A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables

Jeffrey M. Wooldridge*

Department of Economics, Michigan State University, East Lansing, MI 48824-1038, USA

Received 5 August 1999; accepted 21 October 1999

Abstract

I show how to construct the likelihood function for the conditional maximum likelihood estimator in dynamic, unobserved effects models where not all conditioning variables are strictly exogenous. The method for handling the initial conditions problem appears to be novel, and offers a flexible, relatively simple alternative to existing methods. © 2000 Elsevier Science S.A. All rights reserved.

Keywords: Panel data; Dynamic model; Unobserved effects; Initial conditions

JEL classification: C33

1. Introduction

In most applications of dynamic, unobserved effects panel data models, conditioning variables other than lagged dependent variables are assumed to be strictly exogenous, if not fixed in repeated samples. As discussed in Wooldridge (1997) in the context of multiplicative, unobserved effects panel data models, strict exogeneity generally rules out feedback from unexpected movements in the outcome variable to future values of the explanatory variables. As the explanatory variables are often choice variables of an economic agent, strict exogeneity is sometimes untenable.

As an example, consider an unemployment duration model. A variable such as a person’s age clearly meets the strict exogeneity requirement because changes in age have nothing to do with unemployment duration, or anything else. But other variables typically included as controls, such as marital status, may depend in part on past unemployment spells. Ham and Lalonde (1996) consider

*Tel.: +1-517-353-5972; fax: +1-517-432-1068.
E-mail address: wooldril@pilot.msu.edu (J.M. Wooldridge)
unemployment and employment duration models where the key explanatory variable is participation in a job training program. As Ham and Lalonde point out, while program eligibility reasonably can be assumed to be strictly exogenous (because it is randomly assigned), actual participation cannot.

In specific cases, it is fairly well known how to deal with lack of strict exogeneity, although the focus has largely been on lagged dependent variables under the assumption that other conditioning variables are strictly exogenous. In a linear, additive model, one typically chooses a transformation, such as first differencing, to remove the unobserved effect, and then chooses instruments based on sequential conditional moment assumptions. See, for example, (Holtz-Eakin et al., 1988) and Ahn and Schmidt (1995).

Chamberlain (1992) and Wooldridge (1997) contain transformations and instrumental variable methods that can be used for multiplicative panel data models; these methods do not require full distributional assumptions.

Apparently, there is no general treatment for how to handle feedback when, as is often necessary to facilitate empirical work in nonlinear models, distributional assumptions are made on the dynamic distributions. This note attempts to fill this gap at a general level. A useful innovation is the treatment of the initial conditions.

2. Constructing the conditional log-likelihood function

The setup is as follows. Let $y_t$ be a $G$-vector of response variables that we would like to explain; typically, $y_t$ is a scalar, such as a discrete variable indicating labor force status, or a count variable for number of patents awarded to a firm, but the general case is no more difficult. Let $w_t$ be a $J$-vector of explanatory variables that, as we will see, are not assumed to be strictly exogenous. The $K$-vector $z_t$ is a vector of strictly exogenous explanatory variables, in a sense to be made precise. All of these are observed in all time periods.

We assume that the conditional distributions of interest are

$$D(y_t|w_t, z_t, X_{t-1}, c), \quad t = 1, 2, \ldots, T,$$  

(1)

where $X_{t-1} = (x_{t-1}, \ldots, x_0)$, $x_t = (y_t, w_t)$, and $c$ is unobserved heterogeneity. We assume that we observe $x_0 = (y_0, w_0)$, which simply means that $t=0$ is the first period we have observations on $y$ and $w$. (Because we will assume that $z_t$ is strictly exogenous, $z_t$ can contain lags of variables; we just represent contemporaneous and lagged values by $z_t$ to simplify the notation.)

If $w_t$ were not in the model, (1) would be the basis for a standard dynamic unobserved effects analysis with strictly exogenous variables (and $X_{t-1}$ would simply be $Y_{t-1} = (y_{t-1}, \ldots, y_0)$). But, without assuming strict exogeneity of $w_t$, we need to model its conditional distribution, $D(w_t|Z_T, X_{t-1}, c)$. The sense in which $\{z_t\}$ is strictly exogenous is

$$D(y_t|w_t, Z_T, X_{t-1}, c) = D(y_t|w_t, z_t, X_{t-1}, c), \quad t = 1, \ldots, T$$  

(2)

and

$$D(w_t|Z_T, X_{t-1}, c) = D(w_t|z_t, X_{t-1}, c), \quad t = 1, \ldots, T$$  

(3)
where $Z_T = (z_T, \ldots, z_1)$. Eq. (2) means that, once current $w_t$, current $z_t$, past $y_t$ and $w_t$, and $c$ are controlled for, $z_t, s \neq t$, has no affect on the distribution of $y_t$. Eq. (3) has a similar interpretation.

In practice, we parameterize the conditional densities: let

$$f_t(\cdot|w_t, z_t, X_{t-1}, c; \theta_0),$$

and

$$g_t(\cdot|z_t, X_{t-1}, c; \gamma_0),$$

be the densities corresponding to (2) and (3), respectively, where $\theta_0$ and $\gamma_0$ are finite dimensional parameters.

By the usual product law for conditional densities, the joint density of $(y_t, w_t)$ given $(Z_T, X_{t-1}, c)$ is

$$p_t(x_t|z_t, X_{t-1}, c; \delta_0) = f_t(y_t|w_t, z_t, X_{t-1}, c; \theta_0)g_t(w_t|z_t, X_{t-1}, c; \gamma_0),$$

where $x_t$, $y_t$, and $w_t$ are placeholders and $\delta_0$ is the vector of all parameters. It follows that the density of $(x_T, x_1)$ given $(Z_T, x_0, c)$ is

$$p(x_T, \ldots, x_1|Z_T, x_0, c; \delta_0) = \prod_{t=1}^{T} p_t(x_t|z_t, X_{t-1}, x_0, c; \delta_0).$$

Because this density depends on $c$, which is unobserved, using (6) to estimate $\theta_0$ (and $\gamma_0$) requires some care. Without $w_t$ in the model, one suggestion has been to treat $c_t$, the random draw on $c$ for cross section unit $i$, as a parameter to estimate for each $i$. (Unfortunately, this has been labelled the “fixed effects” approach.) While estimating the $c_t$ produces consistent estimators of $\theta_0$ in some special cases, including the linear unobserved effects model and the Poisson regression model, both with strictly exogenous explanatory variables, it generally produces inconsistent estimators with fixed $T$ and $N \to \infty$, where $N$ is the size of the cross section.

Generally, to consistently estimate $\delta_0$ we need to integrate $c$ out of the density. This has been the focus of much of the literature on dynamic unobserved effect models. For binary response models without $w_t$ in the model (so that the density in (5) can be set to unity), the recommended solutions have been: (i) For each cross section observation $i$, treat the initial conditional, $y_{i0}$, as nonrandom, and assume that $c$ is independent of $Z_T$. (ii) Specify the distribution of $y_{i0}$ given $(Z_T, c)$, which allows us to obtain $D(y_{i0}|Z_T, c)$; we can then specify the distribution of $c$ given $Z_T$ or, more typically, assume $c$ and $Z_T$ are independent, and then integrate out $c$ to obtain the density for $D(y_{i0}|Z_T)$. This can be used in conditional maximum likelihood estimation (CMLE), with the conditioning on $Z_T$ (Traditionally, this is viewed as “unconditional” MLE because the $z_t$ are treated as nonrandom). (iii) Approximate $D(y_{i0}|Z_T, c)$ and specify a distribution for $c$ (assuming $c$ and $Z_T$ are independent); this was first suggested by Heckman (1981). See (Hsiao, 1986), Section 7.4, for further discussion.

The above approaches have their drawbacks. First, treating $y_{i0}$ as nonrandom means it is independent of $Z_{iT}$ and $c_t$, usually an untenable assumption. The second approach is very complicated and often infeasible because it requires finding a steady state distribution for $y_t$; the inclusion of $z_t$ complicates matters even more. Heckman (1981) approach is the most flexible but, as we will now see, it is more complicated, and more restrictive, than necessary.

My suggestion is to model $D(c|Z_T, x_0)$, and then construct the density of $(x_T, \ldots, x_1)$ given $(Z_T,$
improvement over assuming depend on the initial condition as well as the strictly exogenous variables, and represents a clear

\[\text{in}(9),\text{and}\ T\text{fixed}\]

As a practical matter, we can specify a parametric density

\[h(c|Z_T, x_0; \lambda_0),\]

where \(\lambda_0\) is a vector of nuisance parameters. For example, we might assume normality, where the conditional mean, and possibly the conditional variance, are flexible functions of \((Z_T, x_0)\). Or, we might use the semi-nonparametric approach of Gallant and Nychka (1987) to obtain a flexible density. It is important to see that the models in (4) and (5) place no restrictions on \(h(c|Z_T, x_0; \lambda)\). (This is because \(D(c|Z_T)\) is unrestricted.) Once we have specified \(h(c|Z_T, x_0; \lambda_0)\), we obtain the density of \((x_T, \ldots, x_1)\) given \((Z_T, x_0)\) by integrating out \(c\):

\[
\int_{\mathbb{R}^m} p(x_T, \ldots, x_1|Z_T, x_0, c; \delta_0)h(c|Z_T, x_0; \lambda_0)\nu(dc),
\]

where \(m\) is the dimension of \(c\) and \(\nu(\cdot)\) is a suitable measure. Given (6) and (8), the log-likelihood function for cross section observation \(i\) is

\[
\ell(X_{iT}, Z_{iT}; \delta, \lambda) = \log\left[\int_{\mathbb{R}^m} \prod_{t=1}^{T} p_i(y_{it}, w_{it}|z_{it}, X_{i,t-1}, c; \delta)h(c|Z_T, x_0; \lambda)\nu(dc)\right]
\]

where \(X_{iT} = (x_{iT}, x_{i,t-1}, \ldots, x_{i0})\) and \(Z_{iT} = (z_{iT}, \ldots, z_{i1})\).

We can easily sketch the consistency of the conditional maximum likelihood estimator that uses (9) as the generic log likelihood. Because (8) is the density of \((x_T, \ldots, x_1)\) given \((Z_T, x_0)\), it follows from the Kullback–Leibler information inequality (for example, Manski (1988, Section 5.1)) that for all \((\delta, \lambda) \neq (\delta_0, \lambda_0)\),

\[E[\ell(X_{iT}, Z_{iT}; \delta_0, \lambda_0)|Z_T, x_0] \geq E[\ell(X_{iT}, Z_{iT}; \delta, \lambda)|Z_T, x_0].\]

By the law of iterated expectations,

\[E[\ell(X_{iT}, Z_{iT}; \delta_0, \lambda_0)] \geq E[\ell(X_{iT}, Z_{iT}; \delta, \lambda)], \quad (\delta, \lambda) \neq (\delta_0, \lambda_0)
\]

where these expectations are with respect to the joint distribution of \((X_{iT}, Z_{iT})\). This ensures that the true parameters \((\delta_0, \lambda_0)\) solve the relevant population maximization problem, but they might not be the unique solutions. For identification, we must assume, or show, that the inequality in (10) is strict.

If we have random sampling in the cross section dimension and standard regularity conditions, with fixed \(T\) the CMLE for \(\delta_0\) and \(\lambda_0\) will be consistent and \(\sqrt{N}\)-asymptotically normally distributed. (See Newey and McFadden (1994) for sufficient regularity conditions.)

Without \(w_i\) in the structural model, \(p_i(y_{it}, w_{it}|z_{it}, X_{i,t-1}, c; \delta)\) gets replaced with \(f_i(y_{it}|z_{it}, Y_{i,t-1}, c; \theta)\) in (9), and \(h(c|Z_T, x_0; \lambda)\) becomes \(h(c|Z_T, y_0; \lambda)\). This explicitly allows the distribution of \(c\) to depend on the initial condition as well as the strictly exogenous variables, and represents a clear improvement over assuming \(y_0\) is nonrandom.

It is crucial to see that this analysis does not treat the \(x_{i0}\) as nonrandom. By modeling the density
for \( D(c|Z_T, x_0) \), we are able to obtain the density for \( D(x_T, \ldots, x_1|Z_T, x_0) \), which is enough to apply CMLE. With CMLE there is no presumption that the conditioning variables, \( Z_T \) and \( x_0 \) in this case, are fixed in repeated samples; in fact, with random sampling in the cross section, they must be allowed to be random. When \( x_{t0} \) is treated as fixed, the practical implication is that the density \( h(c|Z_T, x_0; \lambda_0) \) is replaced with the unconditional density, \( h(c; \lambda_0) \), which is the same as assuming \( c \) is independent of \( (Z_T, x_0) \). This distinction seems to be unappreciated, even in the recent literature. For example, Chay and Hyslop (1998, pp. 12–13) and Vella and Verbeek (1999, p. 243) effectively claim that Heckman (1981) approach is the only practical solution unless we assume the initial conditions are nonrandom.

3. Examples

I now give a couple of examples to illustrate the general method. For notational simplicity, I drop the “o” subscript from the true parameters. First, let \( y_t \) be a binary response, and let \( w_t \) also be a binary explanatory variable. Consider the unobserved effects probit model

\[
P(y_t = 1|w_t, z_t, X_{t-1}, c; \theta) = \Phi(\theta_1 w_t + z_t \theta_2 + \theta_3 y_{t-1} + c),
\]

(11)

so that, once \( w_t, z_t, y_{t-1} \), and \( c \) are controlled for, no further lags of \( w_t, z_t, \) or \( y_t \) affect the response probability. If we want to allow feedback from \( y_{t-1} \) to \( w_t \), even after conditioning on \( c \), we need to specify a model for \( w_t \). A dynamic probit model is natural:

\[
P(w_t = 1|z_t, X_{t-1}, c; \gamma) = \Phi(z_t \gamma_1 + \gamma_2 w_{t-1} + \gamma_3 y_{t-1} + \gamma_4 c),
\]

(12)

where the coefficient \( \gamma_4 \) on \( c \) allows the unobserved heterogeneity to affect \( y_t \) and \( w_t \) differently. We can easily modify (12) to account for institutional constraints. For example, if a firm received a job training grant in year \( t-1 \), it is not eligible for a grant in year \( t \). (Then, \( P(w_t = 1|z_t, w_{t-1} = 1) = 0 \).) Given (11) and (12), we have enough to obtain the density in (6). To round out the analysis, we need a density for \( c \) given \( (y_0, w_0, Z_T) \). One possibility is

\[
c|(y_0, w_0, Z_T) \sim \text{Normal}(\lambda_1 + \lambda_2 y_0 + \lambda_3 w_0 + \bar{z}\lambda_4, \lambda_5),
\]

(13)

where \( \bar{z} \) is the average of \( z_t \) over \( t \); we might even interact \( y_0 \) and \( w_0 \), so that the model is saturated in the two binary outcomes (and possibly interact these with \( \bar{z} \)). This leads to a tractable log-likelihood function. Without \( w_t \), we obtain a new way to estimate a dynamic unobserved effects probit model with strictly exogenous \( z_t \).

As a second example, consider a count data model,

\[
E(y_t|w_t, z_t, X_{t-1}, c) = \mu(\theta_1 w_t + z_t \theta_2, \theta_3 y_{t-1}, c),
\]

(14)

where we also assume that the conditional distribution is Poisson. We allow \( \theta_3 y_{t-1} \) to appear as a separate argument because there is no set way of including lagged dependent variables in count models. Suppose that \( w_t \) is roughly continuous, and we assume conditional normality:

\[
w_t|(z_t, X_{t-1}, c) \sim \text{Normal}(z_t \gamma_1 + \gamma_2 w_{t-1} + \gamma_3 y_{t-1} + \gamma_4 c, \gamma_5).
\]

(15)

We can combine (14) and (15) with (13) and use (9) to construct the log-likelihood for observation \( i \).
4. Open issues

This note is theoretical, and, in effect, suggests the possibility of estimating very general, nonlinear, dynamic, unobserved effects panel data models with feedback. In any application, many issues would have to be resolved. First, under what assumptions are the parameters identified? (They should be fairly generally in the examples given in Section 3.) Second, how can we actually compute the conditional MLE estimates, along with asymptotic standard errors? In complicated models, simulation methods — see, for example, Keane (1993) — might be needed. Third, how do we report the results? In many nonlinear models, just having the parameters estimates is insufficient. For example, in the probit example from Section 3, we might want to estimate the average treatment effect, which requires integrating out $c$ in the right way. Finally, how sensitive are the estimates of the important quantities to our specifications of (5) and (7), which, in many cases, are auxiliary assumptions. (In some models, such as linear models with additive effects, and exponential models with multiplicative effects, full distributional assumptions are not needed for consistent estimation.)

References