NBER TECHNICAL PAPER SERIES

THE ESTIMATION OF DISTRIBUTED LAGS IN SHORT PANELS

Zvi Griliches

Ariel Pakes

Technical Paper No. 4

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02138

October 1980

This paper was presented at the NBER's 1980 Summer Institute in Productivity. We are grateful to Gary Chamberlain, Lung-Fei Lee and Christopher Sims for helpful comments and to John Bound and Brownyn Hall for research assistance and valuable suggestions. The work is part of the NBER's program on Productivity and Technical Change and has been supported by NSF Grants PRA-13740 and SOC79-04279. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #4
October, 1980

The Estimation of Distributed Lags in Short Panels

ABSTRACT

In this paper, we investigate the problem of estimating distributed lags in short panels. Estimates of the parameter of distributed lag relationships based on single time-series of observations have been usually rather imprecise. The promise of panel data in this context is in the N repetitions of the time that it contains which should allow one to estimate the identified lag parameters with greater precision. On the other hand, panels tend to track their observations only over a relatively short time interval. Thus, some assumptions will have to be made on the contributions of the unobserved presample x's to the current values of y before any lag parameters can be identified from such data. In this paper we suggest two such assumptions; both of which are, at least in part, testable, and outline appropriate estimation techniques. The first places reasonable restrictions on the relationship between the presample and insample x's, while the second imposes conventional functional form constraints on the lag coefficients associated with the presample x's.

Zvi Griliches
Ariel Pakes

National Bureau of Economic Research
1050 Massachusetts Avenue
Cambridge, Massachusetts 02138

(617) 868-3921

Zvi Griliches
Ariel Pakes
July, 1980
NBER
Harvard University and
The Hebrew University

This is a preliminary version which undoubtedly contains mistakes and does not contain an empirical example.

The Estimation of Distributed Lags in Short Panels

Introduction*

The problem we will be dealing with in this paper arises, as is often the case in econometrics, because we do not have all the data that we would like to have. In many contexts, we expect our independent variables to "work" only after a relatively long lag while at the same time wishing to discover such effects from a relatively short time series. Consider a very simple model

$$y_t = w_0 x_t + w_1 x_{t-1} + \ldots + w_m x_{t-m} + u_t$$

where we have suppressed the constant term for simplicity of exposition. Here m + 1 is the total length of the lag structure and the problem arises if m is large relative to T, the total number of observations on y, or relatively to T + θ, where θ is the available number of lagged observations on x. A typical example might be T = 20, m = 6, and θ = 2. I.e., 20 observations on y, 22 observations on x, and the necessity to estimate 7 w coefficients. In this case one would have to give up 4 observations on y to get the whole lag "in," and estimate 7 parameters on the basis of 13 degrees of freedom (allowing for the loss of 1 d.f. for the

*This work is part of the NBER's Program on Productivity and Technical Change and has been supported by NSF Grants PRA 79-13740 and SOC 79-04279. We are grateful to Gary Chamberlain, Lung-fei Lee and Christopher Sims for helpful comments and to John Bound and Bronwyn Hall for research assistance and valuable suggestions.

estimation of the constant term). Obviously the results will not be very good. One solution is to impose additional structure on the w's, choosing the form of the lag a priori -- polynomial, geometric, etc. Another solution, and the one we will be concerned with in this paper, is to increase the sample in another dimension, adding more individuals, states or commodities. The sample becomes then a time-series "panel" with N x T degrees of freedom.

There has recently been a significant increase in the number of panel data sets available. Though N in these data sets is often quite large, it is not clear just yet whether such data will result in "better" estimates of distributed lag coefficients. Part of the problem lies in the fact that panel data sets usually track their observations only over a rather short time interval. Thus, if in the example introduced above $T + \Theta$ was less than seven while $m > 7$, then there would be no way we could use the panel data to estimate all $m + 1$ w coefficients. Nor could we estimate even seven coefficients consistently (as N grows large) without either simplifying the model further (assuming a particular lag structure) or bringing in additional information about the missing observations on the independent variable. On the other hand, if it were reasonable to assume some form for the relationship between the missing and the observed data, and if the assumed form allowed for the identification of at least some of the lag coefficients, then the large size of N in panel data may allow for more precise estimates of the identifiable lag coefficients than would be possible from the information in a single time series. Of course, if we are going to build such a functional relationship into our estimating technique then it is desirable to have some way of testing whether or not it is a reasonable approximation for the problem at hand.

Let us now go back to the substantive issue of what types of questions about the w's one may want answers to. One is, what is the "total" effect of a change in x on y? That is, what is $\sum_{j=0}^{m} w_j$? Unfortunately, as Sims (1974) has shown long ago, that is not really knowable without the imposition of a priori constraints. A more modest question is: can one estimate any (how many?) of the w's consistently even if one cannot estimate the unseen tail of the lag distribution? This question is of interest because the shape of the lag structure is of interest on its own merit; indeed, it may be essential to the understanding of the phenomena we are interested in. Moreover, if one could get good estimates of the individual w_j's one might discover that they do follow a particular pattern which is consistent with a particular lag structure, and one might be willing to impose this structure in further investigations to provide an answer also to the first question.

To recapitulate we are interested in estimating at least some of the distributed lag parameters when the tail of the lag distribution is not directly observable. For a single time series this issue has been discussed previously and solved by Klein (1958), Dhrymes (1971), Madalla and Rao (1971) and Pesaran (1973) in the context of an assumed given lag structure form (e.g., geometric or Pascal). Because panel data tend to have many more degrees of freedom, we would like to solve the same problem without imposing a particular functional form a priori. This transforms the problem into one of estimating a system of multivariate regressions when there is a left out variable in each one and using an assumption about the relationship between the left out and the observed variables to

identify some of the coefficients of interest. Thus our models fit rather nicely into the general panel data analysis framework outlined by Chamberlain (1978). If $\underset{\sim}{y}_i$ is the T element vector of observations on y for individual i, and $\underset{\sim}{x}_i$ is the T + θ element vector of observations on x for the same individual, and if we assume $(\underset{\sim}{y}_i, \underset{\sim}{x}_i)$ are random drawings from a larger population of interest, then a fairly general starting point for panel data analysis is the multivariate 'wide-sense' expectation or projection of y on x; that is, $E^*[y|x] = \underset{\sim\sim}{\Pi} x$. $\underset{\sim}{\Pi}$ has T(T + θ) elements. In the models we discuss $\underset{\sim}{\Pi}$ will be composed of a mixture of two types of coefficients. One set consists of parameters from the lag distribution we are interested in. The other consists of coeficients in the regression of the "left out" contribution of the unobserved presample x's to the current value of y, on the observed or insample x's. The specific form of this regression function depends, of course, on how the unobserved x's and w's are generated. The assumptions which we will consider lead always to a null hypotheses that restrict $\underset{\sim}{\Pi}$ to have less than T(T + θ) free parameters and are therefore testable, at least in part, by comparing the restricted to the unrestricted estimates of $\underset{\sim}{\Pi}$.

The rest of this paper is organized as follows. The next section considers first the case of estimating the parameters of an unrestricted lag structure when the process generating the x's can be assumed to have a finite memory. By this we mean that the distribution of the recent x's conditioned on the whole past history of x depends only on a small number of such x's. Further on we add an individual specific constant term to allow for unobserved heterogeneity in the x-process. These assumptions will be appropriate if

the x's can be approximated by an integrated autoregressive process (where the order of the integration plus the order of the autoregression is p) or by a mixture process consisting of individual and time specific factors plus an autoregressive deviate. The autoregressive process need not be homogeneous over time; i.e., its parameters can change from year to year. We show that in this case one can test for p, the length of the memory in the x process, and that the model is capable of producing consistent estimates of the $T + \theta - p$ leading lag coefficients. Implicitly it amounts to finding the "backcasting" function which determines the projection of the unobserved presample x's onto the observed or insample ones, and using this function to solve the truncation remainder problem.

The third section considers the case when the "unseen" w's have geometric or some other relatively simple autoregressive structure. Implicitly this amounts to backcasting the w's rather than the x's. We show that in the panel data context this model reduces itself to the single or several left-out factors case considered by Griliches (1974) and Keifer (1979). More generally, we show that a "solution" to the truncation remainder problem requires either simple assumptions about the lag structure or alternatively about the stochastic process generating the x's. The virtue of our approach is that it provides us with ways of testing some of these assumptions before proceeding to impose them on the data.

II. Distributed Lags from Panel Data with Prior Structure on the x-Process.

All models discussed in this paper are special cases of the distributed lag model:

$$y_{i,t} = \sum_{\tau=0}^{m} w_\tau x_{i,t-\tau} + \alpha_i + d_t + u_{i,t}$$

where

$$E[u_{i,t} | d_t, \alpha_i \text{ and } x_{i,s} \text{ for all } i \text{ and } s] = 0 \qquad (1)$$

and

$$E[\underset{\sim}{u}_i \underset{\sim}{u}_j'] = \begin{cases} \Omega_u & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$$

where

$$\underset{\sim}{u}_i = [u_{i1}, \ldots, u_{iT}]$$

and

$$i=1,\ldots,N \quad \text{and} \quad t=1,\ldots,T \ .$$

A brief discussion of the maintained hypothesis in (1) is worthwhile before proceeding. First, it is not the most general of distributed lag models which could be estimated from panel data. The fact that there are repetitions on the distributed lag relationship for the same individual over different time periods, and for the same time period over different individuals, implies that one need not assume constancy of the lag parameters either over time or over individuals. The model does allow, however, for both individual and time specific constant terms. The presence of the individual specific (time specific) effects allows for

the possibility of unobserved factors which are fairly constant over time (over individuals) and impact both on y and on the choice of x. The time specific effects (the d_t) pose no particular estimation problems in what follows. The relevant limiting dimension for deriving the asymptotic properties of estimators from panel data sets is N, the number of individuals, and as N grows large we shall always be able to estimate the d_t exactly. To keep our notation as simple as possible, then, we shall not consider the d_t explicitly in the discussion below.[1] Allowing for the individual specific factors (the α_i) does, however, complicate matters. In both this and the next section, we will first develop estimation techniques for the simpler case where there are no individual specific effects ($\alpha_i=\alpha$ for all i) and then indicate the extension required to allow for this factor. The extended models allow one to test the hypothesis that in fact there are no such effects in the data.

Two points should be noted about the assumptions we have made on the disturbances. First we are assuming that conditional on the unobserved factors strict exogeneity holds. Second the disturbance vectors [the $\underset{\sim}{u}_i$, $i=1,\ldots N$] are assumed to be a random drawing from a larger population of such vectors but are allowed to possess a free covariance matrix. Thus the nature of the panel allows for a relatively straightforward solution to the rather troublesome problem of determining the time-dependence structure of the disturbances in distributed lag models. One

simply allows that structure to be free and estimates it along with the other parameters of the model.

In some applications the maintained assumption of (conditional) strict exogeneity and of the constancy of the lag parameters may be too restrictive. Since, as we shall show, they are in fact testable assumptions one may wish to relax them in a particular context. Model (1) is adequate, however, for exhibiting our basic results concerning estimation in the presence of truncation remainders.

The essence of the truncation remainder problem can be seen rather clearly from (1). Panels are usually quite short. If θ is the number of observations on x before the first observation on y and one is unwilling to assume that $m \leq T+\theta$ then, none of the lag parameters can be estimated consistently (as N grows large) from (1) without some further restrictions on the model.[2] On the other hand, if there are a set of reasonable restrictions which do allow for the identification of some of the lag coefficients then the fact that N is usually quite large may allow one to estimate them fairly precisely. This section will consider the estimation of lag coefficients when it is reasonable to restrict the auxiliary regression of the presample x's ($x_{i,-\theta-q}$ for all $0<q<m-\theta$) on a set of individual effects and the vector of observed or insample x's ($\underset{\sim}{x}_i$) to have zero coefficients on the last $T+\theta-p$ observed x's. This restriction captures the notion frequently found in time series models of economic variables that (conditional on the heterogeneity introduced by the individual specific effects) linear predictors of future x's

depend only on the x's observed in the recent past. The next section will consider estimating distributed lags from panel data when it is reasonable to restrict the lag parameters themselves.

We begin by considering the simplest case where there are no individual effects ($\alpha_i = \alpha$ for all i). Then if E^* is the "wide-sense" expectation, or projection,operator our restriction on the auxiliary regression of the unobserved or presample x's on $\underset{\sim}{x}_i$ is written as:[*3]

$$E^*(x_{i,-\theta-q} | \underset{\sim}{x}_i) = \sum_{j=1}^{p} \rho_j^{(q)} x_{i,j-\theta-1} \qquad (2)$$

for $0 < q < m-\theta$, and $i = 1,\ldots,N$.

If the x sequences are random drawings from a larger population of interest then a sufficient condition for (2) to be true is that in this population the distribution of $x_T, \ldots, x_{-\theta+p}$, given $x_{-\theta+p-1}, \ldots, x_{-\theta}$ and $x_{-\theta-q}$ is independent of $x_{-\theta-q}$ for all $0<q\leq m-\theta$; that is the distribution of future x's depends only upon the last p realizations of x. This does not require the process to be homogeneous over time. Thus changes in the economic environment may cause <u>different</u> functions of the last p observed values of x to be used as the predictors of future x's in different years. If, however, the x's are generated by a time-invariant process then

(2) will be true under the familiar assumptions of x's that follow an integrated autoregressive process where the order of the integration is k while the order of the autoregression is p-k.[4]

It will now be shown that if (2) is true then one can estimate lag coefficients from (1) without imposing any structure on either the form or the length of the lag distribution. The underlying logic of this point is really quite simple. Divide the x's appearing as independent variables in the equation determining y_t into the observed or insample x's and the presample x's. Now consider the projection of y_t on all the observed x's. This will not identify the desired lag coefficients due to the correlation between the presample and the insample x's. However it follows from (2) that the partial correlation of the presample x's with all but the last p observed values of x will be zero. Thus the distributed lag coefficients corresponding to the leading lag coefficients will be identified.

To be more explicit rewrite (1) as:

$$y_{i,t} = \sum_{\tau=0}^{t+\theta} w_\tau x_{i,t-\tau} + b_{i,t} + u_{i,t} \qquad (3)$$

where

$b_{i,t} = \sum_{\tau=\theta+1}^{m-t} w_{t+\tau} x_{i,-\tau}$, that is $b_{i,t}$ is the contribution of the presample x's to $y_{i,t}$, the truncation remainder in period t.

Now consider the projection of $y_{i,t}$ on the vector of all insample observations on x for individual i, $\underset{\sim}{x}_i$. It will contain the term $E^*[b_{i,t}|\underset{\sim}{x}_i]$ which is just a linear combination of the $E^*[x_{i,-q+\theta}|\underset{\sim}{x}_i]$. Dropping the subscript i, for convenience, we have from (2) above that:

$$E^*[b_t|\underset{\sim}{x}] = \sum_{j=1}^{p} \phi_{t,j-\theta-1} x_{j-\theta-1} \tag{4}$$

where

$$\phi_{t,j-\theta-1} = \sum_{q=1}^{m-(t+\theta)} w_{t+\theta+q} \rho_j^{(q)}$$

for $j = 1,\ldots,p$ and $t = 1,\ldots,T$.

The projection of any unobserved or presample x, and therefore of any linear combination of all the presample x's, on the insample x's has zero coefficients on all but the first p observed values of x. One can now use (4) to derive the projection of $\underset{\sim}{y}' = [y_1,\ldots,y_T]$ or $\underset{\sim}{x}$, that is;

$$E^*[y_t|\underset{\sim}{x}] = \sum_{\tau=0}^{t+\theta-p} w_\tau x_{t-\tau} + \sum_{j=1}^{p} \psi_{t,j-\theta-1} x_{j-\theta-1} \tag{5}$$

where

$$\psi_{t,j-\theta-1} = w_{t+\theta+1-j} + \phi_{t,j-\theta-1}$$

for

$$t = 1,\ldots,T.$$

(5) shows how the panel can be used to derive consistent estimates of the T + θ - p leading coefficients from an unconstrained lag distribution. A model is constructed where each y_t is regressed on all previous insample values of the x's. The last p values of x in each equation are correlated with the contribution of the

presample x's to the current value of y. Hence their coefficients do not provide consistent estimates of parameters from the lag distribution. The first $t + \theta - p$ regression coefficients estimated in the equation for y_t are, however, consistent estimates of the leading $t + \theta - p$ lag coefficients. As t increases, the lag structure we want to estimate is gradually built up.

Of course if $T + \theta$ is small relative to p we will not be able to build up much information on the tail of the lag distribution. This simply reflects the fact that short panels, by their very nature, do not contain unconstrained information on that tail. However, even in cases with small $T + \theta$ the initial consistent estimates of the first few lag coefficients provided by (5) may contain enough information on the lag structure to allow one to restrict it to be a member of a family of distributions which depend on a small number of parameters and concentrate on estimating these parameters thereafter.

If one is willing to assume that $E^*[\underset{\sim}{y}_i|\underset{\sim}{x}_i] = E[\underset{\sim}{y}_i|\underset{\sim}{x}_i]$ and that $V(\underset{\sim}{y}_i|\underset{\sim}{x}_i) = \Omega$, a constant matrix (this will be true, for example if, given (1), the x's distribute joint normally) then consistent and efficient estimators of (5) are extremely easy to obtain. In this case (5) is just a linear multivariate regression model and can be programmed into most existing econometric software packages

(such as TSP or SAS). Of course one can derive more efficient parameter estimates if, in addition, it can be assumed that the x's are generated by one of the simpler stochastic processes which generate equation (2). In this case there will be information in the observed x sequences which helps to estimate the lag coefficients and there will, in general, be a set of nonlinear constraints connecting the Ψ and the w coefficients. Then it will be reasonable to consider maximizing the joint likelihood of $\underset{\sim}{y}_i$ and $\underset{\sim}{x}_i$ rather than the conditional likelihood of $\underset{\sim}{y}_i$ given $\underset{\sim}{x}_i$, and to impose this set of nonlinear constraints.[5] If the assumptions that $E^*[\underset{\sim}{y}_i|\underset{\sim}{x}_i] = E[\underset{\sim}{y}_i|\underset{\sim}{x}_i]$ and $V(\underset{\sim}{y}_i|\underset{\sim}{x}_i) = \Omega$ are not reasonable for the problem at hand, then, though the maximum likelihood (or asymptotically equivalent) estimators discussed above will still provide consistent estimates of the lag coefficients, a more efficient estimator can be obtained by minimizing the distance function discussed in Chamberlain (1980).

How does the model of equations (5) fit into the general panel data analysis framework discussed in the introduction? Letting $\underset{\sim}{\Pi}$ be defined by the projection $E^*(\underset{\sim}{y}_i|\underset{\sim}{x}_i) = \underset{\sim}{\Pi}\underset{\sim}{x}_i$, then the model in (5) implies that

$$\underset{\sim}{\Pi} = \underset{\sim}{W} + \underset{\sim}{\Psi} \tag{6}$$

where $\underset{\sim}{W}$ is the matrix of lag coefficients that would be obtained if none of the truncation remainders were correlated with the

observed x's, that is;

$$\underset{\sim}{W} = \begin{bmatrix} 0 \ldots . 0 \; w_0 \ldots w_\theta & & \\ \vdots & \ddots & \\ 0 & & \\ w_0 \ldots\ldots\ldots\ldots & & w_{p+T-1} \end{bmatrix}$$

and

$\underset{\sim}{\Psi}$ contains the coefficients of the auxiliary regression of the truncation remainders on the observed x's; that is, given equation (2), $\underset{\sim}{\Psi}$ has zero vectors in the first $T+\theta-p$ of its columns;

$$\underset{\sim}{\Psi} = \begin{bmatrix} \overbrace{\underset{\sim}{0}}^{T+\theta-p} & \vdots & \overbrace{\underset{\sim}{\Psi}_s}^{p} \end{bmatrix}$$

Since the last p lag coefficients $(w_{T+\theta-p-1}, \ldots, w_{T+\theta-1})$ are not identified from (6), $\underset{\sim}{W} + \underset{\sim}{\Psi}$ contains only $T + p(T-1) + \theta$ parameters. As noted above, the unrestricted panel data model contains $(T+\theta)T$ of them, so that (5) places $T(T+\theta-p-1) + p - \theta$ testable restrictions on the $\underset{\sim}{\pi}$ matrix. Moreover, it follows from (6) that both p, the length of the memory of the x process, and m, the length of the lag, can be inferred from the properties of the observed $\underset{\sim}{\pi}$ matrix. p is determined by the number of non-zero columns of $\underset{\sim}{\psi}_s$; while both w_τ and the row vector $\underset{\sim}{\psi}_t$ should be close to zero for $\tau > m$ and $t > m$ respectively.[6]

If, in fact, there are correlated effects then equation (2) is no longer likely to provide a good approximation to the auxiliary regression of the presample on the insample x's. If y conditional

on all past x's has an individual specific mean, and the diverse unobserved factors which lead to that mean have an impact on the insample x's, then these same factors are likely to have an impact on presample x's. The simplest way for allowing for this phenomenon is to add a component to the auxiliary regression of the presample on the insample x's which does depend on α_i. Though this procedure would allow for individual specific differences in both the y and x sequences, it would restrict the differences in these two processes to be perfectly correlated. For many examples this would seem to be unduly restrictive, at least as an a priori assumption.[7]
A more general way for allowing for individual specific heterogeneity in the auxiliary regression of the presample on the insample x's is to introduce a second factor into that auxiliary regression, say m_i, and replace (2) with:

$$E^*[x_{i,-\theta-q}|\underset{\sim}{x}_i, \alpha_i, m_i] = c_q m_i + \sum_{j=1}^{p} \rho_j^{*(q)} x_{i,-\theta+j-1} \qquad (7)$$

for $0 < q \leq m - \theta$ and $i = 1,\ldots,n$.

One can now test whether in fact the unobserved heterogenity in the y and x processes are proportional by testing whether $m_i = \kappa\alpha_i$

A sufficient condition for (7) to be true is that the conditional distribution of $x_T,\ldots,x_{-\theta+p}$ given m, $x_{-\theta+p-1},\ldots,x_{-\theta}$ is independent of $x_{-\theta-q}$ for all $0 < q \leq m - \theta$; that is given the individual specific differences that cause heterogeneity in the process generating the x's, the distribution of future x's only depends on the last p observed values of x. Thus, although (7) does not require any form of stationarity, it will be true for many of the mixture models often used to describe the evolution of a

multiple time series of economic variables when there is a need to allow for individual specific heterogeneity. Two of the more familiar of these models occur when the process generating the x's consists of an individual specific mean and a deviate which follows a p^{th} order autoregression (Lillard and Weiss, 1979, use this process with p = 1) and when the x process consists of an individual specific mean plus a deviate which follows an integrated autoregressive process where the order of the integration is k while the order of the integration is p - k.[8] In what follows we will work with the simpler model where $m_i = \kappa\alpha_i$ and then indicate the extension required to deal with the case of free m_i's.

Given (7), a modified version of the argument which led to the identification of the $T + \theta - p$ leading lag coefficients from the distributed lag model without correlated effects can now be used to identify these coefficients in the presence of such effects. Project both sides of (1) onto $\underset{\sim}{x}_i$ and α_i, and define $\underset{\sim}{\zeta}_i = \underset{\sim}{y}_i - E^*[\underset{\sim}{y}_i|\underset{\sim}{x}_i,\alpha_i]$, where $E^*[\underset{\sim}{\zeta}_i|x_i,\alpha_i] = \underset{\sim}{0}$ by construction. Then using (7) we have:

$$y_{i,t} = c_t\alpha_i + \sum_{\tau=0}^{t+\theta-p} w_\tau x_{i,t-\tau} + \sum_{j=1}^{p} \psi^*_{t,j-\theta-1} x_{i,j-\theta-1} + \zeta_{it} \qquad (8)$$

where

$c_t = [1 + \sum_{q=1}^{m-t-\theta} w_{t+\theta+q}c_q]$ and the $\psi^*_{t,j-\theta-1}$ are defined analogously to the $\psi_{t,j-\theta-1}$ in equation (4) above.

Compared to traditional panel data models which allow for correlated effects, the only real novelty in (8) is that the correlated effect will have a coefficient, or a price, which changes

over time. Note that for $t > m - \theta$. There is no truncation remainder and therefore the coefficient of the unobserved α_i (c_t) will be unity. If, however, $T < m - \theta$ then (8) is a single factor model with free "factor loadings." Such a model is only identified up to a single normalization, and a normalization which is consistent with the possibility that $T > m - \theta$, and therefore the normalization we choose in what follows, is $c_T = 1$.

To see that the leading lag coefficients are in fact identified from (8) note that we can use y_{i1} as an error-ridden indicator of the α_i, and rewrite the system in (8) as:

$$y_{i,t} - \bar{c}_t y_{i,1} = \sum_{\tau=0}^{t-1} w_\tau x_{i,t-\tau} + \sum_{\tau=t}^{1+\theta-p} (w_\tau - \bar{c}_t w_{\tau-t+1}) x_{i,t-\tau} \quad (9)$$

$$+ \sum_{j=1}^{p} (\psi^*_{t,j-\theta-1} - \bar{c}_t \psi^*_{1,j-\theta-1}) x_{i,j-\theta-1} + \zeta_{it} - \bar{c}_t \zeta_{i1}$$

where

$$\bar{c}_t = c_t / c_1$$

for $t = 2,\ldots,T$.

(9) is a standard simultaneous equations model and it will identify the leading lag coefficients provided that there exists, for at least one t, an instrument which is excluded from the equation determining $y_t - \bar{c}_t y_1$, is uncorrelated with $\zeta_t - \bar{c}_t \zeta_1$, but is correlated with y_1. If $T \geq 3$, $x_3,\ldots,x_T$ are excluded from the equation determining $y_2 - \bar{c}_2 y_1$, and provided that they are correlated with the α_i, that is

provided there are correlated effects, these variables will possess the properties required of instruments.

(8) defines $E^*[\underset{\sim}{y}_i|\alpha_i,\underset{\sim}{x}_i]$. However, since α_i is not observed the matrix of coefficients from this projection cannot be directly estimated. The coefficient matrix which can be estimated is $E^*[\underset{\sim}{y}_i|\underset{\sim}{x}_i] = E^*_\alpha\{E^*[y_i|\alpha_i,\underset{\sim}{x}_i]|x_i\}$. Taking the wide-sense expectation of α given $\underset{\sim}{x}_i$ in (8) we find that:

$$E^*[y_{it}|\underset{\sim}{x}_i] = c_t[\sum_{r=-\theta}^{T}\alpha_r x_{i,r}] + \sum_{\tau=0}^{t+\theta-p} w_\tau x_{it-\tau} + \sum_{j=1}^{p}\psi^*_{t,j-\theta-1}x_{i,j-\theta-1} \tag{10}$$

where the α_r are defined by the projection

$$E^*[\alpha_i|\underset{\sim}{x}_i] = \sum_{r=-\theta}^{T}\alpha_r x_{ir}.$$

Here again if one is willing to assume that $E^*[\underset{\sim}{y}_i|\underset{\sim}{x}_i]=E[\underset{\sim}{y}_i|\underset{\sim}{x}_i]$ and $V(\underset{\sim}{y}_i|\underset{\sim}{x}_i) = \Omega$ (and, given equation (1), this will be true if α_i and the sequence of x's distribute joint normally) then maximum likelihood (or an asymptotically equivalent estimation procedure) will yield consistent and efficient estimates of the parameter in (10). Though (10) is nonlinear it can easily be programmed into several existent maximum likelihood estimation packages (see, for example, Joreskog and Sordom, 1976, and Hall, 1979) and so should not be too difficult to estimate in practice.

It should be noted, however, that maximum likelihood, if applied, ought to be applied to the model in (10) and not to the model in (8). The system in (8) contains a set of "incidental paramters," a set of parameters whose members enter into the probability laws governing a finite number of sample points, the numbers of members of that set growing with sample

size (the α_i), and therefore the maximum likelihood estimators of (8) do not have desirable properties. By going from (8) to (10) we have, in effect, constructed the density of $\underset{\sim}{y}_i$ conditional on $\underset{\sim}{x}_i$ and the vector $\underset{\sim}{\gamma}$ of all the structural parameters of the model (the parameters which enter into the density function governing the distribution of each sample point which includes the w_τ, the $\psi_{t,j-\theta-1}$, the c_t and the α_r parameters), but which is marginal to the model's incidental parameters. That is we have implicitly produced the density of $\underset{\sim}{y}_i$ conditional on $\underset{\sim}{x}_i$ and $\underset{\sim}{\gamma}$ by taking the density of $\underset{\sim}{y}_i$ conditional on $\underset{\sim}{x}_i$, $\underset{\sim}{\gamma}$ and the α_i, multiplying it by the density of α_i conditional on $\underset{\sim}{x}_i$ and $\underset{\sim}{\gamma}$, and then integrating out the α_i. The maximum likelihood estimator obtained from (10) are found by maximizing this latter density with respect to $\underset{\sim}{\gamma}$, and are therefore identical to the marginal maximum likelihood estimator for the incidental parameter problem introduced by Keifer and Wolfowitz (1956).[9,10]

If $E^*[\underset{\sim}{y}_i|\underset{\sim}{x}_i] \neq E[\underset{\sim}{y}_i|\underset{\sim}{x}_i]$ and/or $V(\underset{\sim}{y}_i|\underset{\sim}{x}_i)$ is not constant over individuals, then the maximum likelihood estimators of (10) though still consistent, will no longer be efficient. Once again we refer the reader to the minimum distance estimator discussed in Chamberlain (1980) to derive efficient parameter estimates in this case.

It will be useful to provide an explicit formula for the $\underset{\sim}{\Pi}$ matrix when correlated effects are allowed for. To do so we first relax the assumption that there is only a single unobserved factor that affects both y and the choice of x; that is relax the assumption that $m_i = \kappa\alpha_i$ for all i. To find the $\underset{\sim}{\Pi}$ matrix for the two factor model simply substitute (4) into

(1) and take $E^*_{\alpha,m}\{E^*[\underset{\sim}{y}_i|\underset{\sim}{x}_i, \alpha_i, m_i]|\underset{\sim}{x}_i\}$ of the resulting equations. In this model it will be the m_i which receives a variable coefficient over time and for $t > m - \theta$, that coefficient, say c_t, will equal zero. Thus a normalization consistent with the possibility that $T > m - \theta$ in the two factor model is $c_1^* = 1$. In the two factor model, however, we require two normalizations, so that we will also set $m_T = 1$, where m_T is defined by the projection $E^*[m_i|\underset{\sim}{x}_i] = \sum_{r=-\theta}^{T} m_r x_r$.[11] Given this notation the $\underset{\sim}{\Pi}$ matrix for the two factor model is written as:

$$\underset{\sim}{\Pi} = \underset{\sim}{W} + \underset{\sim}{\Psi}^* + \underset{\sim}{i}\underset{\sim}{\alpha}' + \underset{\sim}{c}^*\underset{\sim}{m}' \tag{12a}$$

where $\underset{\sim}{i}$ is a $T + \theta$ element unit vector $\underset{\sim}{W}$ and $\underset{\sim}{\Psi}^*$ are defined analogously to $\underset{\sim}{W}$ and $\underset{\sim}{\Psi}$ in equation 6, $\underset{\sim}{m}' = [m_{-\theta},\ldots m_{T-1}, 1)$ $\underset{\sim}{c}^{*\prime} = [1, c_2^*,\ldots, c_T^*]$, and $\underset{\sim}{\alpha}' = [\alpha_{-\theta},\ldots,\alpha_T]$.

If a one-factor model is an adequate summary of the unobserved heterogeneity in the y and x processes, that is if $m_i = \kappa\alpha_i$, then (12a) collapses into:

$$\underset{\sim}{\Pi} = \underset{\sim}{W} + \underset{\sim}{\Psi}^* + \underset{\sim}{c}\underset{\sim}{\alpha}' \tag{12b}$$

where $\underset{\sim}{c}' = [c_1,\ldots, c_{T-1}, 1]$.

The leading $T + \theta - p$ lag coefficients will be identified from (12a) provided $T \geq 4$, and from (12b) provided that $T \geq 3$. Neither of the models will ever identify the last p of the w, the α or the m coefficients.

Given these points it is straightforward to compare the $\underset{\sim}{\Pi}$ matrices in cases which do allow for unobserved heterogeneity in the x process [(12a) and (12b)] to the $\underset{\sim}{\Pi}$ matrix in the case which does not [(6)]. If there are no individual specific factors generating the x's then the elements of $\underset{\sim}{\Pi} = [\pi_{tr}]$ should be zero for all $r > t$, and should have a "stationary" structure, that is should depend only on $t - r$, for $r \geq p - \theta$. Thus, if there were no heterogeneity and we project each $y_{i,t}$ on $\underset{\sim}{x}_i$, then the "leads," the $x_{i,r}$ for $r > t$, should have small insignificant coefficients, and the coefficients of $x_{i,t-\tau}$ for $\tau > t+\theta-p$ should be approximately the same for different t. If there are individual specific differences in the process generating the x's then one would expect both significant leads and non-stationary. The two factor model (12a) allows this non-stationary to take on a more complicated form than the one factor model (12b).

More formally, for $T \geq 4$ all these distributed lag models are nested to the general panel data model with a free $\underset{\sim}{\Pi}$ matrix. Thus one would first test to see if (12a) were a reasonable approximation to the data, and then test whether, given (12a), (12b) is a simplification that the data accepts. If it does we would go on to test whether there was reason to allow for any unobserved heterogeneity in the auxiliary regression of the presample on the insample x's; that is we would test (6) maintaining the null hypothesis in (12a).

Though all our discussion has been concerned with a distributed lag model which contains only a single regressor, it is straightforward to generalize our results to the case

of many regressors. If there were no individual specific effects the model for the (multivariate) auxiliary regression of the vectors of the presample $\underset{\sim}{x}$'s at different dates on the insample $\underset{\sim}{x}$'s would be $E^*[\underset{\sim}{x}_{i,-\theta-q}|\underset{\sim}{x}_{i,-\theta},\ldots,\underset{\sim}{x}_{i,T}] = E^*[\underset{\sim}{x}_{i,-\theta-q}|\underset{\sim}{x}_{i,-\theta},\ldots,\underset{\sim}{x}_{i,-\theta+p}]$. A time inhomogeneous p^{th} order vector autoregression, or integrated vector autoregression where the sum of the order of the integration plus the order of the autoregression was p, would be examples of stochastic processes which generate this auxiliary regression. Similarly, if we were to allow for individual specific heterogeneity in the process generating $\underset{\sim}{x}$, the auxiliary regression would be written as $E^*[\underset{\sim}{x}_{i,-\theta-q}|\underset{\sim}{\alpha}_i,\underset{\sim}{x}_{i,-\theta},\ldots,\underset{\sim}{x}_{i,\tau}] = E^*[\underset{\sim}{x}_{i,-\theta-q}|\underset{\sim}{\alpha}_i,\underset{\sim}{x}_{i,-\theta+p}]$.[12] In either of these cases the $\underset{\sim}{\Pi}$ matrix would allow for free coefficients on the earliest p values of each regressor in each year. This would be true whether or not there was more than one distributed lag in the original model.

We began this paper by noting that the recent proliferation of panel data sets holds out the promise of estimating lag coefficients which can be identified more precisely than had been possible in the past. A major problem with using these panels to estimate distributed lags is that they are characteristically rather short. Thus some assumption must be made on the relationship between the truncation remainder and the observed x's before any lag coefficients can be consistently estimated from panel data. This section showed that if the distribution of future x's conditional on all past x's (or if there is heterogeneity, then conditional on all past x's plus the source of this heterogeneity) depends only on the last p realizations of x (or of the last p realizations plus the source of the heterogeneity) then the panel can be used to

estimate the leading $T + \theta - p$ lag coefficients from an unconstrained lag distribution. The assumption on the x process ought to be a fair approximation for most economic time series.[13] Thus the procedure outlined in this section should prove useful when there is little prior information on the structure of the lag coefficients. There is the problem, though, that in short panels this procedure will not allow one to infer much about the tail of the lag distribution. Unfortunately, short panels, by their very nature, do not contain unconstrained information on this tail. If, on the other hand, one is willing to impose enough prior structure on this tail then panel data can be used to estimate the entire lag distribution. As we show in the next section, when such prior information on the lag coefficients is available there is less need to assume a special structure for the relationship between the presample and the insample x's.

III. Distributed Lags from Panel Data When the Truncation Remainder Follows an Exact Autoregression

In this section we describe a technique for estimating distributed lag coefficients from panel data when, after a few free lags, the lag distribution can be assumed to follow an exact autoregression. We focus on this lag structure because of its previous wide use (see, for example, Nerlove [1972], Dhrymes [1971] and Griliches [1967]). The simplest example is, of course, the modified geometric or Koyck lag which can be written as:

$$w_\tau = \begin{cases} w_\tau & \text{for } \tau \leq \theta \\ \delta w_{\tau-1} & \text{for } \tau > \theta \end{cases} \qquad (13)$$

and $|\delta| < 1$.[14]

If we use (13) to constrain the coefficients in the general distributed lag model without correlated effects (equation [1]), it is clear that with the modified geometric lag structure the contribution of the presample x's to the current value of y, or the truncation remainder, becomes an unobserved factor which follows an exact first order autoregression (i.e., $b_{it} = \lambda b_{i,t-1}$). That is given (13), (1) may be written as:[15]

$$y_{i,t} = \alpha + \sum_{\tau=0}^{\theta} w_\tau x_{i,t-\tau} + w_\theta \sum_{\tau=\theta+1}^{\theta+t} \delta^{\tau-\theta} x_{i,t-\tau} + \delta^{t-1} b_i + u_{i,t} \qquad (14)$$

where

$$b_i = w_\theta \sum_{\tau=\theta+1}^{\infty} \delta^{\tau-\theta} x_{i,1-\tau},$$

that is b_i is the truncation remainder in period (1).

Generalizations to (13) which allow the lag coefficients to follow second or higher order autoregressions, and therefore allow for more flexible lag structures, are written as:

$$w_\tau = \begin{cases} w_\tau & \text{for } \tau \leq \theta \\ \sum_{q=1}^{Q} \delta_q w_{\tau-q} & \text{for } \tau > \theta \end{cases} \qquad (15)$$

where the pairs of conjugate roots to the equation $1 - \sum_{q=1}^{Q} \delta_q \lambda^q$ all lie outside the unit circle. If (15) is substituted into the distributed lag model for panel data the contribution of presample x's to the current value of y can be shown to follow an exact Q^{th} order autoregression. For our purposes, however, it is simpler to treat the truncation remainder in this case as a system of Q difference equations each of which follows an exact first order autoregression. Since the econometric issues involved in estimating the model with Q geometrically decaying factors are, for the most part, straightforward extensions of the model with a single factor, we will concentrate on the simpler model below. The model obtained from using (15) to restrict the lag parameter will be identified provided that $T \geq Q+1$ when there are no correlated effects, and provided that $T \geq Q+2$ when there are such effects.[16]

The problems which arise in estimating (modified) geometric lag structures from panel data differ somewhat from the problems which arise in estimating them from a single time series of observations on x and y. In the context of a single time series the relevant limiting dimension for deriving the asymptotic

properties of one's estimators is T, and as T gets longer the contribution of the initial value of the truncation remainder, b_i, to $\sqrt{T}^{-1}$ times the derivative of the log-likelihood function goes to zero. Thus the value chosen for b_i will not, in this case, affect the maximum likelihood (ML) estimates of the main parameters of interest [see Dhrymes, 1971]. This conclusion, however, is crucially dependent on the length of T.[17] In the panel data context T is characteristically quite short and therefore the contribution of b_i to the likelihood of any given $\underset{\sim}{y}_i$ is non-negligible. Of course, the relevant limiting dimension in panel data is N and as N grows large there is a set of the b_i; but given the fact that T is short it is not surprising that the properties of their <u>distribution</u> will affect the properties of the estimators of the lag parameters.

We begin our discussion of estimating equation (14) by providing a simple, consistent (as N grows large), estimator of its parameters. Later, it will be shown that if it is reasonable to assume that $(\underset{\sim}{x}_i, \underset{\sim}{y}_i, b_i)$ distribute joint normally, then this estimator will also be asymptotically efficient.

Subtract the equation for δy_{t-1} from that for y_t for $t = 2,\ldots,T$ in equation system (14). This panel data modification of Liviatan's (1963) suggestion for estimating distributed lags produces the system of equations:

$$y_t - \delta y_{t-1} = w_0 x_t + \sum_{\tau=1}^{\Theta} (w_\tau - \delta w_{\tau-1}) x_{t-\tau} + u_t - \delta u_{t-1} \qquad (16)$$

for $t = 2,\ldots,T$.

If one adds the projection of y_1 onto all the insample x's (and allows the disturbance from that projection to be freely correlated with the $\varepsilon_t - \delta\varepsilon_{t-1}$) to (16) then this system becomes a standard simultaneous equations model in the parameters $w_0, w_1 - \delta w_0, \ldots, w_{\theta+1} - \delta w_\theta, \delta$, the coefficients of the y_1 projection, and the variance covariance matrix of disturbances. These parameters will be identified provided there exists at least one variable which is excluded from the equation determining y_t, uncorrelated with the disturbances, and is correlated with y_{t-1}. However, if $T > 1$, that is if the proxy for b_t is observed for at least one t, then $x_{t-\theta-1}$ is observed for all $t > 1$ and satisfies the requirements of an appropriate instrument. Thus as long as $T > 1$ one can obtain consistent, as N grows large, estimates of $w_0, w_1 - \delta w_0, \ldots, w_0 - \delta w_{\theta-1}, \delta$ and a one to one transformation of these estimators provides consistent estimates of the parameters needed to define the entire lag distribution.

Had we started out with the model which allowed for correlated effects then one would begin by first differencing the y sequence to eliminate the α_i, and then use $\delta(y_{t-1} - y_{t-2})$ as an indicator for $b_t - b_{t-1}$ in the equation determining $y_t - y_{t-1}$. The system of equations derived in this manner is:

$$(y_t - y_{t-1}) - \delta(y_{t-1} - y_{t-2}) = w_0 x_t + [w_1 - w_0(1+\delta)]x_{t-1} + \sum_{\tau=2}^{\theta} [w_\tau - (1+\delta)w_{\tau-1} + w_{\tau-2}]x_{t-\tau} + u_t - (1+\delta)u_{t-1} + \delta u_{t-2} \qquad (17)$$

for $t = 3, \ldots, T$.

Now one must add the equations for the projections of both $y_2 - y_1$ and y_1 onto all the insample x's to the system in (17), and allow the disturbances from these projections to be freely correlated with the other disturbances in the system in order to derive a standard simultaneous equations model in the parameters δ, w_0, $w_1 - w_0(1+\delta)$, $w_2 - (1+\delta)\, w_1 + w_0, \ldots,$

$w_\theta - (1+\delta)w_{\theta-1} - w_{\theta-2}$, the projection coefficients, and the variance-covariance matrix of disturbances. This system will be identified provided that the proxy for $b_t - b_{t-1}$, that is $y_{t-1} - y_{t-2}$, exists for at least one t, that is provided $T > 2$.[18]

We now consider the implications of the modified geometric lag structure on the $\underset{\sim}{\Pi}$ matrix. For these models the $\underset{\sim}{\Pi}$ matrix is obtained by substituting equation (13) into (1) and then finding the projection of the resulting system of equations onto $\underset{\sim}{x}_i$, or by finding $E^*_{\alpha,b}\{E^*(\underset{\sim}{y}_i|\underset{\sim}{x}_i, \alpha_i, b_i)|\underset{\sim}{x}_i\}$. For the model without correlated effects ($\alpha_i = \alpha$ for $i = 1,\ldots,N$) the $\underset{\sim}{\Pi}$ matrix will be the sum of two terms, one of which, say $\underset{\sim}{W}^*$, contains the matrix of coefficients that would be obtained from that projection if the truncation remainders were uncorrelated with the observed x's, while the other contains the coefficients obtained from the projection of the truncation remainders onto $\underset{\sim}{x}_i$. Since $b_{i,t} = \delta b_{i,t-1}$ for all i and t, the subsequent rows of the latter matrix will all be proportional to each other with factor of proportionality δ. Thus in this case:

$$\underset{\sim}{\Pi} = \underset{\sim}{W^*} + \underset{\sim}{\delta}\underset{\sim}{\beta}'$$

where

$$\underset{\sim}{W^*} = \begin{bmatrix} 0 \ldots . 0 \; w_0 \ldots . w_\theta & & & \\ 0 \ldots 0 w_0 \ldots . . w_\theta \delta w_\theta & & & \\ 0 \cdot \cdot \cdot & & & \\ w_\theta \cdot \ldots w_\theta \delta w_\theta \ldots \delta^{T-1} w_\theta & & & \end{bmatrix} \tag{18}$$

and

$$\underset{\sim}{\delta}' = [1, \delta, \delta^2, \ldots . \delta^{T-1}]$$

$$\underset{\sim}{\beta}' = [\beta_{-\theta}, \ldots . . . . , \beta_T]$$

where the β's are defined by the projection

$$E^*[b_i | \underset{\sim}{x}_i] = \sum_{r=-\theta}^{\tau} \beta_r x_{i,r} \quad .$$

Note that the model leading to (18) does not restrict the auxiliary regression of the presample onto the insample x's in any way. Thus, in general, if we projected each y_{it} onto the $\underset{\sim}{x}_i$ we would expect to pick up both nonzero leads and a nonstationary coefficient matrix simply because we do not have enough lagged x's in the estimating equation. If, however, it were reasonable to assume that the presample x's have a nonzero partial correlation with only the earliest p observed values of x (equation (2) of the last section) then, it can be shown that $\beta_T = \beta_{T-1} = , \ldots , = = \beta_{-\theta+p} = 0$. In this case the coefficients of the x's which lead y ought to be near zero, and (18) will be nested to (6), the $\underset{\sim}{\Pi}$ matrix for the model which allows for a free lag distribution, but assumes that the x's have the Markovian property of equation (2).

If we allow for correlated effects in the model which assumes a modified geometric lag structure, then an additional term must be added to the $\underset{\sim}{\Pi}$ matrix in (18) in order to pick up the correlation of α_i and the observed x's. Thus, in this case we have:

$$\underset{\sim}{\Pi} = W^* + \underset{\sim\sim}{\delta\beta}' + \underset{\sim\sim}{i\alpha}'$$

where

$\underset{\sim}{i}$ is the $T + \Theta$ element unit vector

and

$$\underset{\sim}{\alpha}' = [\alpha_{-\Theta}, \ldots, \alpha_T] \tag{19}$$

where the α's are defined by the projection

$$E^*[\alpha_i | \underset{\sim}{x}_i] = \sum_{r=-\Theta}^{T} \alpha_r x_{i,r} .$$

Once again (19) imposes no restrictions on the auxiliary regression of presample on insample x. It is interesting to note however, that, even so, (19) is a restricted version of (12a), that is of the model which allows for a free lag structure but assumes that the auxiliary regression of the presample on the insample x's and the m_i (the unobserved individual specific heterogeneity in the x-process) to have zero coefficients on all but the earliest p observed values of x (equation (7) of the last section). This, however, would not be true if we were to allow for a more complicated structure for the lag coefficients

(equation (15) with $Q > 1$). Still, provided that the auxiliary regression of the presample x's on the insample x's and the m_i is well approximated by (7), then the models generated by (15) will always be nested to (12a); and we will be able to derive more precise estimates of the lag coefficients in those models by imposing the restrictions implied by (7). With $Q = 1$ these restrictions are $\beta_T = \beta_{T-1} = ,\ldots, = \beta_{-\theta+p} = 0$.

Next consider maximum likelihood estimators of the parameters of the model. These estimators will be efficient if $E^*[\underset{\sim}{y}_i|\underset{\sim}{x}_i] = E[\underset{\sim}{y}_i|\underset{\sim}{x}_i]$ and $V[\underset{\sim}{y}_i|\underset{\sim}{x}_i]$ is a constant matrix; they will be efficient, then, if, given (1), $\underset{\sim}{x}_i$, α_i and b_i distribute joint normally. (19) contains several nonlinear constraints on the parameters of interest, and therefore maximum likelihood estimates of the $\underset{\sim}{\Pi}$ matrix in this case will be difficult to obtain. Thus, estimation will be simpler if we go back tc the transformation in equation (17),that is if we consider the model generated by $\underset{\sim}{y}^*_i = [y_{i,1}, y_{i,2} - y_{i,1}, y_{i,3} - y_{i,2} - \delta(y_{i,2} - y_{i,1}),\ldots,$ $y_{i,T} - y_{i,T-1} - \delta(y_{i,T-1} - y_{i,T-2})]$. Note that $\underset{\sim}{y}^*_i = A\underset{\sim}{y}_i$ where the determinant of A is independent of the parameters of the model (it equals unity), and recall both that there are no nonlinear constraints connecting the parameters defining the distribution of $\underset{\sim}{y}^*_i$, and that a one to one transformation of those parameters defines all the parameters of the model. Thus efficient estimates of the model's parameters can be obtained by maximizing the likelihood of $\underset{\sim}{y}^*_i$, and that can be done in any standard simultaneous equations package.[19] Two additional points are worthy of note.

First the likelihood function we are maximizing is the likelihood of $\underset{\sim}{y}_i$ conditional on $\underset{\sim}{x}_i$ but marginal to the α_i and the b_i, that is marginal to the two sets of incidental parameters in the model. Since no incidental parameters appear in this likelihood, the estimators we obtain will possess all the standard desirable properties of maximum likelihood estimators (see Keifer and Wolfowitz (1956)). Second, in the panel data context the distribution of the truncation remainders, or of the "impact of past history" will affect the ML estimations of the lag parameters unless $\underset{\sim}{\beta} = 0$, that is unless b_i has a zero partial correlation with each observed x, a condition which is very close to the requirement that the x's be serially uncorrelated. Finally, once again we note that if $E^*[\underset{\sim}{y}_i|\underset{\sim}{x}_i] \neq E[\underset{\sim}{y}_i|\underset{\sim}{x}_i]$ and/or $V[\underset{\sim}{y}_i|\underset{\sim}{x}_i]$ depends on the index i, then more efficient estimates of the models parameters can be obtained by using the minimum distance estimation described in Chamberlin (1980).

FOOTNOTES

[1] Thus all equations could be interpreted as applying to variables that have been prefiltered to eliminate all time specific means.

[2] If $\theta \leq m \leq T+\theta$ then, without further restrictions, only the fraction $m-\theta/T$ of the NT observations available could be used to derive consistent estimators of the lag coefficients. That is, one would have to drop $m-\theta$ observations for each of the N individuals.

[3] Equation (2) may have a set of time-specific constant terms that will be removed by filtering out year specific averages, see footnote 2, above.

[4] There is, of course, the issue of determining whether (2) is a good approximation for the problem one is investigating. Formally the only tests of (2) which will be considered in this paper are tests which are conditional on the distributed lag model we are estimating, that is on equation (1) in the text. Given (1), (2) implies certain restrictions on the multivariate projection of $\underset{\sim}{y}_i$ on $\underset{\sim}{x}_i$ (see the discussion below). If, however, one is willing to assume that the leads which receive zero coefficients in the projections of past x's onto future x's are independent of time then one could test (2) rather directly, by regressing the early observed values of x on the later ones. This type of preliminary analysis of the properties of the x-sequence is likely to be quite useful as it will also suggest a plausible range of values for p.

[5] In all these cases the joint density of $\underset{\sim}{y}_i$ and $\underset{\sim}{x}_i$ is obtained rather straightforwardly by multiplying the density of $\underset{\sim}{y}_i$ conditional on $\underset{\sim}{x}_i$, (the density of the model in (5)), and the marginal density of the $\underset{\sim}{x}_i$, which will depend on the process assumed to generate the x's. For example, if the x's can be assumed to be generated by a stationary autoregressive process of order p, with partial correlations $\rho_1,\ldots,\rho_p$, and variance σ^2, then one simply multiplies the density of the model in (5) by the density of N realizations of $T+\theta$ consecutive observations on an AR(p) process; see Anderson, 1978. In the autoregressive case Box and Jenkins, 1976, provide the formula for the 'backcasting' function which determines the $\rho_j^{(q)}$ coefficients in equation (2), and therefore the nonlinear constraints in the model. These coefficients will follow the recursion $\rho_j^{(q+1)} = \rho_j\rho_1^{(q)} + \rho_{j+1}^{(q)}$, where it is understood that $\rho_j^{(\cdot)} = 0$ for all $j>p$. The p initial conditions required to solve this system are $\rho_j^{(1)} = \rho_j$, for $j = 1,\ldots,p$.

[6] In any given application one is likely to have fairly narrow a priori bounds on both p and m. For p the bounds can be derived from the properties of the observed x-sequences (see footnote 4 above); while one can only determine whether m is a member of the closed set $[0, T + \theta - p]$ or if it is outside of it. Thus maximum likelihood testing criteria for the choice of p and m ought to be adequate in most cases. If, however, p and/or m can take on a fairly wide range of values then these criteria will tend to choose values of p and m that are too large in the sense that the probability of choosing the true model will be assympto-

tically bounded away from unity. Gewke and Meese (1979) discuss this point in more detail and provide alternative testing criteria for the more difficult cases.

[7] In some cases, however, it will not. Take the example where the relationship of interest is a Cobb-Douglas production function, y is output, x is (perhaps a vector of) input, and the α_i are interfirm productivity differences that are fairly constant over time, known to the manager of the firm, but unobserved by the econometrician. Then the theory of derived demand for inputs implies that the heterogeneity in the x and the y processes are exactly proportional.

[8] Both these processes may have time specific means, as can equation (7), since these means will be filtered out together with the year-specific averages.

[9] The classic reference for the incidental parameter problem is Neyman and Scott (1948). In our case the estimates of $\underset{\sim}{\gamma}$ found by maximizing the likelihood of $\underset{\sim}{y}_i$ condtional on $\underset{\sim}{x}_i$, $\underset{\sim}{\gamma}$ and α_i with respect to $\underset{\sim}{\gamma}$ and α_i will be inconsistent in $\underset{\sim}{\gamma}$ as well as in the α_i. Intuitively, the reason this occurs is that the conditional maximum likelihood estimator is based on finding certain weighted individual specific means of all variables and then maximizing a model based on deviations from these means. For short panels, however, the new variables, or the deviations from the means, will, by construction, have moving average

components. Once such a component exists presample values of the independent variables will have a nonzero partial correlation with all insample values of the independent variables.

[10] If it is reasonable to assume that the x's are generated by one of the simple mixture processes which lead to equation (7) then more efficient parameters estimates can be obtained by maximizing the joint likelihood of $\underset{\sim}{y}_i$ and $\underset{\sim}{x}_i$ with respect to $\underset{\sim}{\gamma}$, and by imposing certain nonlinear constraints between the various elements of the $\underset{\sim}{\gamma}$ vector. For example, if the x's are generated by a mixture process consisting of individual specific means and a p^{th} order autoregressive deviate, then the constraints discussed in footnote, 5, above continue to hold and, in addition, $\alpha_r = \alpha$ for $T-p>r>p-\theta$, while $\alpha_{-\theta+j} = \alpha_{T-j}$ for $0 \leq j \leq p$. The constraints on the α_r are obtained by noting that in this case the vector $\underset{\sim}{\alpha}$ will be proportional to $\underset{\sim}{\Sigma}^{-1}\underset{\sim}{i}$, where $\underset{\sim}{\Sigma}$ is the covariance matrix of the vector $[x_{-\theta},\ldots,x_T]$ and $\underset{\sim}{i}$ is the $T+\theta$ unit vector. Here $\underset{\sim}{\Sigma} = \underset{\sim}{\Sigma}^* + \sigma^2_\alpha \underset{\sim}{ii}'$ where $\underset{\sim}{\Sigma}^*$ is the covariance matrix of $T+\theta$ consecutive observations on a p^{th} order autoregressive process. A simple extension of the results of Lillard and Weiss, 1979, proves that $\underset{\sim}{\Sigma}^{-1} = (\underset{\sim}{\Sigma}^{*-1} - \theta_1\sigma^2_\alpha \underset{\sim}{\Sigma}^{*-1}\underset{\sim}{ii}'\underset{\sim}{\Sigma}^{*-1}\underset{\sim}{ii}'\underset{\sim}{\Sigma}^{*-1})^{-1}$; from which it follows that $\underset{\sim}{\Sigma}^{-1}\underset{\sim}{i} = \underset{\sim}{\Sigma}^{*-1}\underset{\sim}{i}\theta_1$ where $\theta_1 = (1 + \sigma^2_\alpha \underset{\sim}{i}'\underset{\sim}{\Sigma}^{*-1}\underset{\sim}{i})^{-1}$. The formulae for the elements of $\underset{\sim}{\Sigma}^{*-1}$ are provided in Galbraith and Galbraith, 1974.

[11] To see the need for this second normalization note that we can always write $m_i = \kappa\alpha_i + l_i$ where $\mathrm{Cov}(l_i, \alpha_i) = 0$ by construction. Projecting both sides of this equation on $\underset{\sim}{x}_i$ we have

$\sum_{r=-\theta}^{T} m_r x_r = \kappa \sum_{r=-\theta}^{T} \alpha_r x_r + \sum_{r=-\theta}^{T} l_r x_r$, so that $m_r = \kappa\alpha_r + l_r$ for $r = -\theta,\ldots,T$. Now from the covariance restriction it is clear that for any variance-covariance matrix of $\underset{\sim}{x}$ and any vector $\underset{\sim}{\alpha}$, $\sum_{r=-\theta}^{T} \sum_{r=-\theta}^{T} \alpha_j l_r \operatorname{Cov}(x_{i,r}, x_{i,j}) = 0$. That is, given $V(\underset{\sim}{x})$ and given $\underset{\sim}{\alpha}$ the model is only free to choose $T+\theta-1$ elements of $\underset{\sim}{l}$, and therefore $T+\theta-1$ elements of $\underset{\sim}{m}$.

12 Here, one would undoubtedly like to simplify and see whether it were reasonable to allow for a smaller number of separate unobserved factors. This is the multivariate analog of the problem of testing whether $m_i = \kappa\alpha_i$ in (7), and is discussed in more detail in Keifer (1979).

13 Of course if a different assumption on the x process were relevant then one could work out its implications on the auxiliary regression of the presample on the insample x's. Our point is simply that once an appropriate approximation to the x process is found, then one can frequently use it to provide consistent estimates of lag coefficients from an unconstrained lag distribution.

14 $|\delta| < |$ is required for the finiteness of the truncation remainder for reasonable assumptions on the x sequence, while $0 \leq \delta < 1$ may be required for lag structure to have a sensible economic interpretation.

[15] We begin, as we did in the last section, with the simpler model which does not allow for correlated effects.

[16] The form of the model with Q geometrically decaying factors is obtained by solving the difference equation in (15) for its roots and its initial conditions, and then substituting that solution into equation (1). The identification conditions can be derived from the matrix of coefficients obtained from the projection of $\underset{\sim}{y}_i$ onto $\underset{\sim}{x}_i$. Intuitively, one cross-section is required to estimate the projection of each of the initial values of the factors onto $\underset{\sim}{x}_i$, one cross-section is required to estimate the projection of α_i onto $\underset{\sim}{x}_i$, and one cross-section is required to estimate the parameters of the lag distribution.

[17] When T is short the particular way one specifies the value of b_i will effect the ML estimates of the main parameters of interest even in the time series context. See the articles by Pesaran (1973) and Glesjer (1977).

[18] In fact, the system is overidentified with $T > 2$ the number of overidentifying restrictions being $(T + \Theta)(T - 2) - (\Theta + 2)$. In the model without correlated effects (equation 16) the number of overidentifying restrictions is $(T - 1)(T + \Theta) - (\Theta + 2)$.

[19] Maximum likelihood of the parameters of the modified geometric lag structure model without correlated effects, the model in (18), are derived in a similar fashion. The simplifying transformation in this case is defined in (16), and the model in (18) is, of course, nested to the model with correlated affects.

References

Anderson, T.W. (1978), "Repeated Measurements on Autoregressive Processes," Journal of the American Statistical Association, June, Vol. 73, #362; Theory and Methods Section, . 371-378.

Box, G.E.P. and G.M. Jenkins, (1970), Time Series Analysis: Forecasting and Control, Holden-Day Inc., San Francisco.

Chamberlain, Gary (1978), "Analysis of Covariance with Qualitative Data," Harvard Institute of Economic Research Discussion Paper #678.

Chamberlain, Gary (1980), "Multivariate Regression Models for Panel Data," June 1980, Unpublished paper, presented at the NBER Summer Conference on Productivity.

Dhrymes, Phoebus J. (1971), Distributed Lags, Holden-Day, San Francisco.

Galbraith, R.F. and J.I. Galbraith (1974), "On the Inverses of Some Patterned Matrices Arising in the Theory of Stationary Time Series," Journal of Applied Probability, Vol. 11, pp. 63-71.

Gewke, John, and Richard Meese, (1979), "Estimating Regression Models of Finite but Unknown Order", Social System Research Institute, The University of Wisconsin, Madison, Discussion Paper #7925.

Glejser, H., (1977), "On Two New Methods to Deal with Truncation Remainders in Small Sample Distributed Lag Models with Autocorrelated Disturbances," International Economic Review, Vol. 18, No. 3, pp. 783-789.

Griliches, Zvi, (1967), "Distributed Lags: A Survey," Econometrica, Vol. 35, No. 1, pp. 16-49.

Griliches, Zvi, (1974), "Errors in Variables and Other Unobservables," Econometrica, Vol. 42, pp. 971-988.

Hall, Bronwyn H. ,(1979), Moments , the Moment Matrix Processor's Users Manual, Version 1.1, Stanford California.

Joreskog, K.G. and D. Sordom, (1976), Estimation of Linear Structural Equation Systems by Maximum Likelihood Methods -- A Fortran IV Program. National Education Resources Inc.

Keifer, N.M., (1979),"Regression Systems with Fixed Effects with a Factor Structure," May, Unpublished paper.

Keifer, J. and J. Wolfowitz (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," Annals of Mathematical Statistics, Dec., Vol. 27, No. 4, pp. 887-906

Klein, L.R., (1958), "The Estimation of Distributed Lags," Econometrica, Vol. 26, No. 4, pp. 553-565.

Lillard, Lee A. and Yoram Weiss (1979) "Components of Variation in Panel Earnings Data: American Scientists," Econometrica, Vol. 47, No. 2, pp. 437-454.

Liviatan, N., (1963), "Consistent Estimation of Distributed Lags," International Economic Review, Vol. 4, pp. 44-52.

Madalla, G.S. and Rao, A.S. (1971), "Maximum Likelihood Estimation of Solow's and Jorgenson's Distributed Lag Models," Review of Economics and Statistics, February, pp. 80-88.

Nerlove, Marc (1972) "Lags in Economic Behavior," Econometrica. Vol. 40, No. 2, pp. 221-225.

Neyman, Jerzy and Scott, E.L. (1948), "Consistent Estimates Based on Partially Consistent Observations," Econometrica, Vol. 16, #1, pp. 1-32.

Pesaran, M. Hashem (1973), "The Small Sample Problem of Truncation Remainders in the Estimation of Distributed Lag Models with Autocorrelated Errors," International Economic Review, Vol. 14, No. 1, pp. 120-131.

Sims, Christopher (1974), "Distributed Lags," in Intriligator and Kendrick (eds.) Frontiers of Quantitative Economics, Vol. II, North-Holland, Amsterdam and New York.