipulation shows that the addition of Y_1 to the right-hand side of (12) produces an equation that is equivalent to (11). The problem with this reasoning is that regression to the mean does not necessarily create estimation problems (see below) and is not a proper justification for including Y_1 as a predictor on the right-hand side.

2. *First-difference approach.* Suppose theory suggests that X_1 has a lagged effect on Y_2. Will the method of first differences help in estimating the parameters of a model consistent with that theory? We can represent the model as

$$Y_2 = \delta_1 X_1 + \delta_2 Z + u_2 \tag{13}$$

where Z represents unmeasured background variables (e.g., childhood experiences and genetic structure). If this process also occurs with the same structure between time 0 and time 1, we can write:

$$Y_1 = \delta_1 X_0 + \delta_2 Z + u_1. \tag{14}$$

Since Z are unmeasured variables, their effects cannot be directly modeled and if they are related to X, δ_1 will be a biased estimate. The cross-lagged panel design (11) uses Y_1 as a surrogate for these unmeasured variables which we argue below leads to new biases. By contrast, taking first differences leads to unbiased estimates as

$$Y_2 - Y_1 = \delta_1(X_1 - X_0) + u_2 - u_1. \quad (15)$$

There are three important points about (15). First, the unchanging Z variables have been differenced out so it is no longer problematic that they are unmeasured. Second, a third observation on $X(X_0)$ is needed. Third, Y_1 does not appear as a predictor variable on the right-hand side of (15). That is, regression to the mean is not presumed to be a concern and there is no attempt to estimate the "stability" of Y. Since the question of whether Y_1 belongs on the right-hand side of any of Eqs. (10) through (15) is so fundamental, this issue is examined below in detail.

The Role of Initial Level as a Predictor Variable

1. Regression to the mean in models of change. When specifying a model with a differenced dependent variable, many researchers express concern about potential bias from regression to the mean. Bohrnstedt (1969, 1975) argues that the negative correlation between initial level and change should be controlled by adding the initial level of the dependent variable as an explanatory variable. In this section we show that this is not only unnecessary, it is likely to introduce bias rather than eliminate it.

The phenomenon of regression to the mean, first documented by scientists studying the intergenerational progressions of vegetable sizes and of human heights, is the process by which a data series averages out to a mean level or growth rate. When compared to his family's progression in height, an offspring who is unusually tall will generally have children whose height represent smaller deviations from the trend. If change in height is charted, the tall offspring will be associated with a larger than average change while his children will be associated with smaller than average or even negative change.

The presence of random shocks in a model, whether they arise from measurement error or from parts of the process which through lack of knowledge cannot be modeled, implies that looking at just one observation per unit of analysis can be misleading. In our example on heights, the change in height between the unusually tall father and the grandfather yields a misleading indicator of the family's average change in height. This problem is not unique to differenced observations; the presence of random shocks is as much of a problem when studying the level of a variable, e.g., if we just observe the fathers' height our inference about the family's average height will be inaccurate.

Consider for a moment the process summarized in (1), in which the dependent variable Y is explained or described by X. Dropping the Z variables yields the following:[2]

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t. \tag{16}$$

Maintaining the assumption of unchanging parameters between t and $t + 1$ gives the following first difference form:

$$Y_{t+1} - Y_t = \beta_1(X_{t+1} - X_t) + \varepsilon_{t+1} - \varepsilon_t$$

or

$$\Delta Y = \beta_1 \Delta X + \Delta\varepsilon. \tag{17}$$

The differenced equation looks like the model in Eq. (16) with redefined variables. What causes some concern is that the error term in (17) is correlated with the initial level of Y, i.e., $E(\Delta\varepsilon\ Y_t) \neq 0$. The ready availability of Y_t and its integral place in the model seems to lead to the conclusion that Y_t should be used to control for the disturbance term, $\Delta\varepsilon$. As a result, the initial level of the dependent variable is often included as an explanatory variable in the change equation. However, correlation with the disturbance term is clearly not an acceptable justification for including a variable in the model, despite the fact that *for a given individual* that information improves *predictions*. We could always find variables that are correlated with the error term, even if it is only the time of day that the information was coded, but we would never search for such variables let alone argue that they belong in the model.

At the core of this controversy is *explanation* versus *prediction*. With most data *any particular variable is well predicted by past values of itself and yet is rarely caused by its past values* so we do not advocate adding lagged values unless there is a causal link. Despite the widespread use of initial level in models of change, few examples come to mind of processes where change is caused by current level. In this respect, including initial level on the right-hand side of a differenced model is no more or less valid than including lagged dependent variables as predictors in the undifferenced model. For example, one's socioeconomic origin will improve prediction of intergenerational mobility; however, we are generally less interested in predicting an individual's mobility pattern than in understanding the process by which mobility is enhanced or constrained.

The argument to leave out initial level, unless it has a direct effect on change, may seem like advice to throw out perfectly good information. After all, although a father who exceeds his father's height by quite a bit does not cause his children to have less change, they are likely to

[2] The set of variables X_t that describe Y at time t could include lagged observations of X.

exceed him by less than he exceeded his father. Similarly, someone who earns an unusually large raise one year is likely to experience somewhat smaller earnings increases in subsequent years. This brings up a second important point. Typically, *we are not considering observations on only a single case.* Information on initial level is not needed to control for random shocks, even for prediction, as can be seen by the following example.

Consider the relationship between earnings and psychological distress (PD) shown in Figs. 1 and 2. Suppose the relationship between the initial level of earnings and PD, depicted in Fig. 1, is causal with PD related to earnings as

$$\mathrm{PD}_t = \alpha + \beta \, \mathrm{Earn}_t + \varepsilon_t. \tag{18}$$

There will be individuals with positive residuals above the regression line and those with negative residuals below it.

Consider individuals with a given level of earnings and positive residuals, e.g., those labeled ε_j in Fig. 1. On average, the residuals for these individuals at time $t + 1$ will be less positive and their residuals in the change equation will be negative. Individuals with negative residuals in the first period will most likely have positive residuals in the change equation (ε_k in Figs. 1 and 2). However, as long as these two sets of individuals are distributed independently of *changes* in earnings, they will not affect the estimate of β in the change equation.

In general, if the sample has been drawn randomly from the population on which the model is based, the model is properly specified and the distribution of the disturbance term is symmetric, then controlling in the change equation for the presence of positive or negative disturbances in the initial period is unnecessary because the effects of the disturbance should average out over the population.

Two possible counterarguments are that we do not have this faith in our model or that we use unweighted data in which individuals with either positive or negative disturbances in the initial period are oversampled, presumably by oversampling people with high or low initial PD. In these cases it might seem useful to include the initial level of PD in the change equation as a proxy for the disturbance term in the initial period.[3] To see that there are more appropriate proxies, consider the implications of including ε_t as an explanatory variable in the change equation.

[3] Individuals with positive disturbances at time t have, as a group, above-average levels of Y at time t; those with negative disturbances have lower average levels of Y. Conversely higher initial levels of Y imply a somewhat higher probability of positive residuals and therefore a somewhat higher probability of a small change.

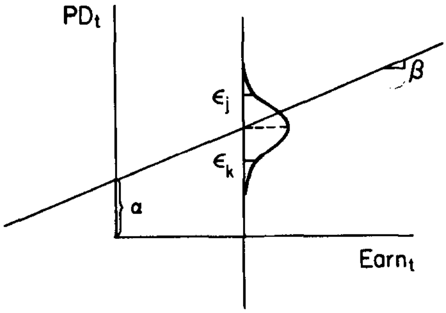

FIG. 1. Cross-sectional relationship between earnings and psychological distress.

The disturbance from Eq. (16) can be expressed as

$$\varepsilon_t = Y_t - \beta_0 - \beta_1 X_t. \tag{19}$$

Including ε_t in the change equation to control for regression to the mean does imply that Y_t enters as an explanatory variable in the change equation, but it also implies that X_t enters as well. The change equation with ε_t added as an explanatory variable is

$$\Delta Y = \beta_1 \Delta X + \phi \varepsilon_t + \Delta \varepsilon$$

or

$$\Delta Y = \beta_1 \Delta X + \phi Y_t - \phi \beta_0 - \phi \beta_1 X_t + \Delta \varepsilon \tag{20}$$

To estimate this equation, an instrumental variable must be used for Y_t because it is generally correlated with the disturbance term ($\Delta\varepsilon$). Including only Y_t as a proxy for the initial disturbance term (ε_t) is inappropriate. While it is true that Y_t is correlated with ε_t, controlling for ε_t implies far more about the equation to be estimated than the mere inclusion of

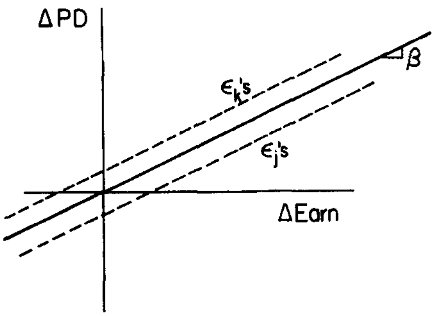

FIG. 2. Longitudinal relationship between changes in earnings and changes in psychological distress.

Y_t in the change equation.[4] Thus, the conventional form of 2W2V models, with the initial value of the dependent variable on the right-hand side, cannot be justified by arguments about "regression to the mean."

2. *Heterogeneity and state dependence.* Arguments for the inclusion of initial level in a change equation sometimes center on the correlation of observations over time. Two distinct theories of why observations on a given individual are highly correlated over time are discussed in this section. The first, heterogeneity, proposes that unmeasured, individual-specific factors that lead to certain behaviors or attitudes at one point in time persist and influence future behaviors or attitudes. We have already seen that the effects of unchanging, unmeasured characteristics on the *level* of psychological distress are differenced out in a change equation. Thus, this kind of heterogeneity is handled well in a differenced equation in which the initial value of the dependent variable does not appear on the right-hand side.

Unmeasured differences between individuals may affect not only initial level but also change. For example, particularly resourceful people may be especially healthy psychologically and also continue to develop psychologically at a faster rate than others less able to cope with life's challenges.[5] In this case, direct effects of unmeasured characteristics on *change* will not be eliminated in first-difference equations. Some researchers argue that the initial level of the dependent variable should be included in the change equation to account for these individual differences.

Even if individual-specific variables are not observed, or for other reasons cannot be included in the change equation, it is still inappropriate to include the initial level of PD on the right-hand side of a change equation. The initial level of PD is an inappropriate proxy for heterogeneity, even though it may be highly correlated with the excluded variables, because it is correlated with the disturbance term. Its inclusion in the equation will result in inconsistent estimates. A more appropriate procedure for controlling for heterogeneity in this case is to apply the method of instrumental variables and use an instrument for PD on the right-hand side of the change equation.

The second explanation of the correlation of individual observations over time is that of *state dependence.* Proponents of this view postulate that something about the initial level of the dependent variable has direct, causal bearing on subsequent change. For example, individuals with high

[4] Recall from the discussion of (1) and of (2) that any unobserved Z-type variables that are part of the level equation drop out of the difference equation. In (20), where the initial residual is added to the change equation, these variables no longer drop out and, if correlated to other included variables, the equation is not estimable.

[5] Note that this would lead to increased variance in psychological well being over time with the healthy and distressed growing increasingly far apart. Panel data in our experience do not support this view.

prior levels of PD may not be motivated to engage in training programs that might increase their earnings and reduce distress over personal failings. One may wish to build this assumption into a change model, in which case initial level of PD enters the change equation directly, as part of the theoretical model. This situation is often called state dependence because there is something about the initial state that causes future states and therefore change.

Even though state dependence provides a justification for including initial level as an explanatory variable in a model describing change, the likely correlation between initial level and the disturbance term means the model as stated will be estimated with bias. To obtain consistent estimates of the parameters, an instrumental variable must be used in place of PD. The proxy variable should be one that is correlated with PD in the initial period but is not correlated with the disturbance term in either period. One possible proxy would be a still earlier value of PD; however, if serial correlation is present in the disturbance term of the level equation, prior values of PD will not be appropriate instruments.

3. Dynamic models. The most sophisticated dynamic models are considered to be those that use differential equations to specify continuous change. These models are in principle better able to capture the intricacies of the theoretical process than simple difference models when the process is continuous over time. If one's psychological well being adapts gradually to new circumstances and measurements on psychological characteristics are frequent enough to capture this process, then a specification relating adjustment in psychological well being to a function of time and other explanatory variables may be useful.

In the context of discussing the role of initial level as a predictor variable, it is important to point out that initial level does not automatically enter the specification when differential equations are used. Coleman (1968) makes this point clear by prefacing the discussion on differential equation models by the comment "we might assume that the rate of increase of the number of actions of a given kind in a group would be proportional to the number of actions already taken." However, Coleman gives other examples where change is constant over time and does not depend on the level of the dependent variable.

Researchers often seem to assume that since differential equations used in the physical sciences are frequently written as functions that relate change to initial level, this type of specification is necessary, or at least usually appropriate. This is not the case. The appropriateness of a differential equation for either physical or social sciences depends on its intended use—explanation or prediction. In describing the rate at which water flows out of a tank, one could model the change in water level between two points in time as a function of the amount of pressure and the size of the opening. However, since pressure is exactly related to the water level for a given container, one could posit a *predictive* equation in which the change in water level is a function of the initial water level. Water level does not help explain the process, but since it is perfectly related to the causal factor (water pressure), it predicts very well.

In relationships outside the physical sciences there are several reasons why following this model may be problematic. First, in contrast to the water example, initial level is typically not perfectly related to the causal processes in social science examples, so it is not a perfect predictor of change. Second, social science phenomena usually depend on interrelations between many factors, so explanation of the process is needed in order to make accurate predictions as factors change; we do not have controlled processes. Third, because our relationships are generally not fully explained, but rather include a stochastic element, statistical problems ensue when initial level is included in the model.

This is not to say that initial level of the dependent variable never belongs in dynamic models, only that it does not enter *automatically* when a dynamic model is specified. There are dynamic models that do include initial level as a predictor because it is part of the causal process. Possible theoretical formulations that would justify the inclusion of the initial level of the dependent variable on the right-hand side of the change equation are partial adjustment models and certain distributed lag models. See Kessler and Greenberg (1981) or Kmenta (1971) for a discussion of these models.

APPLICATIONS TO STRESSFUL LIFE EVENT MODELS

The method of first differences has been presented as an approach to reducing certain sources of bias in cross-sectional equations. This procedure will not eliminate specification errors arising from misrepresentation of dynamic processes.

Under what conditions do first-difference equations accurately represent dynamic processes and under what conditions are they misrepresented? This section contrasts two theories of the relationship between socioeconomic factors and psychological distress to illustrate the conditions under which first differences yield valid and invalid specifications.

Reactive Theories

The study of socioeconomic status in relation to psychological distress has a longstanding tradition among sociologists. The argument is that environmental conditions associated with class position influence psychological states. For example, poor housing, lack of money, blocked opportunities, and the like cause lower SES individuals to feel alienated, depressed, and develop other symptoms that we will generally classify as symptoms of psychological distress (PD). This can be modeled in a cross-sectional formulation as

$$PD_t = \beta_0 + \beta_1 SES_t + \beta_2 Z_t + \varepsilon_t. \quad (21)$$

This formulation will generally yield a significant effect of SES (β_1) if the sample size is large (although the magnitude of the effect is generally quite modest). Such cross-sectional estimates have been criticized by those who argue other variables, typically unmeasured (Z_t), such as nutritional deprivation in infancy, inadequate parental role models, and even congenital disorders lead to both low SES and PD. Another argument is that the symptoms of distress, PD, actually cause people to be selected into relatively low SES categories—the social drift argument (Kessler and Cleary, 1980).

An immediate problem with the formulation in Eq. (21) is that the argument does not directly refer to SES per se, but to a set of environmental conditions that are associated with SES. Hence, SES is really a surrogate measure that is likely to contain considerable measurement error. Setting aside this measurement issue, how would first differences help us to model this process? The first difference equation would appear as

$$\Delta PD = \Delta\beta_0 + \beta_1 \Delta SES + \beta_2 \Delta Z + \Delta\varepsilon. \quad (22)$$

If the Z characteristics are unchanging, as they are in the examples above, ΔZ will equal zero and drop out of the equation. If there are consistent measurement biases in PD, as might be the case if self-reported symptoms are used, then these will be eliminated in the differenced equation. Moreover, the social causation argument falls into the class of "reactive" formulations which Augustyniak, Duncan, and Liker (1985) demonstrate is well represented by a simple differencing process.

The social causation argument is that psychological states respond to environmental conditions much as mercury in a thermometer responds to a given temperature level. A unit change in temperature will result in a unit change in the density of the mercury which we will see as a change in level on the thermometer. A pure environmental argument would imply that a unit change in SES will lead to a unit change in PD, which is directly reflected in the differenced Eq. (22). One might argue that PD will be affected with some time lag in which case this can be easily built into the equation as discussed above. We call this a "reactive" model since a person is psychologically reacting to some environmental state.

Estimation of this equation is still problematic for three reasons: (1) limited variance in ΔSES measure if measures at time 1 and 2 are both taken in adulthood when little change occurs; (2) measurement error in ΔSES; and (3) the influence of PD on SES is not taken into account.

The first two problems can be resolved by measuring the actual conditions for which SES is a surrogate. While occupational prestige and formal education (traditional SES components) do not change much in adulthood, environmental conditions such as income, housing, welfare receipt, and the like change a great deal (Duncan *et al.*, 1984) and may be more accurate indicators of the conditions that influence psychological health.

The issue of social causation versus social selection cannot be easily resolved as discussed above. One might make an argument about the time lags involved. For example, psychological states may respond to a new environmental condition rapidly while the reverse process takes more time. For example, even though one comes out of a bout of depression it will take time to undo the damage done while depressed, e.g., finding a job or a better job may require an investment in training. Hence, if one were to observe a correlation between changes in job status and changes in psychological states within a short time span it might be inferred that this reflects the effects of job status on psychological status, not the reverse.

The problem of reciprocal causation cannot be resolved by simply controlling for the initial level of PD as Wheaton (1978) suggests. In a widely cited paper, Wheaton (1978) develops a conventional 2W2V model, and it is instructive to compare that model to Eq. (22) above. The equation for "current" psychological health from Wheaton's Fig. 1 can be written as

$$PD_{1971} = \beta_0 + \beta_1 PD_{1967} + \beta_2 SES_{1965} + \beta_3 SES_{1960} + \beta_4 FASES + \varepsilon \quad (23)$$

where PD and SES are as defined above and FASES is father's SES. Wheaton uses confirmatory factor analysis with multiple indicators of psychological distress and adjusts for correlated specific variance in each multiple indicator over time, but this does not change the structural model above.

Equation (23) is algebraically equivalent to

$$PD_4 - PD_3 = \beta_0 + \beta_1^* PD_3 + \beta_2^*(SES_2 - SES_1) + \beta_3^* SES_1 + \beta_4^* FASES + \varepsilon \quad (24)$$

where $\beta_1^* = \beta_1 - 1$, $\beta_2^* = \beta_2$, $\beta_3^* = \beta_2 + \beta_3$, and $\beta_4^* = \beta_4$. For simplicity, we have used the subscripts 1, 2, 3, and 4 to refer to the years of measurement 1960, 1965, 1967, and 1971.

If we believed the level of SES affected PD with a lag of several years, a differenced equation would then look like Eq. (24) with the constraint that $\beta_1^* = \beta_3^* = \beta_4^* = 0$. Before relaxing these constraints, we must ask the question: Why should the initial level of psychological distress, a prior level of SES, and father's SES cause psychological stress from 1967 to 1971 to change in a particular way? If father's status had an effect on the subject's health apart from the subject's status, and this effect did not change over time, father's status would cancel out in a change equation since father's status represents an unchanging background variable. The discussion of the possible justifications for including initial level of the dependent change score as a right-hand side variable applies here.

We do not mean to assert that Wheaton (1978) was using first differences to derive Eq. (23). However, it is not clear from his paper what the justification is for including three SES measures and 1967 PD as causal factors in 1971 PD. One might argue that this is some form of a dynamic model that makes different assumptions from the "reactive" model, but it is not clear what those assumptions might be.

Adaptive Theories

The "reactive" model can be adequately modeled by simple first differences. However, there are other assumptions about the dynamic process relating environmental conditions to psychological states that cannot be handled in this way.

The theoretical framework of most recent studies of stressful life events assumes that people become distressed because of the disorientation created by a large degree of change. The essence of the argument is that people grow accustomed to the way things are and react adversely when their steady life flow is seriously disrupted. At the extreme, any change, positive or negative, could be stressful (Holmes and Rahe, 1967), since some degree of readjustment is required. Modifications of this theory, however, suggest that only particular types of change cause psychological distress (Mueller, Edwards, and Yarvis, 1977) and add the assumption that people recover their prior psychological state with time (Surtes and Ingham, 1980). That is, after some period of time people adjust to their new situation, which eventually becomes their new accustomed way of life. We refer to this as an "adaptive" model.

The contrast between the "reactive" and "adaptive" theories is illustrated in Fig. 3. Suppose that personal earnings are the sole determinant of psychological distress (PD). Figure 3 shows the time path of psychological distress for 1969 to 1974 under the assumption that earnings in 1969 are $12,000, that earnings increase to $15,000 in 1970, 1971, and 1972, and then fall to $10,000 in 1973 and 1974. In the reactive model, increases in earnings change PD levels by a given number of units, but PD does not change again until earnings change, since only the income level affects PD. Note that we are making assumptions about the timing of influences—PD responds to new levels of earnings within a 1-year period, and also about the duration of influences—there is no residual influence of having been at a specific income level. A person who earned $12,000 last year and $10,000 this year will have the same level of efficacy this year as a person who earned $5,000 last year and earns $10,000 this year.

In the "adaptive" model PD is a short-term response, but then after some time people adapt and return to some steady state level of PD. A

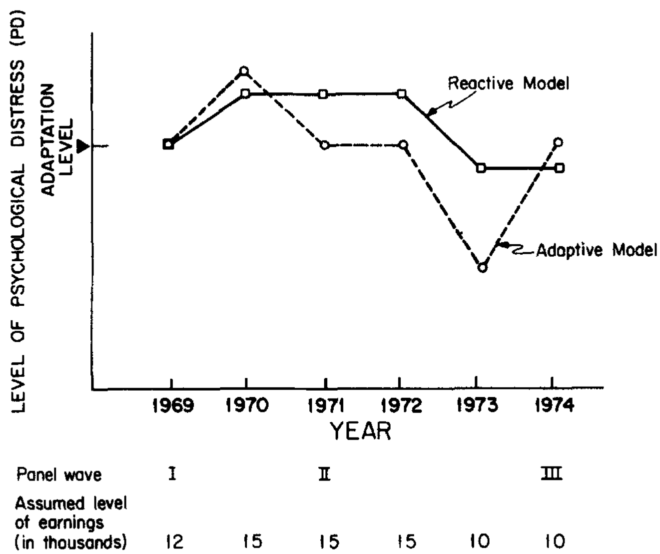

FIG. 3. Changes in psychological distress in the reactive and adaptive models.

process in which adaptations take between 1 and 2 years to complete is depicted by the dotted line in Fig. 3. There is some steady-state "adaptation level" of efficacy to which people return after a temporary disruption caused by a change in earnings. Until aspirations have been properly adjusted to this new earnings level, there is a discrepancy between aspirations and achievements that causes efficacy level to be above or below this adaptation level. Note that the timing of measurement becomes particularly important if this model applies. If one were to have measures at waves I, II, and III in Fig. 3, it would appear that the changes in earnings between these waves had no effect at all on efficacy. It is only when measures are taken before a change occurs and then again before the affected individual has had time to adapt that the maximum impact can be assessed. More than two observations are needed to plot the full process of adaptation.

We can represent the adaptive process more formally as

$$PD_t = \beta_0 + \beta_1(SES_t - SES_{t-1}) + \beta_2 Z_t + \varepsilon_t. \quad (25)$$

If we were to observe the psychological distress of a sample at a later point in time $(t + 1)$ we would observe

$$PD_{t+1} = \beta_0 + \beta_1(SES_{t+1} - SES_t) + \beta_2 Z_{t+1} + \varepsilon_{t+1}. \quad (26)$$

Taking first differences would result in

$$PD_{t+1} - PD_t = \Delta\beta_0 + \beta_1(SES_{t+1} - 2SES_t + SES_{t-1}) + \Delta\varepsilon. \quad (27)$$

An intuitive explanation for these results is as follows: If psychological distress is the result of a change in life conditions (i.e., the first derivative of change in a continuous process), then a change in the level of distress is the result of a change in the rate of change in life conditions (i.e., the second derivative). Equation (27) results from an additional algebraic step which emphasizes the fact that we have been taken beyond the realm of two waves of panel data, and a third wave of data is needed (or change must be measured retrospectively in at least one wave).

Most articles using panel data to assess the effects of changes in circumstances on psychological health differ from all three equations (25–27). Generally, they are our Eq. (25) with the addition of PD_{t-1} on the right side (Myers *et al.*, 1974; Pearlin, Marton, Lieberman, Menagham, and Mullan, 1981; Thoits, 1982; Williams, Ware, and Donald, 1983). Again we suggest that the implications of including the initial level as a right-hand side variable have not been carefully considered in these cases.

When we begin to consider adjustment processes such as the life event formulations it becomes increasingly implausible that discrete equations like (25–27) are adequate representations. To observe psychological distress that results from a change in environmental conditions (or persistent environmental conditions for that matter), one would have to measure psychological states before the individual had time to recover. Moreover, if recovery is a gradual process then the level of distress we observe will depend on the point in the recovery process that we observe the process. An appropriate differential equation model is necessary to represent the process as it unfolds over time.

SUMMARY AND CONCLUSION

This paper focuses on the method of first differences as a way of developing models of change. A great deal of professional effort has gone into specifying cross-sectional equations and differencing them at two time points can improve estimates in certain instances. These include (1) instances where certain explanatory variables are unmeasured and unchanging; (2) instances where observed X variables are measured with errors that persist over time; and (3) instances where panel data give more reliable measurement of changes in X between t and $t + 1$ than in the level of X at time t.

Another problem that has been discussed in the literature is the greater measurement errors in change scores. We have shown that this is not necessarily a problem in a differenced equation if the variables are measured with errors that persist over time; indeed, the model with change scores may even reduce biases caused by measurement error. Moreover, a simple algebraic manipulation will convert conventional two-wave, two-variable models into models with change scores and demonstrate that they do not reduce any measurement biases contained in models based on change scores.

The conventional two-wave, two-variable model used extensively by sociologists seems to have become part of the "culture" of linear panel analysis. With proper constraints this model appears to provide a means to model complex causal processes involving two-way causation. Duncan (1969) emphasizes that these constraints on the underidentified two-wave, two-variable model are critical and can dramatically alter substantive conclusions. We have gone further in questioning the basic structure of the model.

When we take a common cross-sectional model and difference it, we derive an equation that is substantially different from the two-wave, two-variable model. The main differences are that the differenced model does not include roles for the initial level of the dependent variable, nor for the initial level of the independent variable. We argue that including the initial level of the dependent variable is seldom justified on statistical grounds, although it may be justified on theoretical grounds. Even when theory calls for inclusion of this initial level, statistical problems are likely and an instrumental variable substitute should be constructed. A rationale for including the initial level of the independent variable is also necessary and we have not seen such a stated rationale in the substantive literature relating socioeconomic status and stressful life events to psychological distress discussed here.

In many cases cross-sectional equations are simply inadequate representations of dynamic processes. Our "adaptive" model is one such case which calls for some form of continuous time model. In this case, taking first differences of incorrectly specified cross-sectional equations will result in a poorly specified difference equation. Differential equation models appear to be a fruitful direction for future research on dynamic processes. These models will be no less susceptible to misspecification than two-wave, two-variable models, and we recommend caution in adopting functional forms used in the physical sciences that may not apply to social processes.

REFERENCES

Augustyniak, S., Duncan, G. J., Liker, J. K. (1985), "Income dynamics and self-conceptions: linking theory and method in models of change," *in* Life Course Dynamics: Trajectories and Transitions from 1968 to 1980 (G. Elder, Ed.), Cornell Univ. Press, Ithaca, N.Y.

Bohrnstedt, G. W. (1969), "Observations on the measurement of change," *in* Sociological Methodology (E. F. Borgatta, Ed.), pp. 1113–1136, Jossey-Bass, San Francisco.

Bohrnstedt, G. W. (1975), Structural Equations and the Analysis of Change, presented at the 28th annual meetings of the Gerontological Society, Louisville, Ky.

Brim, O., and Kagan, J. (1980), Constancy and Change in Human Development, Harvard Univ. Press, Cambridge.

Cronbach, L. J., and Furby, L. (1970), "How we should measure 'change'—or should we?" Psychological Bulletin **24,** 68–80.

Coleman, J. S. (1968), "The mathematical study of change," *in* Methodology in Social Research (H. M. Blalock, Jr., and A. B. Blalock, Eds.), pp. 428–478, McGraw–Hill, New York.

Corcoran, M., Duncan, G. J., and Ponza, M. (1983), "A longitudinal analysis of white women's wages," Journal of Human Resources **18**, 497–520.

Duncan, G. J. et al. (1984), Years of Poverty, Years of Plenty, Institute for Social Research, The University of Michigan, Ann Arbor.

Duncan, O. D. (1969), "Some linear models for two-wave, two-variable panel analysis," Psychological Bulletin **72**, 177–182.

Elder, G. H., Jr. (1979), "Social structure and personality: a life course perspective," *in* Life-Span Development and Behavior (P. Baltes and O. Brim, Eds.), Vol 2, Academic Press, New York.

Hausman, J. A., and Taylor, W. F. (1980), "Panel data and unobservable individual effects," Econometrica **49**, 1377–1398.

Holmes, T. H., and Rahe, R. H. (1967), "The social readjustment rating scale," Journal of Psychosomatic Research **11**, 213–218.

Howard, K. I., and Krause, M. S. (1980), "Some comments on techniques for estimating the source and direction of influence in panel data," Psychological Bulletin **74**, 219–224.

Kessler, R., and Cleary, P. (1980), "Social class and psychological distress," American Sociological Review **45**, 463–478.

Kessler, R., and Greenberg, D. F. (1981), Linear Panel Models, Academic Press, New York.

Kmenta, J. (1971), Elements of Econometrics, Macmillan Co., New York.

Mueller, D. P., Edwards, D. W., and Yarvis, R. M. (1977), "Stressful life events and psychiatric symptomology: change or undesirability?," Journal of Health and Social Behavior **18**, 307–317.

Myers, J. K., Lindenthal, J. J., and Peppar, M. P. (1974), "Social class, life events and psychiatric symptoms: a longitudinal study," *in* Stressful Life Events (B. P. Dohrenwend and B. S. Dohrenwend, eds.), pp. 191–206, Wiley, New York.

Pearlin, L. I., Lieberman, M. A., Menagham, E. G., and Mullan, J. T. (1981), "The stress process," Journal of Health and Social Behavior **22**, 337–356.

Schoenberg, R. (1977), "Dynamic models and cross-sectional data: the consequences of dynamic misspecification," Social Science Research **6**, 133–144.

Surtes, P. G., and Ingham, J. G. (1980), "Life stress and depressive outcome: application of a dissipation model to life events," Social Psychiatry **15**, 21–31.

Thoits, P. A. (1982), "Conceptual, methodological, and theoretical problems in studying social support as a buffer against life stress," Journal of Health and Social Behavior **23**, 145–159.

Wheaton, B. (1978), "The sociogenesis of psychological disorder: reexamining the causal issues with longitudinal data," American Sociological Review **43**, 383–403.

Williams, A. W., Ware, J. E., Jr., and Donald, C. A. (1981), "A model of mental health, life events, and social supports applicable to general populations," Journal of Health and Social Behavior **22**, 324–336.

Wonnacott, R. J., and Wonnacott, T. J. (1970), Econometrics, Wiley, New York.