JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **20**: 39–54 (2005)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jae.770

SIMPLE SOLUTIONS TO THE INITIAL CONDITIONS PROBLEM IN DYNAMIC, NONLINEAR PANEL DATA MODELS WITH UNOBSERVED HETEROGENEITY

JEFFREY M. WOOLDRIDGE*
Department of Economics, Michigan State University, USA

SUMMARY

I study a simple, widely applicable approach to handling the initial conditions problem in dynamic, nonlinear unobserved effects models. Rather than attempting to obtain the joint distribution of all outcomes of the endogenous variables, I propose finding the distribution conditional on the initial value (and the observed history of strictly exogenous explanatory variables). The approach is flexible, and results in simple estimation strategies for at least three leading dynamic, nonlinear models: probit, Tobit and Poisson regression. I treat the general problem of estimating average partial effects, and show that simple estimators exist for important special cases. Copyright © 2005 John Wiley & Sons, Ltd.

1. INTRODUCTION

In dynamic panel data models with unobserved effects, the treatment of the initial observations is an important theoretical and practical problem. Much attention has been devoted to dynamic linear models with an additive unobserved effect, particularly the simple AR(1) model without additional covariates. As is well known, the usual within estimator is inconsistent, and can be badly biased. [See, for example, Hsiao (1986, section 4.2).]

For linear models with an additive unobserved effect, the problems with the within estimator can be solved by using an appropriate transformation—such as differencing—to eliminate the unobserved effects. Then, instrumental variables (IV) can usually be found for implementation in a generalized method of moments (GMM) framework. Anderson and Hsiao (1982) proposed IV estimation on a first-differenced equation, while several authors, including Arellano and Bond (1991), Arellano and Bover (1995), Ahn and Schmidt (1995), improved on the Anderson–Hsiao estimator by using additional moment restrictions in GMM estimation. More recently, Blundell and Bond (1998) and Hahn (1999) have shown that imposing restrictions on the distribution of initial conditions can greatly improve the efficiency of GMM over certain parts of the parameter space.

Solving the initial conditions problem is notably more difficult in nonlinear models. Generally, there are no known transformations that eliminate the unobserved effects and result in usable moment conditions, although special cases have been worked out. Chamberlain (1992) finds moment conditions for dynamic models with a multiplicative effect in the conditional mean, and Wooldridge (1997) considers transformations for a more general class of multiplicative models. Honoré (1993) obtains orthogonality conditions for the unobserved effects Tobit model with a lagged dependent variable. For the unobserved effects logit model with a lagged dependent

* Correspondence to: Professor Jeffrey M. Wooldridge, Department of Economics, Michigan State University, East Lansing, MI 48824-1038, USA. E-mail: wooldri1@msu.edu

Received 12 July 2002
Revised 11 September 2003

variable, Honoré and Kyriazidou (2000) find an objective function that identifies the parameters under certain assumptions on the strictly exogenous covariates.

The strength of semiparametric approaches is that they allow estimation of parameters without specifying a distribution for the unobserved effect. Unfortunately, semiparametric identification hinges on some strong assumptions concerning the strictly exogenous covariates; for example, time dummies are not allowed in the Honoré and Kyriazidou (2000) approach. Honoré and Kyriazidou also reduce the sample to cross-sectional units with no change in any discrete covariates over the last two time periods.

Another practical limitation of the Honoré (1993) and Honoré and Kyriazidou (2000) estimators—and one that often goes unnoticed—is that partial effects on the response probability or conditional mean are not identified. Therefore, the absolute importance of covariates, or the amount of state dependence, cannot be determined from semiparametric approaches.

In this paper I reconsider the initial conditions problem in a parametric framework for nonlinear models. A parametric approach has its usual drawbacks because I specify an auxiliary conditional distribution for the unobserved heterogeneity; misspecification of this distribution generally results in inconsistent parameter estimates. Nevertheless, in some leading cases the approach I take leads to some remarkably simple maximum likelihood estimators. Further, I show that the assumptions are sufficient for uncovering the quantities that are usually of interest in nonlinear applications: partial effects on the mean response, averaged across the population distribution of the unobserved heterogeneity.

Previous research in parametric, dynamic nonlinear models has focused on three different ways of handling initial conditions; these are summarized by Hsiao (1986, section 7.4). The first approach is to treat the initial conditions for each cross-sectional unit as nonrandom variables. Unfortunately, nonrandomness of the initial conditions, $\mathbf{y}_{i0}$, implies that $\mathbf{y}_{i0}$ is independent of unobserved heterogeneity, $\mathbf{c}_i$. Even when we observe the entire history of the process $\{\mathbf{y}_{it}\}$, the assumption of independence between $\mathbf{c}_i$ and $\mathbf{y}_{i0}$ is very strong. For example, suppose we are interested in modelling earnings of college graduates once they leave college, and $\mathbf{y}_{i0}$ is earnings in the first post-school year. That we observe the start of this process is logically distinct from the strong assumption that unobserved 'ability' and 'motivation' are independent of initial earnings.

A better approach is to allow the initial condition to be random, and then to use the joint distribution of *all* outcomes on the response—including that in the initial time period—conditional on unobserved heterogeneity and observed strictly exogenous explanatory variables. The main complication with this approach is specifying the distribution of the initial condition given unobserved heterogeneity. Some authors insist that the distribution of the initial condition represent a steady-state distribution. While the steady-state distribution can be found in special cases—such as the first-order linear model without exogenous variables [see Bhargava and Sargan (1983) and Hsiao (1986, section 4.3)] and in the unobserved probit model without additional conditioning variables [see Hsiao (1986, section 7.4)]—it cannot be done for even modest extensions.

For the dynamic probit model with covariates, Heckman (1981) proposed approximating the conditional distribution of the initial condition. This avoids the practical problem of not being able to find the conditional distribution of the initial value. But, as we will see, it is computationally more difficult than necessary for obtaining both parameter estimates and estimates of averaged effects in nonlinear models.

The approach I suggest in this paper is to model the distribution of the unobserved effect conditional on the initial value and any exogenous explanatory variables. This suggestion has been made before for particular models. For example, Chamberlain (1980) mentions this possibility for the linear AR(1) model without covariates, and Blundell and Smith (1991) study the conditional maximum likelihood estimator of the same model; see also Blundell and Bond (1998). For the binary response model with a lagged dependent variable, Arellano and Carrasco (2003) study a maximum likelihood estimator conditional on the initial condition, where the distribution of the unobserved effect given the initial is taken to be discrete. When specialized to the binary response model, the approach here is more flexible, and computationally much simpler: the response probability can have the probit or logit form, strictly exogenous explanatory variables are easily incorporated along with a lagged dependent variable, and standard random effects software can be used to estimate the parameters and averaged effects.

Specifying a distribution of heterogeneity conditional on the initial condition has several advantages. First, we can choose the auxiliary distribution to be flexible, and view it as an alternative approximation to Heckman's (1981). Second, in several leading cases—probit, ordered probit, Tobit and Poisson regression—an auxiliary distribution can be chosen that leads to straightforward estimation using standard software. Third, partial effects on mean responses, averaged across the distribution of unobservables, are identified and can be estimated without much difficulty. I show how to obtain these partial effects generally in Section 4, and Section 5 covers the probit and Tobit models.

2. EXAMPLES

We introduce three examples in this section to highlight the important issues; we return to these examples in Section 5.

Example 1 (Dynamic Probit Model with Unobserved Effect): For a random draw i from the population and $t = 1, 2, \ldots, T$:

$$\mathrm{P}(y_{it} = 1|y_{i,t-1}, \ldots, y_{i0}, \mathbf{z}_i, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\gamma} + \rho y_{i,t-1} + c_i) \tag{1}$$

This equation contains several assumptions. First, the dynamics are first order, once $\mathbf{z}_{it}$ and c_i are also conditioned on. Second, the unobserved effect is additive inside the standard normal cumulative distribution function, Φ. Third, the $\mathbf{z}_{it}$ satisfy a strict exogeneity assumption: only $\mathbf{z}_{it}$ appears on the right-hand side, even though $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT})$ appears in the conditioning set on the left. Naturally, $\mathbf{z}_{it}$ can contain lags, and even leads, of exogenous variables.

As we will see in Sections 3 and 4, the parameters in equation (1), as well as average partial effects, can be estimated by specifying a density for c_i given $(y_{i0}, \mathbf{z}_i)$. A homoscedastic normal distribution with conditional mean linear in parameters is very convenient, as we will see in Section 5. □

Example 2 (Dynamic Tobit Model with Unobserved Effect): Consider

$$y_{it} = \max[0, \mathbf{z}_{it}\boldsymbol{\gamma} + \mathbf{g}(y_{i,t-1})\boldsymbol{\rho} + c_i + u_{it}] \tag{2}$$

$$u_{it}|y_{i,t-1}, \ldots, y_{i0}, \mathbf{z}_i, c_i \sim \text{Normal}(0, \sigma_u^2) \tag{3}$$

for $t = 1, 2, \ldots, T$. This model applies to corner solution outcomes, where y_{it} is an observed response that equals zero with positive probability but is continuously distributed over strictly positive values. It is not well suited to true data censoring applications, as in that case we would want a lagged value of the latent variable underlying equation (2) to appear. The function $\mathbf{g}(\cdot)$ allows the lagged value of the *observed* response to appear in a variety of ways. For instance, we might have $\mathbf{g}(y_{-1}) = \{1[y_{-1} = 0], 1[y_{-1} > 0]\log(y_{-1})\}$, which allows the effect of lagged y to be different depending on whether the previous response was a corner solution (zero) or strictly positive. In this case, $\boldsymbol{\rho}$ is 2×1.

Honoré (1993) proposes orthogonality conditions that identify the parameters, but partial effects are unidentified. We will show how to obtain $\sqrt{N}$-consistent estimates of the parameters and average partial effects in Section 5. □

Example 3 (Dynamic Unobserved Effects Poisson Model): For each $t = 1, \ldots, T$, y_{it} given $(y_{i,t-1}, \ldots, y_{i0}, \mathbf{z}_i, c_i)$ has a Poisson distribution with mean

$$\mathrm{E}(y_{it}|y_{i,t-1}, \ldots, y_{i0}, \mathbf{z}_i, c_i) = c_i \exp[\mathbf{z}_{it}\boldsymbol{\gamma} + \mathbf{g}(y_{i,t-1})\boldsymbol{\rho}] \tag{4}$$

Again, we allow for the lagged dependent variable to appear in a flexible fashion, perhaps as a set of dummy variables for specific outcomes on $y_{i,t-1}$. The null hypothesis of no state dependence is H_0: $\boldsymbol{\rho} = \mathbf{0}$. Chamberlain (1992) and Wooldridge (1997) have proposed orthogonality conditions based only on equation (4), where no conditional distributional assumptions are needed for y_{it} or c_i. Unfortunately, because the moment conditions have features similar to using first differences in a linear equation, the resulting GMM estimators can be very imprecise (even though the parameters would be generally identified). In Section 5 we show how a particular model for a conditional distribution for c_i leads to a simple maximum likelihood analysis. □

3. GENERAL FRAMEWORK

Let i index a random draw from the cross-section, and let t denote a particular time period. Initially, we assume that we observe $(\mathbf{z}_{it}, \mathbf{y}_{it})$, $t = 1, \ldots, T$, along with $\mathbf{y}_{i0}$. We are interested in the conditional distribution of $\mathbf{y}_{it} \in \mathbb{R}^{\mathrm{G}}$ given $(\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \mathbf{c}_i)$, where $\mathbf{z}_{it}$ is a vector of conditioning variables at time t and $\mathbf{c}_i \in \mathbb{R}^{\mathrm{J}}$ is unobserved heterogeneity. (The dimension of $\mathbf{z}_{it}$ could be increasing with t, but in our examples its dimension is fixed.) We denote the conditional distribution by $\mathrm{D}(\mathbf{y}_{it}|\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \mathbf{c}_i)$. The asymptotic analysis is with the number of time periods, T, fixed, with the cross-section sample size, N, going to infinity.

We make two key assumptions on the conditional distribution of interest. First, we assume that the dynamics are correctly specified. This means that at most one lag of $\mathbf{y}_{it}$ appears in the distribution given outcomes back to the initial time period. Second, $\mathbf{z}_i = \{\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT}\}$ is appropriately strictly exogenous, conditional on $\mathbf{c}_i$. Both of these can be expressed as follows.

Assumption 1: For $t = 1, 2, \ldots, T$:

$$\mathrm{D}(\mathbf{y}_{it}|\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \mathbf{c}_i) = \mathrm{D}(\mathbf{y}_{it}|\mathbf{z}_i, \mathbf{y}_{i,t-1}, \ldots, \mathbf{y}_{i0}, \mathbf{c}_i) \tag{5}$$

We next assume that we have a correctly specified parametric model for the density representing equation (5) which, for lack of a better name, we call the 'structural' density.

Assumption 2: For $t = 1, 2, \ldots, T$, $f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta})$ is a correctly specified density for the conditional distribution on the left-hand side of equation (5), with respect to a σ-finite measure $\nu(d\mathbf{y}_t)$. The parameter space, $\boldsymbol{\Theta}$, is a subset of $\mathbb{R}^{\mathrm{P}}$. Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_{\mathbf{o}} \in \boldsymbol{\Theta}$.

The requirement that we have a density with respect to a σ-finite measure is not restrictive in practice. If $\mathbf{y}_t$ is purely discrete, ν is a counting measure. If $\mathbf{y}_t$ is continuous, ν is a Lebesgue measure. An appropriate σ-finite measure can be found for all of the possible response variables of interest in economics.

Most specific analyses of dynamic, nonlinear unobserved effects models begin with assumptions similar to 1 and 2. Together, Assumptions 1 and 2 imply that the density of $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ given $(\mathbf{y}_{i0} = \mathbf{y}_0, \mathbf{z}_i = \mathbf{z}, \mathbf{c}_i = \mathbf{c})$ is

$$\prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}_{\mathrm{o}}) \tag{6}$$

where we drop the i subscript to indicate dummy arguments of the density. In using equation (6) to estimate $\boldsymbol{\theta}_{\mathrm{o}}$, we must confront the fact that it depends on the unobservables, $\mathbf{c}$. One possibility is to construct the log-likelihood function that treats the N unobserved effects, $\mathbf{c}_i$, as (vectors of) parameters to be estimated. This leads to maximizing the function

$$\sum_{i=1}^{\mathrm{N}} \sum_{t=1}^{\mathrm{T}} \log f_t(\mathbf{y}_{it}|\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \mathbf{c}_i; \boldsymbol{\theta}) \tag{7}$$

over $\boldsymbol{\theta}$ and $(\mathbf{c}_1, \ldots, \mathbf{c}_N)$. While this approach avoids having to restrict the distribution of $\mathbf{c}_i$, it suffers from an incidental parameters problem with fixed T: except in very special cases, the estimator of $\boldsymbol{\theta}_{\mathrm{o}}$ is inconsistent.

The alternative is to 'integrate out' the unobserved effect. As we discussed in the Introduction, there have been several suggestions for doing this. A popular approach is to find the density of $(\mathbf{y}_{i0}, \mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ given $\mathbf{z}_i$. If we specify $f(\mathbf{y}_0|\mathbf{z}, \mathbf{c})$ then

$$f(\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_T|\mathbf{z}, \mathbf{c}) = f(\mathbf{y}_1, \ldots, \mathbf{y}_T|\mathbf{y}_0, \mathbf{z}, \mathbf{c}) \cdot f(\mathbf{y}_0|\mathbf{z}, \mathbf{c}) \tag{8}$$

Next, we specify a density $f(\mathbf{c}|\mathbf{z})$. We can then integrate equation (8) with respect to this density to obtain $f(\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_T|\mathbf{z})$.

Rather than trying to find the density of $(\mathbf{y}_{i0}, \mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ given $\mathbf{z}_i$, my suggestion is to use the density of $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ conditional on $(\mathbf{y}_{i0}, \mathbf{z}_i)$. Because we already have the density of $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ conditional on $(\mathbf{y}_{i0}, \mathbf{z}_i, \mathbf{c}_i)$—given by equation (6)—we need only specify the density of $\mathbf{c}_i$ conditional on $(\mathbf{y}_{i0}, \mathbf{z}_i)$. Because this density is not restricted by the specification in Assumption 2, we can choose it for convenience, or flexibility or, hopefully, both. As in Chamberlain's (1980) analysis of unobserved effects probit models with strictly exogenous explanatory variables, we view the device of specifying $f(\mathbf{c}|\mathbf{y}_0, \mathbf{z})$ as a way of obtaining relatively simple estimates of $\boldsymbol{\theta}_{\mathrm{o}}$. Specifying a model for $f(\mathbf{c}|\mathbf{y}_0, \mathbf{z})$ seems no worse than having to specify models for $f(\mathbf{y}_0|\mathbf{z}, \mathbf{c})$, which must themselves be viewed as approximations, except in special cases where steady-state distributions can be derived.

Assumption 3: $h(\mathbf{c}|\mathbf{y}_0, \mathbf{z}; \boldsymbol{\delta})$ is a correctly specified model for the density of $\mathrm{D}(\mathbf{c}_i|\mathbf{y}_{i0}, \mathbf{z}_i)$ with respect to a $\boldsymbol{\sigma}$-finite measure $\eta(d\mathbf{c})$. Let $\boldsymbol{\Delta} \subset \mathbb{R}^{\mathrm{M}}$ be the parameter space and let $\boldsymbol{\delta}_{\mathrm{o}}$ denote the true value of $\boldsymbol{\delta}$.

Assumption 3 is much more controversial than Assumptions 1 and 2. Ideally, we would not have to specify anything about the relationship between $\mathbf{c}_i$ and $(\mathbf{y}_{i0}, \mathbf{z})$, whereas Assumption 3 assumes we have a complete conditional density correctly specified. In some specific cases—linear models, logit models, Tobit models and exponential regression models—consistent estimators of $\boldsymbol{\theta}_o$ are available without Assumption 3. But these estimators are complicated and need not have good precision.

Under Assumptions 1, 2 and 3, the density of $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ given $(\mathbf{y}_{i0} = \mathbf{y}_0, \mathbf{z}_i = \mathbf{z})$ is

$$\int_{\mathbb{R}^J} \left(\prod_{t=1}^{T} f_t(\mathbf{y}_t | \mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}_o) \right) h(\mathbf{c} | \mathbf{y}_0, \mathbf{z}; \boldsymbol{\delta}_o) \eta(d\mathbf{c}) \tag{9}$$

which leads to the log-likelihood function conditional on $(\mathbf{y}_{i0}, \mathbf{z}_i)$ for each observation i:

$$\ell_i(\boldsymbol{\theta}, \boldsymbol{\delta}) = \log \left[\int_{\mathbb{R}^J} \left(\prod_{t=1}^{T} f_t(\mathbf{y}_{it} | \mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \mathbf{c}; \boldsymbol{\theta}) \right) h(\mathbf{c} | \mathbf{y}_{i0}, \mathbf{z}_i; \boldsymbol{\delta}) \eta(d\mathbf{c}) \right] \tag{10}$$

To estimate $\boldsymbol{\theta}_o$ and $\boldsymbol{\delta}_o$, we sum the log-likelihoods in equation (10) across $i = 1, \ldots, N$ and maximize with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$. The resulting conditional MLE is $\sqrt{N}$-consistent and asymptotically normal under standard regularity conditions. In dynamic unobserved effects models, the log-likelihoods are typically very smooth functions, and we usually assume that the needed moments exist and are finite. From a practical perspective, identification is the key issue. Generally, if $\mathrm{D}(\mathbf{c}_i | \mathbf{y}_{i0}, \mathbf{z}_i)$ is allowed to depend on all elements of $\mathbf{z}_i$ then the way in which any time-constant exogenous variables can appear in the structural density is restricted. To increase explanatory power, we can include time-constant explanatory variables in $\mathbf{z}_{it}$, but we will not be able to separately identify the partial effect of the time-constant variable from its partial correlation with $\mathbf{c}_i$.

The log-likelihood in equation (10) assumes that we observe data on all cross-sectional units in all time periods. Nevertheless, for unbalanced panels under certain sample selection mechanisms, we can use the same conditional log-likelihood for the subset of observations constituting a balanced panel. Let s_i be a selection indicator: $s_i = 1$ if we observe data in all time periods (including $\mathbf{y}_{i0}$), and zero otherwise. Then, if $(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ and s_i are independent conditional on $(\mathbf{y}_{i0}, \mathbf{z}_i)$, the MLE using the balanced panel will be consistent, and the usual asymptotic standard errors and test statistics are asymptotically valid. Consistency follows from the general argument in Wooldridge (2002, section 17.2.2).

When attrition is an issue, obtaining the density conditional on $(\mathbf{y}_{i0}, \mathbf{z}_i)$ has some advantages over the more traditional approach, where the density would be conditional only on $\mathbf{z}_i$. In particular, the current approach allows attrition to depend on the initial condition, $\mathbf{y}_{i0}$, in an arbitrary way. For example, if y_{i0} is annual hours worked, an MLE analysis based on equation (10) allows attrition probabilities to differ across initial hours worked. In the traditional approach, one would have to explicitly model attrition as a function of $\mathbf{y}_{i0}$ and figure out the appropriate Heckit-type analysis.

Of course, reducing the data set to a balanced panel can discard useful information. But available semiparametric methods have the same feature. For example, the objective function in Honoré and Kyriazidou (2000) includes differences in the strictly exogenous covariates for $T = 3$. Any observation where $\Delta \mathbf{z}_{it}$ is missing for $t = 2$ or 3 cannot contribute to the analysis.

4. ESTIMATING AVERAGE PARTIAL EFFECTS

In nonlinear models we often need to go beyond estimation of the parameters, $\boldsymbol{\theta}_{\mathrm{o}}$, and obtain estimated partial effects. Typically, we would like the effect on a mean response after averaging the unobserved heterogeneity across the population. I now show how to construct consistent, $\sqrt{N}$-asymptotically normal estimators of these average partial effects (APEs).

Let $q(\mathbf{y}_t)$ be a scalar function of $\mathbf{y}_t$ whose conditional mean we are interested in at time t. The leading case is $q(y_t) = y_t$ when y_t is a scalar, but $q(\cdot)$ could be an indicator function if we are interested in probabilities. Generally, we are interested in

$$\begin{aligned} m(\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}_{\mathrm{o}}) &= \mathrm{E}[q(\mathbf{y}_{it})|\mathbf{z}_{it} = \mathbf{z}_t, \mathbf{y}_{i,t-1} = \mathbf{y}_{t-1}, \mathbf{c}_i = \mathbf{c})] \\ &= \int_{\mathrm{IR}^{\mathrm{G}}} q(\mathbf{y}_t) f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}_{\mathrm{o}}) \nu(\mathrm{d}\mathbf{y}_t) \end{aligned} \tag{11}$$

where $\mathbf{z}_t$, $\mathbf{y}_{t-1}$ and $\mathbf{c}$ are values that we must choose. Unfortunately, since the unobserved heterogeneity rarely, if ever, has natural units of measurement, it is unclear which values we should plug in for $\mathbf{c}$. Instead, we can hope to estimate the partial effects averaged across the distribution of $\mathbf{c}_i$. That is, we estimate

$$\boldsymbol{\mu}(\mathbf{z}_t, \mathbf{y}_{t-1}) = \mathrm{E}[m(\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}_i; \boldsymbol{\theta}_{\mathrm{o}})] \tag{12}$$

where the expectation is with respect to $\mathbf{c}_i$. (For emphasis, variables with an i subscript are random variables in the expectations; others are fixed values.) Under Assumptions 1, 2 and 3, we do not have a parametric model for the unconditional distribution of $\mathbf{c}_i$, and so it may seem that we need to add additional assumptions to estimate equation (12). Instead, we can obtain a consistent estimator of equation (12) using iterated expectations:

$$\begin{aligned} \mathrm{E}[m(\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}_i; \boldsymbol{\theta}_{\mathrm{o}})] &= \mathrm{E}\{\mathrm{E}[m(\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}_i; \boldsymbol{\theta}_{\mathrm{o}})|\mathbf{y}_{i0}, \mathbf{z}_i]\} \\ &= \mathrm{E}\left[\left(\int_{\mathrm{IR}^{\mathrm{G}}} q(\mathbf{y}_t) f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}_{\mathrm{o}}) \nu(\mathrm{d}\mathbf{y}_t)\right) h(\mathbf{c}|\mathbf{y}_{i0}, \mathbf{z}_i; \boldsymbol{\delta}_{\mathrm{o}}) \eta(\mathrm{d}\mathbf{c})\right] \end{aligned} \tag{13}$$

where the outside, unconditional expectation is with respect to the distribution of $(\mathbf{y}_{i0}, \mathbf{z}_i)$. While equation (13) is generally complicated, it simplifies considerably in some leading cases, as we will see in Section 5. In effect, we first compute the expectation of $q(\mathbf{y}_{it})$ conditional on $(\mathbf{z}_{it}, \mathbf{y}_{i,t-1}, \mathbf{c}_i)$, which is possible because we have specified the density $f_t(\mathbf{y}_t|\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}; \boldsymbol{\theta}_{\mathrm{o}})$. Then, we (hopefully) integrate $m(\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c})$ against $h(\mathbf{c}|\mathbf{y}_{i0}, \mathbf{z}_i; \boldsymbol{\delta}_{\mathrm{o}})$ to obtain $\mathrm{E}[m(\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}_i; \boldsymbol{\theta}_{\mathrm{o}})|\mathbf{y}_{i0}, \mathbf{z}_i]$.

One point worth emphasizing about equation (13) is that $\boldsymbol{\delta}_{\mathrm{o}}$ appears explicitly. In other words, while $\boldsymbol{\delta}_{\mathrm{o}}$ may be properly viewed as a nuisance parameter for estimating $\boldsymbol{\theta}_{\mathrm{o}}$, $\boldsymbol{\delta}_{\mathrm{o}}$ is not a nuisance parameter for estimating APEs. Because the semiparametric literature treats $h(\mathbf{c}|\mathbf{y}_0, \mathbf{z})$ as a nuisance function, there seems little hope that semiparametric approaches will deliver consistent estimates of APEs in dynamic, unobserved effects panel data models.

Given equation (13), a consistent estimator of $\boldsymbol{\mu}(\mathbf{z}_t, \mathbf{y}_{t-1})$ follows immediately:

$$\hat{\boldsymbol{\mu}}(\mathbf{z}_t, \mathbf{y}_{t-1}) \equiv N^{-1} \sum_{i=1}^{N} r(\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{y}_{i0}, \mathbf{z}_i; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\delta}}) \tag{14}$$

where $r(\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{y}_{i0}, \mathbf{z}_i; \boldsymbol{\theta}_o, \boldsymbol{\delta}_o) \equiv \mathrm{E}[m(\mathbf{z}_t, \mathbf{y}_{t-1}, \mathbf{c}_i; \boldsymbol{\theta}_o)|\mathbf{y}_{i0}, \mathbf{z}_i]$. Under weak assumptions, $\hat{\boldsymbol{\mu}}(\mathbf{z}_t, \mathbf{y}_{t-1})$ is a $\sqrt{N}$-asymptotically normal estimator of $\boldsymbol{\mu}(\mathbf{z}_t, \mathbf{y}_{t-1})$, whose asymptotic variance can be obtained using the delta method.

5. THE EXAMPLES REVISITED

We reconsider the examples from Section 2, showing how we can apply the results from Sections 3 and 4. Our focus is on choices of the density $h(\mathbf{c}|\mathbf{y}_0, \mathbf{z}; \boldsymbol{\delta})$ that lead to computational simplicity. For notational convenience, we drop the 'o' subscript on the true values of the parameters.

5.1. Dynamic Probit and Ordered Probit Models

In addition to equation (1), assume that

$$c_i|y_{i0}, \mathbf{z}_i \sim \mathrm{Normal}(\alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2, \sigma_a^2) \tag{15}$$

where $\mathbf{z}_i$ is the row vector of all (nonredundant) explanatory variables in all time periods. If $\mathbf{z}_{it}$ contains a full set of time period dummy variables these elements would be dropped from $\mathbf{z}_i$. The presence of $\mathbf{z}_i$ in equation (15) means that we cannot identify the coefficients on time-constant covariates in $\mathbf{z}_{it}$, although time-constant covariates can be included in $\mathbf{z}_i$ in equation (15).

Given equation (1), we can write

$$\begin{aligned} f(y_1, y_2, \ldots, y_T|y_0, \mathbf{z}, c; \boldsymbol{\beta}) = \prod_{t=1}^{T} \{ & \Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + c)^{y_t} \\ & \times [1 - \Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + c)]^{1-y_t}\} \end{aligned} \tag{16}$$

where $\boldsymbol{\beta} = (\boldsymbol{\gamma}', \rho)'$. When we integrate equation (16) with respect to the normal distribution in equation (15), we obtain the density of $\mathrm{D}(y_{i1}, \ldots, y_{iT}|y_{i0}, \mathbf{z}_i)$.

It turns out that we can specify the density in such a way that standard random effects probit software can be used for estimation. If we write

$$c_i = \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i \tag{17}$$

where $a_i|(y_{i0}, \mathbf{z}_i) \sim \mathrm{Normal}(0, \sigma_a^2)$, then y_{it} given $(y_{i,t-1}, \ldots, y_{i0}, \mathbf{z}_i, a_i)$ follows a probit model with response probability

$$\Phi(\mathbf{z}_{it}\boldsymbol{\gamma} + \rho y_{i,t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i) \tag{18}$$

This is easy to derive by writing the latent variable version of the model as $y_{it}^* = \mathbf{z}_{it}\boldsymbol{\gamma} + \rho y_{i,t-1} + c_i + u_{it}$ and plugging in for c_i from equation (17):

$$y_{it}^* = \mathbf{z}_{it}\boldsymbol{\gamma} + \rho y_{i,t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i + u_{it} \tag{19}$$

where $u_{it}|(\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) \sim \text{Normal}(0, 1)$; equation (18) follows. Thus, the density of $(y_{i1}, \ldots, y_{iT})$ given $(y_{i0} = y_0, \mathbf{z}_i = \mathbf{z}, a_i = a)$ is

$$\prod_{t=1}^{T} \{\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_0 + \mathbf{z}\boldsymbol{\alpha}_2 + a)^{y_t} \\ \times [1 - \Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_0 + \mathbf{z}\boldsymbol{\alpha}_2 + a)]^{1-y_t}\} \tag{20}$$

and integrating equation (20) against the Normal $(0, \sigma_{\text{a}}^2)$ density gives the density of $(y_{i1}, \ldots, y_{iT})$ given $(y_{i0} = y_0, \mathbf{z}_i = \mathbf{z})$:

$$\int_{\mathbb{R}} \left(\prod_{t=1}^{T} \{\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_0 + \mathbf{z}\boldsymbol{\alpha}_2 + a)^{y_t}\right) \\ \times [1 - \Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_0 + \mathbf{z}\boldsymbol{\alpha}_2 + a)]^{1-y_t}\}(1/\sigma_{\text{a}})\phi(a/\sigma_{\text{a}})\text{d}a \tag{21}$$

Interestingly, the likelihood in equation (21) has exactly the same structure as in the standard random effects probit model, except that the explanatory variables at time period t are

$$\mathbf{x}_{it} \equiv (1, \mathbf{z}_{it}, y_{i,t-1}, y_{i0}, \mathbf{z}_i) \tag{22}$$

Importantly, we are not saying that a_i is independent of $y_{i,t-1}$, which is impossible. (Dependence between a_i and $y_{i,t-1}$ means that a pooled probit analysis of y_{it} on $\mathbf{x}_{it}$ is inconsistent for the parameters and the APEs.) Further, the density in equation (21) is not the joint density of $(y_{i1}, \ldots, y_{iT})$ given $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, as happens in the case with strictly exogenous $\mathbf{x}_{it}$. Nevertheless, the way random effects probit works is by forming the products of the densities of y_{it} given $(\mathbf{x}_{it}, a_i)$, and then integrating out using the unconditional density of a_i, and this is what equation (21) calls for. So we add y_{i0} and $\mathbf{z}_i$ as additional explanatory variables in each time period and use standard random effects probit software to estimate $\boldsymbol{\gamma}, \rho, \alpha_0, \alpha_1, \alpha_2$ and σ_{a}^2.

Under the assumptions made, we can easily obtain estimated partial effects at interesting values of the explanatory variables. The average partial effects are based on

$$\text{E}[\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + c_i)] \tag{23}$$

where the expectation is with respect to the distribution of c_i. The general formula in equation (14) turns out to be easy to obtain. Again, replace c_i with $c_i = \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i$, so that expression (23) is

$$\text{E}[\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i)] \tag{24}$$

where the expectation is over the distribution of $(y_{i0}, \mathbf{z}_i, a_i)$; $\mathbf{z}_t$ and y_{t-1} are fixed values here. Now, as in Section 4, we use iterated expectations:

$$\text{E}[\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i)] \\ = \text{E}\{\text{E}[\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i)|y_{i0}, \mathbf{z}_i]\} \tag{25}$$

The conditional expectation inside equation (25) is easily shown to be

$$\Phi(\mathbf{z}_t\boldsymbol{\gamma}_{\text{a}} + \rho_{\text{a}} y_{t-1} + \alpha_{\text{a}0} + \alpha_{\text{a}1} y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_{\text{a}2}) \tag{26}$$

where the 'a' subscript denotes the original parameter multiplied by $(1+\sigma_{\text{a}}^2)^{-1/2}$. Now, we want to estimate the expected value of expression (26) with respect to the distribution of $(y_{i0}, \mathbf{z}_i)$. A consistent estimator is

$$N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{z}_t\hat{\boldsymbol{\gamma}}_{\text{a}}+\hat{\rho}_{\text{a}}y_{t-1}+\hat{\alpha}_{\text{a}0}+\hat{\alpha}_{\text{a}1}y_{i0}+\mathbf{z}_i\hat{\boldsymbol{\alpha}}_{\text{a}2}) \tag{27}$$

where the 'a' subscript now denotes multiplication by $(1+\hat{\sigma}_{\text{a}}^2)^{-1/2}$, and $\hat{\boldsymbol{\gamma}}, \hat{\rho}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\boldsymbol{\alpha}}_2$ and $\hat{\sigma}_{\text{a}}^2$ are the MLEs. We can compute changes or derivatives of expression (27) with respect to $\mathbf{z}_t$ or y_{t-1} to obtain APEs. Thus, we can determine the economic importance of any state dependence.

Allowing for a more flexible conditional mean in equation (15) is straightforward, provided the mean is linear in parameters. For example, including interactions between y_{i0} and $\mathbf{z}_i$ is simple, and would be warranted if we included interactions between the elements of $\mathbf{z}_{it}$ and $y_{i,t-1}$ in the structural model. Allowing for heteroscedasticity in $\text{Var}(c_i|y_{i0},\mathbf{z}_i)$ is more complicated and would probably require special programming. Still, specifying, say, $\text{Var}(c_i|y_{i0},\mathbf{z}_i)=\sigma_{\text{a}0}^2\exp(\gamma_1 y_{i0}+\mathbf{z}_i\boldsymbol{\gamma}_2)$ leads to a tractable log-likelihood function: simply replace σ_{a} in equation (21) with $\sigma_{\text{a}}[\exp(\gamma_1 y_{i0}+\mathbf{z}_i\boldsymbol{\gamma}_2)]^{1/2}$. With $a_i|y_{i0},\mathbf{z}_i \sim \text{Normal}(0,\sigma_{\text{a}0}^2\exp(\gamma_1 y_{i0}+\mathbf{z}_i\boldsymbol{\gamma}_2))$, the conditional expectation in equation (25) is still easy to obtain: $\text{E}[\Phi(\mathbf{z}_t\boldsymbol{\gamma}+\rho y_{t-1}+\alpha_0+\alpha_1 y_{i0}+\mathbf{z}_i\boldsymbol{\alpha}_2+a_i)|y_{i0},\mathbf{z}_i]=\Phi\{(\mathbf{z}_t\boldsymbol{\gamma}+\rho y_{t-1}+\alpha_0+\alpha_1 y_{i0}+\mathbf{z}_i\boldsymbol{\alpha}_2)/(1+\sigma_{\text{a}}^2[\exp(\gamma_1 y_{i0}+\mathbf{z}_i\boldsymbol{\gamma}_2)]^{1/2})\}$—see Wooldridge [2002, problem 15.18(c)] for the static case—and so APEs would be readily computable by averaging across i.

Certain specification tests are easy to compute. For example, after estimating the basic model, terms such as $(\hat{\alpha}_1 y_{i0}+\mathbf{z}_i\hat{\boldsymbol{\alpha}}_2)^2$ and $(\hat{\alpha}_1 y_{i0}+\mathbf{z}_i\hat{\boldsymbol{\alpha}}_2)^3$ could be added and their joint significance tested using a standard likelihood ratio test. Obtaining score tests for exponential heteroscedasticity in $\text{Var}(c_i|y_{i0},\mathbf{z}_i)$ or for nonnormality in $\text{D}(c_i|y_{i0},\mathbf{z}_i)$ are good topics for future research.

The binary probit model extends in a straightforward way to a dynamic ordered probit model. If y_{it} takes on values in $\{0,1,\ldots,J\}$ then we can specify an ordered probit model with J lagged indicators, $1[y_{i,t-1}=j]$, $j=1,\ldots,J$, and strictly exogenous explanatory variables, $\mathbf{z}_{it}$. The underlying latent variable model would be $y_{it}^*=\mathbf{z}_{it}\boldsymbol{\gamma}+\mathbf{r}_{i,t-1}\boldsymbol{\rho}+c_i+e_{it}$, where $\mathbf{r}_{i,t-1}$ is the vector of J indicators, and e_{it} has a conditional standard normal distribution. The observed value, y_{it}, is determined by y_{it}^* falling into a particular interval, where the cut-points must be estimated. If we specify $\text{D}(c_i|y_{i0},\mathbf{z}_i)$ as having a homoscedastic normal distribution, standard random effects ordered probit software can be used. Probably we would allow $h(c|y_0,\mathbf{z};\boldsymbol{\delta})$ to depend on a full set of indicators $1[y_{i0}=j]$, $j=1,\ldots,J$.

Certainly there are some criticisms that one can make about the conditional MLE approach for dynamic probit models. First, suppose that there are no covariates. Then, unless $\alpha_1=0$, equation (15) implies that c_i has a mixture-of-normals distribution, rather than a normal distribution, as would be a standard assumption. But c_i given y_{i0} has some distribution, and it is unclear why an unconditional normal distribution for c_i is *a priori* better than a conditional normal distribution. For cross-sectional binary response models, Geweke and Keane (1999) find that, empirically, mixture-of-normals probit models fit significantly better than the standard probit model. Granted, the mixing probability here is tied to y_0, and the variance is assumed to be constant. But often is econometrics we assume that unobserved heterogeneity has a conditional normal distribution rather than an unconditional normal distribution.

Related to the previous criticism is that, in models without covariates, equation (15) implies a distribution $\mathrm{D}(y_{i0}|c_i)$ different from the steady-state distribution. This is not ideal—in the linear model, one can allow for a non-steady-state distribution while including the steady-state distribution as a special case—but it is only relevant in models without covariates. Plus, even if there are no covariates, it is not clear why imposing a steady-state distribution is better than that implied by equation (15). Dynamic panel data models are really about modelling the conditional distributions in Assumption 1. One can take issue with any set of auxiliary assumptions.

Another criticism is that if $\rho = 0$ then, because c_i given $\mathbf{z}_i$ cannot be normally distributed unless $\alpha_1 = 0$, the model is not compatible with Chamberlain's (1980) static random effects probit model. That the model here does not encompass Chamberlain's is true, but it is unclear why normality of c_i given $\mathbf{z}_i$ is necessarily a better assumption than normality of c_i given $(y_{i0}, \mathbf{z}_i)$. Both are only approximations to the truth and, when estimating a dynamic model, it is much more convenient to use equation (15). Plus, Chamberlain's static model does not allow estimation of either $\boldsymbol{\rho}$ or the amount of state dependence, as measured by the average partial effect.

Assumption (15) is also subject to the same criticism as Chamberlain's (1980) random effects probit model with strictly exogenous covariates. Namely, if we want the same model to hold for any number of time periods T, the normality assumption in equation (15) imposes distributional restrictions on the $\mathbf{z}_{it}$. For example, suppose $\alpha_1 = 0$. Then, for equation (15) to hold for T and $T-1$, $\mathbf{z}_{it}\boldsymbol{\alpha}_{2T}$ given $(\mathbf{z}_{i1}, \ldots, \mathbf{z}_{i,T-1})$ would have to have a normal distribution. While theoretically this is a valid criticism, it is hardly unique to this setting. For example, every time an explanatory variable is added in a cross-sectional probit analysis, the probit model can no longer hold unless the new variable is normally distributed. Yet researchers regularly use probit models on different sets of explanatory variables.

5.2. Dynamic Tobit Models

For the Tobit model the density in Assumption 2 is

$$\begin{aligned} f_t(y_t|\mathbf{z}_t, y_{t-1}, c, \boldsymbol{\theta}) &= 1 - \Phi[(\mathbf{z}_t\boldsymbol{\gamma} + \mathbf{g}(y_{t-1})\boldsymbol{\rho} + c)/\sigma_{\mathrm{u}}], \quad y_t = 0 \\ &= (1/\sigma_{\mathrm{u}})\phi[(y_t - \mathbf{z}_t\boldsymbol{\gamma} - \mathbf{g}(y_{t-1})\boldsymbol{\rho} - c)/\sigma_{\mathrm{u}}], \quad y_t > 0 \end{aligned}$$

To implement the conditional MLE, we need to specify a density in Assumption 3. Again, it is convenient for this to be normal, as in equation (15). For the Tobit case, we might replace y_{i0} with a more general vector of functions, $\mathbf{r}_{i0} \equiv \mathbf{r}(y_{i0})$, which allows c_i to have a fairly flexible conditional mean. Interactions between elements of $\mathbf{r}_{i0}$ and $\mathbf{z}_i$ may be warranted. We can use an argument very similar to the probit case to show that the log-likelihood has a form that can be maximized by standard random effects Tobit software, where the explanatory variables at time t are $\mathbf{x}_{it} \equiv (\mathbf{z}_{it}, \mathbf{g}_{i,t-1}, \mathbf{r}_{i0}, \mathbf{z}_i)$ and $\mathbf{g}_{i,t-1} \equiv \mathbf{g}(y_{i,t-1})$. In particular, the latent variable model can be written as $y_{it}^* = \mathbf{z}_{it}\boldsymbol{\gamma} + \mathbf{g}_{i,t-1}\boldsymbol{\rho} + c_i + u_{it} = \mathbf{z}_{it}\boldsymbol{\gamma} + \mathbf{g}_{i,t-1}\boldsymbol{\rho} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + u_{it}$, where u_{it} given $(\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i)$ has a Normal $(0, \sigma_{\mathrm{u}}^2)$ distribution. Again, we estimate σ_{a}^2 rather than σ_{c}^2, but σ_{a}^2 is exactly what appears in the average partial effects.

Denote $\mathrm{E}(y_{it}|\mathbf{w}_{it} = \mathbf{w}_t, c_i = c)$ as

$$m(\mathbf{w}_t\boldsymbol{\beta} + c, \sigma_{\mathrm{u}}^2) = \Phi[(\mathbf{w}_t\boldsymbol{\beta} + c)/\sigma_{\mathrm{u}}](\mathbf{w}_t\boldsymbol{\beta} + c) + \sigma_{\mathrm{u}}\phi[(\mathbf{w}_t\boldsymbol{\beta} + c)/\sigma_{\mathrm{u}}] \tag{28}$$

where $\mathbf{w}_t = (\mathbf{z}_t, \mathbf{g}_{t-1})$. As in the probit case, for estimating the APEs it is useful to substitute for c_i:

$$\begin{aligned}\mathrm{E}[m(\mathbf{w}_t\boldsymbol{\beta} + c_i, \sigma_{\mathrm{u}}^2)] &= \mathrm{E}[m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i, \sigma_{\mathrm{u}}^2)] \\ &= \mathrm{E}\{\mathrm{E}[m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i, \sigma_{\mathrm{u}}^2)|\mathbf{r}_{i0}, \mathbf{z}_i]\} \end{aligned} \tag{29}$$

where the first expectation is with respect to the distribution of c_i and the second expectation is with respect to the distribution of $(\mathbf{y}_{i0}, \mathbf{z}_i, a_i)$. The second equality follows from iterated expectations. Since a_i and $(\mathbf{r}_{i0}, \mathbf{z}_i)$ are independent, and $a_i \sim \mathrm{Normal}(0, \sigma_{\mathrm{a}}^2)$, the conditional expectation in equation (29) is obtained by integrating $m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i, \sigma_{\mathrm{u}}^2)$ over a_i with respect to the Normal $(0, \sigma_{\mathrm{a}}^2)$ distribution. Since $m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i, \sigma_{\mathrm{u}}^2)$ is obtained by integrating $\max(0, \mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i + u_{it})$ with respect to u_{it} over the Normal $(0, \sigma_{\mathrm{u}}^2)$ distribution, it is easily seen that the conditional expectation in equation (29) is

$$m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2, \boldsymbol{\sigma}_{\mathrm{a}}^2 + \boldsymbol{\sigma}_{\mathrm{u}}^2) \tag{30}$$

A consistent estimator of the expected value of expression (30) is simply

$$N^{-1}\sum_{i=1}^{N} m(\mathbf{w}_t\hat{\boldsymbol{\beta}} + \hat{\alpha}_0 + \mathbf{r}_{i0}\hat{\boldsymbol{\alpha}}_1 + \mathbf{z}_i\hat{\boldsymbol{\alpha}}_2, \hat{\sigma}_{\mathrm{a}}^2 + \hat{\sigma}_{\mathrm{u}}^2) \tag{31}$$

Other corner solution responses can be handled similarly. For example, suppose y_{it} is a fractional variable that can take on the values zero and one with positive probability. Then we can define y_{it} as a doubly-censored version of the latent variable y_{it}^* introduced earlier. Standard software that estimates two-limit random effects Tobit models is readily applied.

5.3. Dynamic Poisson Model

As in Section 2, we assume that y_{it} given $(y_{i,t-1}, \ldots, y_{i0}, \mathbf{z}_i, c_i)$ has a Poisson distribution with mean given in equation (4). For Assumption 3, write

$$c_i = a_i \exp(\boldsymbol{\alpha}_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2) \tag{32}$$

where $\mathbf{r}_{i0}$ is a vector of functions of y_{i0}. Assume that a_i is independent of $(y_{i0}, \mathbf{z}_i)$ and $a_i \sim \mathrm{Gamma}(\eta, \eta)$, which is analogous to Hausman *et al.* (1984). Then, for each t, $y_{it}|(y_{i,t-1}, \ldots, y_{i0}, \mathbf{z}_i, a_i)$ has a Poisson distribution with mean

$$a_i \exp(\mathbf{z}_{it}\boldsymbol{\delta} + \mathbf{g}_{i,t-1}\boldsymbol{\rho} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2) \tag{33}$$

where $\mathbf{r}_{i0}$ denotes a vector function of y_{i0}. Call the mean in expression (33) $a_i m_{it}$. Then the density of $(y_{i1}, \ldots, y_{iT})$ given $(\mathbf{z}_i, y_{i0}, a_i)$ is obtained, as usual, by the product rule:

$$\prod_{t=1}^{T} \exp(-a_i m_{it})(a_i m_{it})^{y_t}/y_t! = \left(\prod_{t=1}^{T} m_{it}^{y_t}/y_t!\right) \exp\left(-a_i \sum_{t=1}^{T} m_{it}\right) a_i^n \tag{34}$$

where $n = y_1 + \cdots + y_T$. When we integrate out a_i with respect to the Gamma (η, η) density, we obtain a density that has the usual random effects Poisson form with Gamma (η, η) heterogeneity, as in Hausman *et al.* [1984, equation (2.3)]. The difference is that the explanatory variables are $(\mathbf{z}_{it}, \mathbf{g}_{i,t-1}, \mathbf{r}_{i0}, \mathbf{z}_i)$. This makes estimation especially convenient in software packages that estimate random effects Poisson models with Gamma heterogeneity. The Chamberlain (1992) and Wooldridge (1997) moment estimators are compatible with this MLE analysis in the sense that the moment estimators only use the conditional mean assumption (4).

6. EMPIRICAL APPLICATION: THE PERSISTENCE OF UNION MEMBERSHIP

Vella and Verbeek (1998) (hereafter, VV) use panel data on working men to estimate the union wage differential, accounting for unobserved heterogeneity. I use their data to estimate a simple model of union membership dynamics. Most of the explanatory variables in VV's data set are constant over time. One variable that does change over time is marital status ($marr_{it}$). A simple dynamic model of union membership is

$$\begin{aligned} \mathrm{P}(union_{it} &= 1|union_{i,t-1}, \ldots, union_{i0}, marr_{i1}, \ldots, marr_{it}, c_i) \\ &= \Phi(\eta_t + \gamma_1 marr_{it} + \rho_1 union_{i,t-1} + c_i), \quad t = 1, \ldots, T \end{aligned} \tag{35}$$

where $t = 1$ corresponds to 1981 and $t = T$ corresponds to 1987. The initial time period is 1980. The unobserved effect, c_i, is assumed to satisfy assumption (15), where $\mathbf{z}_i$ is the $1 \times T$ vector of marital status indicators and $y_{i0} = union_{i0}$. The η_t are unrestricted year intercepts.

Column (1) in Table I contains the maximum likelihood estimates. These were obtained simply by using the Stata®7.0 'xtprobit' command, where a full set of time dummies (not shown), current marital status, lagged union status, union membership status in 1980 ($union_0$) and the marital status dummy variables for 1981 through 1987 ($marr_1$ through $marr_7$) are included as explanatory variables. Asymptotic standard errors are given in parentheses.

Even after controlling for the unobserved effect using the model in Section 5.1, the coefficient on the lagged union status variable is very statistically significant. The initial value of union status is also very important, and implies that there is substantial correlation between the unobserved heterogeneity and the initial condition. In fact, the coefficient on $union_0$ (1.514) is much larger than the coefficient on $union_{t-1}$ (0.875).

Getting married is estimated to have a marginally significant effect on belonging to a union, with a t statistic of about 1.51. The variables $marr_1, \ldots, marr_7$ are included to allow for correlation between c_i and marital status in all time periods. There is no clear pattern to the coefficients, and only $marr_7$ is statistically different from zero at the 5% level.

In order to explicitly control for some observed heterogeneity, column (2) of Table I includes the time-constant variables *educ* and *black*. While we cannot necessarily identify the causal effects of education and race on union membership, we can include them in the model for unobserved heterogeneity in equation (15), which means we just add them as explanatory variables. The coefficient on *educ* is statistically insignificant, while blacks are significantly more likely to belong to a union. Interestingly, even after *educ* and *black* are included, there is much unobserved heterogeneity that cannot be explained by $union_0$, $marr_1, \ldots, marr_7$, *educ* and *black*: $\hat{\sigma}_{\mathrm{a}} = 1.129$ ($t = 11.07$). This means that the unobserved effect $a_i = c_i - \mathrm{E}(c_i|union_{i0}, marr_{i1}, \ldots, marr_{i7}, educ_i, black_i)$ accounts for about 56% of the unexplained variance of the composite error, $a_i + u_{it}$, where u_{it} has a conditional standard normal distribution.

Table I. Dependent variable: *union*$_t$

Explanatory variable	(1)	(2)
$marr_t$	0.168 (0.111)	0.169 (0.111)
$union_{t-1}$	0.875 (0.094)	0.886 (0.094)
$union_0$	1.514 (0.165)	1.477 (0.171)
$marr_1$	0.064 (0.209)	0.055 (0.207)
$marr_2$	−0.071 (0.256)	−0.061 (0.246)
$marr_3$	−0.129 (0.242)	−0.136 (0.242)
$marr_4$	0.025 (0.265)	0.070 (0.268)
$marr_5$	0.407 (0.246)	0.428 (0.244)
$marr_6$	0.109 (0.263)	0.079 (0.263)
$marr_7$	−0.427 (0.211)	−0.388 (0.216)
educ	–	−0.017 (0.036)
black	–	0.535 (0.194)
constant	−1.828 (0.152)	−1.712 (0.449)
$\hat{\sigma}_a$	1.129 (0.102)	1.099 (0.098)
Log-likelihood value	−1,287.48	−1,283.39

Table II. Estimated probability of being in a union, 1987

	In union, 1986	Not in union, 1986
Married, 1987	0.408	0.226
Not married, 1987	0.370	0.197

To get at the magnitudes of the state dependence, we estimate the probability of being in a union in 1987 given that the man is or is not in a union in 1986, broken down also by marital status. As discussed in Section 5.1, we average out c_i using equation (27). Specifically, Table II reports

$$N^{-1}\sum_{i=1}^{N}\Phi[(-1.828+0.0738+0.168marr_t+0.875union_{t-1}+\hat{\alpha}_1 y_{i0}+\mathbf{z}_i\hat{\boldsymbol{\alpha}}_2)/(1+1.275)^{1/2}]$$

for $union_{t-1}=0$ or 1 and $marr_t=0$ or 1, where -0.0738 is the coefficient on the 1987 year dummy and $\hat{\sigma}_a^2=1.275$. The $\hat{\boldsymbol{\alpha}}_j$ are from column (1) of Table I.

For a married man belonging to a union in 1986, the estimated probability of belonging to a union in 1987—averaged across the distribution of c_i—is 0.408. For a married man not belonging to a union in 1986, the estimated probability is 0.226. The difference, 0.182, is an estimate of the state dependence of union membership. The magnitude for unmarried men, 0.173, is similar.

One way to extend the model is to allow for the interaction term $marr_{it} \cdot union_{i,t-1}$ in the structural model. It is then natural to add the interactions between $marr_{ir}$, $r = 1, \ldots, T$ and $union_{i0}$ in the distribution $\mathrm{D}(c_i|union_{i0}, marr_{i1}, \ldots, marr_{i7})$. When added to the model in column (2) of Table I, the coefficient on $married_{it} \cdot union_{i,t-1}$ has a t statistic of only 0.84 and the p-value for exclusion of all eight interaction terms is 0.981.

7. CONCLUSIONS

I have suggested a general method for handling the initial conditions problem in a dynamic, nonlinear, unobserved effects panel data model. The key insight is that, in general nonlinear models, we can use a joint density conditional on the strictly exogenous variables *and* the initial condition. In an application of an early version of this paper, Erdem and Sun (2001) applied the approach to choice dynamics for five different products. The authors find strong evidence of state dependence in product choice.

The conditional density in Assumption 3 can be modelled flexibly, but perhaps the most important contribution of the paper is that it shows how to obtain simple estimators in dynamic probit, Tobit and Poisson unobserved effects models for specific choices of the auxiliary density. Plus, we have shown how to obtain simple estimates of the partial effects averaged across the distribution of the unobserved heterogeneity; hopefully, APEs will be routinely reported in future empirical work.

Many issues can be studied in future research. For one, it is important to know the consequences of misspecifying the density in Assumption 3. Intuitively, as the size of the cross-section increases, we can make $h(\mathbf{c}|\mathbf{y}_0, \mathbf{z}; \boldsymbol{\delta})$ more and more flexible. Unless nonlinearities in the model are caused by true data censoring, any study to evaluate the impact of various choices of $h(\mathbf{c}|\mathbf{y}_0, \mathbf{z}; \boldsymbol{\delta})$ should focus on estimates of APEs, not just on $\boldsymbol{\theta}_o$. As is well known, it often makes no sense to compare parameter estimates across different nonlinear models.

The approach proposed in Section 3 can be modified when some of the explanatory variables fail the strict exogeneity requirement. Wooldridge (2000) lays out a framework for handling models with feedback, but specific implementation issues have yet to be explored.

ACKNOWLEDGEMENTS

I would like to thank two anonymous referees, the participants at the Michigan, Michigan State, North Carolina State and Penn State econometrics workshops, and the attendees of the ESRC Study Group Econometrics Conference in Bristol, UK, July 2000, for many helpful comments and suggestions. I am especially grateful to Brandon Bartels for catching two mistakes in coding some of the variables in my original analysis. This is a revised version of 'The initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity', University of Bristol Department of Economics Working Paper No. 00/496.

REFERENCES

Ahn SC, Schmidt P. 1995. Efficient estimation of models for dynamic panel data. *Journal of Econometrics* **68**: 5–27.

Anderson TW, Hsiao C. 1982. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* **18**: 67–82.

Arellano M, Bond SR. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**: 277–297.

Arellano M, Bover O. 1995. Another look at the instrumental variables estimation of error-component models. *Journal of Econometrics* **68**: 29–51.

Arellano M, Carrasco R. 2003. Binary choice panel data models with predetermined variables. *Journal of Econometrics* **115**: 125–157.

Bhargava A, Sargan JD. 1983. Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* **51**: 1635–1659.

Blundell R, Bond S. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**: 115–143.

Blundell RW, Smith RJ. 1991. Initial conditions and efficient estimation in dynamic panel data models. *Annales d'Economie et de Statistique* **20/21**: 109–123.

Chamberlain G. 1980. Analysis of covariance with qualitative data. *Review of Economic Studies* **47**: 225–238.

Chamberlain G. 1992. Comment: sequential moment restrictions in panel data. *Journal of Business and Economic Statistics* **10**: 20–26.

Erdem T, Sun B. 2001. Testing for choice dynamics in panel data. *Journal of Business and Economic Statistics* **19**: 142–152.

Geweke J, Keane M. 1999. Mixture of normals probit models. In *Analysis of Panels and Limited Dependent Variable Models*, Hsaio C, Lahiri K, Lee L-F, Pesaran MH (eds). Cambridge University Press: Cambridge; 49–78.

Hahn J. 1999. How informative is the initial condition in the dynamic panel data model with fixed effects? *Journal of Econometrics* **93**: 309–326.

Hausman JA, Hall BH, Griliches Z. 1984. Econometric models for count data with an application to the patents–R&D relationship. *Econometrica* **52**: 909–938.

Heckman JJ. 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time–discrete data stochastic process. In *Structural Analysis of Discrete Data with Econometric Applications*, Manski CF, McFadden D (eds). MIT Press: Cambridge, MA; 179–195.

Honoré BE. 1993. Orthogonality conditions for Tobit models with fixed effects and lagged dependent variables. *Journal of Econometrics* **59**: 35–61.

Honoré BE, Kyriazidou E. 2000. Panel data discrete choice models with lagged dependent variables. *Econometrica* **68**: 839–874.

Hsiao C. 1986. *Analysis of Panel Data*. Cambridge University Press: Cambridge.

Vella F, Verbeek M. 1998. Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. *Journal of Applied Econometrics* **7**: 413–421.

Wooldridge JM. 1997. Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory* **13**: 667–678.

Wooldridge JM. 2000. A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables. *Economics Letters* **68**: 245–250.

Wooldridge JM. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press: Cambridge, MA.