Causal Inference with Panel Data

Paul D. Allison
University of Pennsylvania

For nearly half a century, the fundamental problem for statistical analysis in the social sciences has been how to make causal inferences from nonexperimental data (Blalock 1961). For nearly as long, there has been a widespread consensus that the best kind of nonexperimental data for making causal inferences is longitudinal data. Unfortunately, there has not been nearly as much consensus on the best methods for analyzing such data. The literature on longitudinal data analysis is much too vast for a detailed review in this paper, but here are some of the main themes. For psychologists and sociologists, the dominant approach has been some version of the cross-lagged panel model, originating with the two-wave, two-variable model proposed by Duncan (1969) and elaborated by many others (Markus 1979, Kessler and Greenberg 1981, Finkel 1995, Kenny and Judd 1996). A rather different approach has been to model longitudinal data as a multi-level or hierarchical linear model (Bryk and Raudenbusch 1992, Goldstein 1995). One version of this approach is known as random growth curve modeling (Muthén and Curran 1992). Finally, economists have distinguished between fixed- and random-effects models, and have developed several novel estimation methods for handling various elaborations of these models (Wooldridge 2002, Baltagi 1995).

In this article, I consider some linear models for panel data that embody many of the elements of these different approaches. What is particularly attractive about these models is that they protect against the two central threats to valid causal inference: unmeasured confounding variables and reverse causation. The models themselves are not particularly profound or original. What's novel is the estimation method. Although economists have developed

estimation methods that could be used for these models, their methods are generally unknown in the wider social science community and, with a few exceptions have not been incorporated into commercial software. By contrast, I show how these models can be easily estimated using widely available software for structural equation modeling. I also present results from simulations which suggest that these methods have superior statistical properties relative to those in the econometric literature.

DATA AND MODELS

I shall consider data of the following sort. We have a sample of n individuals, each of whom is observed at T points in time (t=1,…,T). Thus, the data set is "balanced" with the same number of observations for each individual. Although the proposed estimation methods can, in principle, be applied to unbalanced data, the initial development is much simpler if we exclude that possibility. At each time point, we observe two quantitative variables, x_{it} and y_{it}, which may have a reciprocal causal relationship. We may also observe a vector of control variables w_{it} which vary over both individual and time, and another vector of control variables z_i which vary over individuals but not over time.

The initial goal is to formulate a linear model that embodies a reciprocal relationship between x and y while controlling for w and z. Consider the following set of equations,

$$\begin{aligned} y_{it} &= \mu_t + \beta_1 x_{i(t-1)} + \beta_2 y_{i(t-1)} + \delta_1 w_{it} + \gamma_1 z_i + \alpha_i + \varepsilon_{it} \\ x_{it} &= \tau_t + \beta_3 x_{i(t-1)} + \beta_4 y_{i(t-1)} + \delta_2 w_{it} + \gamma_2 z_i + \eta_i + \upsilon_{it} \end{aligned}, \quad t = 1,\ldots,T \qquad (1)$$

where μ_t and τ_t, are intercepts that vary with time, β_1, β_2, β_3, and β_4 are scalar coefficients, δ_1, δ_2, γ_1, and γ_2 are row vectors of coefficients, and ε_{it} and υ_{it} are random disturbances. Although all the lags for x and y are shown here as lags of one time unit, the lags could be greater and could be different for each variable. The critical thing is that we do *not* allow for simultaneous

causation, which would require various kinds of problematic and *ad hoc* assumptions in order to estimate and interpret the causal effects. The terms α_i and η_i are "fixed effects", that is, fixed parameters that vary across individuals. They can be thought of as representing the effects on x and y of all unmeasured variables that are both constant over time and have constant effects.

The equations in (1) are essentially a cross-lagged panel model. They differ from the typical cross-lagged panel model by the incorporation of the fixed effects, which allow for the control of unmeasured confounders, and the presumption that the coefficients are constant over time. The latter assumption can certainly be relaxed, but will be maintained for now to keep things as simple as possible.

More needs to be said about the random disturbance terms, ε_{it} and υ_{it}. We'll assume that they are independent of each other (within and between time points) and normally distributed with means of 0 and constant variance (at least across individuals, although we could allow for variances that change over time). We'll also assume that w_{it} is strictly exogenous, meaning that for any t and any u, w_{it} is independent of ε_{iu} and υ_{iu}. With respect to x and y, we cannot assume strict exogeneity because both variables appear as dependent variables. Instead, we shall assume that they are *sequentially* exogenous (Wooldrige 2002). For all $u \geq t$, x_{it} is independent of ε_i and y_{it} is indepenent of υ_{iu}, i.e., the disturbance terms are independent of previous values of x and y.

ESTIMATION

Estimation of (1) is not straightforward for reasons that are well known in the econometric literature. First, the presence of lagged dependent variables as predictors in each equation means that conventional methods will yield biased estimates of the β coefficients under almost any condition. Elswhere (Allison 1990), I have argued that inclusion of lagged dependent variables as predictors may not be sensible in many situations. If we suppress the effects of the

lagged dependent variables (i.e., set β_2 and β_3=0) and if T=3, unbiased estimates of the coefficients can be obtained by taking first differences:

$$y_{i3} - y_{i2} = (\mu_3 - \mu_2) + \beta_1(x_{i2} - x_{i1}) + \delta_1(w_{i3} - w_{i2}) + (\varepsilon_{i3} - \varepsilon_{i2})$$
$$x_{i3} - x_{i2} = (\tau_3 - \tau_2) + \beta_4(y_{i2} - y_{i1}) + \delta_2(w_{i3} - w_{i2}) + (\upsilon_{i3} - \upsilon_{i2})$$

Note that z_i drops out of both equations. In this form, the equations can be estimated by ordinary least squares (OLS). But for T>3, the reciprocal relationship between x and y implies that conventional fixed-effects methods will still yield yield biased estimates.

Arellano and Bond (1991) proposed an instrumental variables estimator for models like those in (1), and this approach (with some variations) has become the standard among ecometricians. The Arellano-Bond estimator is currently available in Stata. In contrast, I show here how to estimate (1) directly using maximum likelihood methods that are readily available by way of conventional structural equation modelling (SEM) software, such as LISREL, EQS, AMOS, MPLUS and PROC CALIS (SAS). I first show how to implement the method by way of example. Then I present results of a simulation study showing that the method appears to do what is promised: produce unbiased estimates of the reciprocal effects of x and y.

As an example, I analyze data for 178 occupations in the U.S. for the years 1983, 1989, 1995 and 2001 (labeled T1-T4). The data come from the March "Current Population Survey: Annual Demographic File" (CPS). The observations in CPS data are individual persons, but the analysis required occupational averages for each year on all the variables. For each year, I calculated the proportion female and the median wage for females for each occupation. This was done only for the 178 occupations that had at least 50 sample members in each of the years. Further details can be found in England et al. (2004). For wages, the variables are labeled MDWGF1-MDWGF4, and for proportion female we have PF1-PF4.

For the model in (1) let y be median wage and let x be proportion female. In 1983, the correlation between these two variables was -.33, which was highly significant. There has been considerable controversy regarding the possible direction of causality between these two variables (England et al. 2004). One argument is that employers devalue occupations that have a high proportion female and, consequently, pay lower wages. The rival hypothesis is that declining wages make occupations less attractive to men; as they leave for better paying work, women fill their vacant positions. I shall assume that changes in either of these variables show up in changes in the other variable six years later. To keep the example simple, we will not consider any w variables (time-varying covariates) or z variables (time-invariant covariates).

By estimating the equations in (1), using conventional SEM software, we can assess each of the two possible causal effects. Although it's possible to estimate the two equations simultaneously, estimating them separately allows for considerably more flexibility in specifying the model. (The problem with simultaneous estimation is that a given variable has to be expressed in the same way as a dependent and independent variable.) Note that to do maximum likelihood estimation, we must strengthen our assumptions by specififying a joint distribution for all variables, in this case multivariate normality.

There are two key devices that are necessary to implement the method. First, the fixed effects in each equation are modeled as a latent variable that is allowed to be correlated with all time-varying predictor variables. The rationale for this method is described in Teachman et al. (2001) and Allison and Bollen (1997). Second, the assumption of sequential exogeneity is modelled by allowing the error term at each point in time to be correlated with *future* values of the time-dependent covariates, but not past values (Wooldridge 2002).

We used the CALIS procedure in SAS to estimate the two equations. Here is the program for estimating the first equation in (1):

```
PROC CALIS DATA=my.occ UCOV AUG;
LINEQS
  mdwgf4= t4 INTERCEPT + b1 pf3  + b2 mdwgf3 + falpha + e4,
  mdwgf3= t3 INTERCEPT + b1 pf2  + b2 mdwgf2 + falpha + e3,
  mdwgf2= t2 INTERCEPT + b1 pf1  + b2 mdwgf1 + falpha + e2;
STD
  falpha=s1, e2-e4=sa:;
COV
  falpha*mdwgf1 pf1 pf2 pf3=c0 c1 c2 c3:,
  e2*pf3 =c4;
RUN;
```

Here is a quick explanation. The UCOV option specifies that the matrix to be analyzed is a sum of squares and cross-products matrix. The AUG option augments this matrix with a column that corresponds to the intercept. These two options are necessary if it is desired to estimate the intercept term. Otherwise one could analyze the covariance matrix (by specifying the COV option). Following LINEQS (for linear equations) there is a separate equation for each dependent variable at each point in time, and those equations correspond directly to the equations in (1). Note that there is no equation predicting median wage or proportion female at time 1 because we do not observe their lagged values six years earlier (1977).

The fixed effects are represented by FALPHA in each equation. The COV statement allows for correlations between FALPHA and the time-varying covariates, thus implementing a fixed-effects model. Note that for the lagged dependent variable, a correlation is only allowed between FALPHA and the value of the variable at time 1. That's because only the time 1 variable is exogenous and correlations are only allowed among exogenous variables. There's actually no need to specify a correlation between FALPHA and the later values of the lagged dependent variable because FALPHA is one of the predictors in the equation for each of these

variables. The COV statement also allows a correlation between E2 and the cross-lagged variable at time 3, which corresponds to the assumption of sequential exogeneity.

To estimate the second equation, we simply repeat the program while interchanging the roles of PF and MDWGF:

```
PROC CALIS DATA=my.occ UCOV AUG;
LINEQS
  pf4= t4 INTERCEPT + b1 mdwgf3 + b2 pf3 + feta + e4,
  pf3= t3 INTERCEPT + b1 mdwgf2 + b2 pf2 + feta + e3,
  pf2= t2 INTERCEPT + b1 mdwgf1 + b2 pf1 + feta + e2;
STD
  feta=s1, e2 e3 e4=sa:;
COV
  feta*pf1 mdwgf1 mdwgf2 mdwgf3=ca:,
  e2*mdwgf3=cb;
RUN;
```

Table 1. Estimates for Reciprocal Model with Fixed and Lagged Effects

mdwgf2	=	-0.0836	*	pf1	+	0.3434	*	mdwgf1	+	7.9837	*	Intercept	+	1.0000	falpha	+	1.0000	e2
Std Err		2.4323		b1		0.0640		b2		1.2411		t2						
t Value		-0.0344				5.3680				6.4329								
pf2	=	0.2994	*	pf1	+	-0.00054	*	mdwgf1	+	0.3353	*	Intercept	+	1.0000	falpha	+	1.0000	e2
Std Err		0.0820		b2		0.00151		b1		0.0384		t2						
t Value		3.6534				-0.3572				8.7220								

Results for the two equations are shown in Table 1. Not surprisingly, each variable has a positive, statistically significant effect on itself six years later. With respect to the "cross-lagged" coefficients, however, there is no evidence for an effect in either direction.

I also estimated a model that removes the lagged dependent variables, and got essentially the same results for the cross-lagged coefficients. Similarly, a model that includes the lagged dependent variables but does *not* include the fixed effects (the classic 2-wave, 2-variable panel model) yields no evidence for a cross-lagged effect in either direction.

SIMULATION STUDY

To evaluate the proposed method, I generated observations from the following model:

$$y_{i0} = \alpha_i + u_{i0}$$
$$x_{i0} = \eta_i + v_{i0}$$
$$y_{it} = \beta_1 x_{i(t-1)} + \beta_2 y_{i(t-1)} + c_i + u_{it}$$
$$x_{it} = \beta_3 x_{i(t-1)} + \beta_4 y_{i(t-1)} + d_i + v_{it}$$

for t=1,…,4.. The stable, unmeasured components α_i and η_i were generated as bivariate standard normal variates with a correlation of .5. The disturbances u_{it} and v_{it} were each standard normal and independent of all exogenous variables. For each condition, I generated 500 samples. The baseline condition had a sample size of N=500, and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = .5$.

Using the SEM method just described, I estimated the parameters of the model for each of the 500 samples, with results reported in the first line of Table 2. It is sufficient to look at just one of the parameters, β_1, which represents the effect of x on y. Table 2 reports the mean of the 500 estimates, the standard deviation of those estimates and the mean of the estimated standard errors. Ideally, the latter two estimates should be very close. In the last column, I report the "coverage" , that is, the proportion of nominal 95% confidence intervals (calculated in each sample using the conventional normal approximation) that contain the true value of the parameter. If the method is performing well, this should be close to .95.

The first row of Table 2 indicates excellent performance. Bias was minimal, the average of the standard errors was close to the standard deviation of the estimates and the coverage was just a little below the nominal 95% level. Next, instead of a full factorial design, I varied one dimension at a time, leaving the others at their baseline values. I first varied the sample size, with N=100 and N=500. Then I varied the values of the parameter of interest, from 0 to -.5 to 1. Next, I eliminated the lagged effects of the dependent variable by setting $\beta_2 = \beta_3 = 0$ (although

the analysis model still estimated these effects). And, finally, I varied the value of the other cross-lagged effect (β_4), setting it to 0 and then to -.5. The only condition under which the performance of the proposed method was less than ideal was when $\beta_4 = 0$. Although the bias was small, the coverage was somewhat deficient at 88 percent.

Table 2. Results of Simulation Study

Model	Parameter[b]	Estimate[c]	Avg. SE[d]	Stan. Dev.[e]	Coverage[f]
Baseline	.50	.500	.028	.030	.93
Baseline + N=100	.50	.504	.062	.067	.94
Baseline + N=500	.50	.500	.020	.022	.94
Baseline + β_1=0	.00	.003	.030	.031	.95
Baseline + β_1 = -.5	-.50	-.500	.031	.034	.93
Baseline + β_1 = 1	1.00	.998	.028	.030	.94
Baseline + $\beta_2 = \beta_3 = 0$	.50	.502	.046	.053	.95
Baseline + β_4 = 0	.50	.519	.091	.010	.88
Baseline + β_4 = -.5	.50	.501	.031	.033	.95

[b]True value of the coefficient in the model producing the data.
[c]Mean of 500 parameter estimates.
[d]Mean of 500 standard error estimates
[e]Standard deviation of 500 parameter estimates
[f]Percentage of nominal 95% confidence intervals that include the true value.

In later versions of this paper, I plan to extend the simulations to compare the performance of the SEM estimator with conventional cross-lagged panel methods, conventional fixed-effects methods, and the Arellano-Bond estimator.

CONCLUSION

This paper proposes an SEM estimator of a linear cross-lagged panel model with fixed effects. Effective estimation of such a model should make it possible to draw valid and reliable conclusions about the relative magnitudes of reciprocal effects of two or more variables. This is because the method protects against both the possible confounding with unmeasured, stable covariates and against the potential biasing effects of reverse causation. By contrast,

conventional fixed-effects estimation methods are unsatisfactory because they do not adequately address the endogeneity of the two variables. The instrumental variables estimator of Arellano and Bond attempts to accomplish the same goal, but the methods described here may be more accessible and familiar to many social scientists.

REFERENCES

Allison, P. D. (1990),"Change scores as dependent variables in regression analysis," in *Sociological Methodology 1990*, ed. C. Clogg, 93-114, Oxford: Basil Blackwell.

Allison, P. D. and. Bollen, K.A (1997) "Change Score, Fixed Effects, and Random Component Models: A Structural Equation Approach.," paper presented at the Annual Meeting of the American Sociological Association.

Baltagi, B. H. (1995), *Econometric Analysis of Panel Data.* New York: John Wiley & Sons.

Blalock, H. (1961). Causal Inferences in Nonexperimental Research. University of North Carolina Press, Chapel Hill, NC.

Bryk, A .S., and Raudenbusch, S .W. (1992), *Hierarchical Linear Models: Application and Data Analysis Methods*. Newbury Park, CA: Sage.

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

England, P., Allison, P.D., Wu, Y. and Ross, M. (2004), "Does Bad Pay Cause Occupations to Feminize, Does Feminization Reduce Pay, and How Can We Tell with Longitudinal Data?," paper prepared for presentation at the Annual Meeting of the American Sociological Association, San Francisco, CA, August 16.

Goldstein, H. (1987), *Multilevel Models in Educational and Social Research*. London: Griffin.

Kenny, D. A., & Judd, C. M. (1996). A general procedure for the estimation of interdependence. *Psychological Bulletin, 119,* 138-148.

Kessler, Ronald C. and David F. Greenberg (1981) *Linear Panel Analysis: Models of Quantitative Change*. New York: Academic Press.

Muthén, B. and Curran, P. (1997), "General Longitudinal Modeling of Individual Differences in Experimental Designs: A Latent Variable Framework for Analysis and Power Estimation," *Psychological Methods*, 2, 371-402.