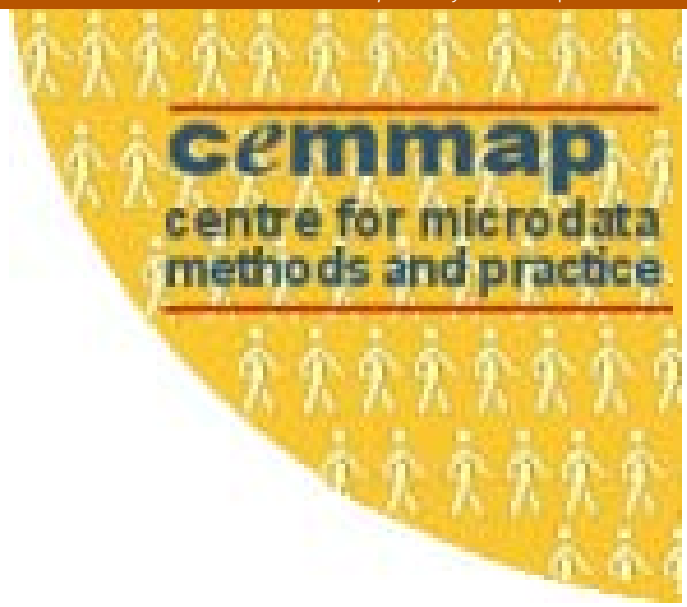

SIMPLE SOLUTIONS TO THE INITIAL CONDITIONS PROBLEM IN DYNAMIC, NONLINEAR PANEL DATA MODELS WITH UNOBSERVED HETEROGENEITY

Jeffrey M Wooldridge

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP18/02

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
East Lansing, MI 48824-1038
(517) 353-5972
WOOLDRI1@PILOT.MSU.EDU

This version: June 2002

Key Words: Panel data, dynamic model, unobserved effects, initial conditions, logit, probit, Tobit, Poisson.

JEL Classification: C33

Acknowledgements: I would like to thank the participants at the Michigan State, North Carolina State, Penn State, and University of Michigan econometrics workshops, and the attendees of the ESRC Study Group Econometric Conference in Bristol, England, July 2000, for many helpful comments and suggestions. 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 Number 00/496.

ABSTRACT

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.

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), and 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 condition 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 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 (although only relative effects can be estimated) without specifying a distribution (conditional or unconditional) of the unobserved effect. Unfortunately, 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, nor are variables that always increase for each cross-sectional unit, such as age or workforce experience. Honoré and Kyriazidou also reduce the sample to cross-sectional units with no change in the discrete covariates over the last two time periods. In practice, this could be a significant reduction in the sample size, especially considering the semiparametric estimators converge at rates less than the standard $\sqrt{N}$, where N is the size of the cross section.

Another practical limitation of the Honoré (1993) and Honoré and Kyriazidou (2000) estimators is that partial effects on the response probability or conditional mean are not identified. Therefore, the economic importance of covariates, or even 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 all of 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. In some leading cases, estimated average partial effects are easy to obtain.

Previous research in parametric, nonlinear models has primarily focused on three different ways of handling initial conditions in dynamic models with unobserved heterogeneity; these are summarized by Hsiao (1986, Section 7.4). The simplest approach is to treat the initial conditions for each cross-sectional unit as nonrandom constants. Unfortunately, this implies an untenable assumption, namely, that the initial outcome of the response variable or variables, $\mathbf{y}_{i0}$, is independent of unobserved heterogeneity, $\mathbf{c}_i$, and any observed exogenous variables. 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 modeling earnings of individuals once they leave school, and y_{i0} is earnings in the first post-school year. The fact that we observe the start of this process is logically distinct from the assumption that unobserved heterogeneity -- containing "ability" and "motivation," say -- is independent of initial earnings.

A better approach is to explicitly 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 (and observed) 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 effects probit model without additional conditioning variables (see Hsiao (1986, Section 7.4)) -- it cannot be done generally.

For the dynamic probit model with covariates, Heckman (1981) proposed approximating the conditional distribution of the initial condition. (Bhargava and Sargan (1983) effectively take this same approach for the linear AR(1) model with strictly exogenous covariates.) 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). (In this paper, I use the phrase "conditional maximum likelihood" in its most general sense: it simply means that the

likelihood function is conditional on a set of variables.) For the binary response model with a lagged dependent variable, Arellano and Carrasco (2002) 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, at least along some dimensions, 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 results in a joint distribution of outcomes *after* the initial period conditional on the initial value and any strictly exogenous variables. This approach has several advantages. First, we are free to choose the auxiliary distribution so as to be flexible or convenient. Because we are not specifying the distribution of the initial value, conditonal on unobserved heterogeneity, we need not even consider the notion of a steady-state distribution. Of course, we might just view the approach here as a different approximation that has some computational advantages. Second, in several leading cases -- probit, ordered probit, Tobit, and Poisson regression -- an auxiliary distribution can be chosen that leads to a straightforward parameterization that can be estimated 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 in order to highlight the important issues; we return to these examples in Section 5. In all of the examples, we assume random sampling in the cross section dimension, where the cross section (N) is large relative to the number of time periods (T). The asymptotic analysis is for fixed T.

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

$$P(y_{it} = 1|y_{i,t-1},\ldots,y_{i0},\mathbf{z}_i,c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\gamma} + \rho y_{i,t-1} + c_i). \qquad (2.1)$$

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. (We could specify the logit function, rather than the probit function, but we focus on probit here.) 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, if appropriate, of exogenous variables.)

As we will see in Sections 3 and 4, the parameters in (2.1), as well as average partial effects, can be estimated by specifying a density for c_i given $(y_{i0},\mathbf{z}_i)$. A homoskedastic normal distribution with conditional mean linear in parameters is especially convenient, as we will see in Section 5. The typical approaches to this model are computationally more difficult; see,

for example, Hsiao (1986, Section 7.4). ■

EXAMPLE 2 (Dynamic Tobit Model with Unobserved Effect): We write a dynamic Tobit model as

$$y_{it} = \max[0, \mathbf{z}_{it}\boldsymbol{\gamma} + \mathbf{g}(y_{i,t-1})\boldsymbol{\rho} + c_i + u_{it}] \tag{2.2}$$

$$u_{it} | y_{i,t-1}, \ldots, y_{i0}, \mathbf{z}_i, c_i \sim \text{Normal}(0, \sigma_u^2), \tag{2.3}$$

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 (such as top-coded data or durations), as in that case we would want a lagged value of the latent variable underlying (2.2) to appear. The function $\mathbf{g}(\cdot)$ is generally a vector function, which allows the lagged value of the *observed* response to appear in a variety of ways. For example, 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.

A maximum likelihood approach that treats the c_i as parameters to estimate is computationally difficult, and inconsistent for fixed T. Little is known about the properties of such an estimator for various T. Honoré (1993) proposes orthogonality conditions that identify the parameters, but average partial effects are apparently unidentified. We will show how to obtain $\sqrt{N}$-consistent estimates of the parameters as well as 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

$$E(y_{it}|y_{i,t-1},\ldots,y_{i0},\mathbf{z}_i,c_i) = c_i\exp[\mathbf{z}_{it}\boldsymbol{\gamma} + \mathbf{g}(y_{i,t-1})\boldsymbol{\rho}]. \qquad (2.4)$$

Again, we allow for the lagged dependent variable to appear in a flexible fashion. For example, this could consist of a set of dummy variables for specific outcomes on $y_{i,t-1}$. To test the null hypothesis of no state dependence, we test H_0: $\boldsymbol{\rho} = \mathbf{0}$. A reasonable analysis allows c_i to be correlated with the initial condition and $\mathbf{z}_i$. Chamberlain (1992) and Wooldridge (1997) have proposed orthogonality conditions based only on (2.4), 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. In Section 5 we show how a particular model for a conditional distribution for c_i leads to a straightforward maximum likelihood analysis. ■

3. GENERAL FRAMEWORK

3.1. Random Sampling

In this section we let i denote a random draw from the cross section, and let t denote a time period. We assume that we observe $(\mathbf{z}_{it},\mathbf{y}_{it})$ for $t = 1,\ldots,T$, and we observe y_{i0}. In the general framework, we are interested in the conditional distribution of $\mathbf{y}_{it} \in \mathbb{R}^G$ 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 $\mathbf{c}_i \in \mathbb{R}^J$ is unobserved heterogeneity. (In the general setup, the dimension of $\mathbf{z}_{it}$ can change with

t, although in our examples the dimension of $\mathbf{z}_{it}$ is fixed.) We denote the conditional distribution by $D(\mathbf{y}_{it}|\mathbf{z}_{it},\mathbf{y}_{i,t-1},\mathbf{c}_i)$. The asymptotic analysis is with the number of time periods, T, fixed, and 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 A.1: For $t = 1,2,\ldots,T$,

$$D(\mathbf{y}_{it}|\mathbf{z}_{it},\mathbf{y}_{i,t-1},\mathbf{c}_i) = D(\mathbf{y}_{it}|\mathbf{z}_i,\mathbf{y}_{i,t-1},\ldots,\mathbf{y}_{i0},\mathbf{c}_i). \quad \blacksquare \qquad (3.1)$$

We could allow for additional lags of $\mathbf{y}_{it}$ in the conditional distribution but, as we will see below in stating Assumption A.3, we would generally need more time periods.

We next assume that we have a correctly specified parametric model for the density representing (3.1) which, for lack of a better name, we call the "structural" density.

ASSUMPTION A.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 (3.1), with respect to a σ-finite measure $\nu(d\mathbf{y}_t)$. The parameter space, Θ, is a subset of $\mathbb{R}^P$. Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_o \in \Theta$. $\blacksquare$

The requirement that we have a density with respect to a σ-finite measure is

not restrictive in practice. (The assumption that this measure does not depend on t is also not very restrictive.) If $\mathbf{y}_t$ is purely discrete, ν is counting measure. If $\mathbf{y}_t$ is continuous, ν is Lebesgue measure. An appropriate σ-finite measure can be found for all of the possible response variables of interest in economics, including those that are neither purely discrete nor purely continuous (such as a Tobit response). In this section, we do not need the measure explicitly, but we do refer to it in Section 4.

Most specific analyses of dynamic, nonlinear unobserved effects models begin with assumptions very similar to A.1 and A.2. (Examples include dynamic logit, probit, and Tobit models. An exception is Honoré and Kyriazidou (2000), who consider the dynamic binary response model without specifying the response probability. But they can only get consistency of the parameters up to scale, the estimator converges at a rate slower than $\sqrt{N}$, and it is very unlikely the estimator has an asymptotic normal distribution when properly scaled.) Together, A.1 and A.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

$$\prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta}_o), \tag{3.2}$$

where we drop the i subscript to indicate dummy arguments of the density. In using (3.2) to estimate $\boldsymbol{\theta}_o$, 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

$$\sum_{i=1}^{N}\sum_{t=1}^{T} \log f_t(\mathbf{y}_{it}|\mathbf{z}_{it},\mathbf{y}_{i,t-1},\mathbf{c}_i;\boldsymbol{\theta}). \tag{3.3}$$

over $\boldsymbol{\theta}$ and $(\mathbf{c}_1,\ldots,\mathbf{c}_N)$. While this approach avoids having to restrict the distribution of $\mathbf{c}_i$ -- conditional or unconditional -- it is computationally difficult. More importantly, with fixed T, it suffers from an incidental

parameters problem: except in very special cases, the estimator of $\boldsymbol{\theta}_o$ 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. The first, which is to treat the $\mathbf{y}_{i0}$ as fixed, is the same as assuming $\mathbf{y}_{i0}$ is independent of $(\mathbf{z}_i,\mathbf{c}_i)$. Generally, this is too strong an assumption.

We can follow the general route of attempting to find the density for $(\mathbf{y}_{i0},\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT})$ given $\mathbf{z}_i$. If we specify $f(\mathbf{y}_0|\mathbf{z},\mathbf{c})$ then

$$f(\mathbf{y}_0,\mathbf{y}_1,\ldots,\mathbf{y}_T|\mathbf{z},\mathbf{c}) = f(\mathbf{y}_1,\ldots,\mathbf{y}_T|\mathbf{y}_0,\mathbf{z},\mathbf{c})\cdot f(\mathbf{y}_0|\mathbf{z},\mathbf{c}). \qquad (3.4)$$

Next, we specify a density $f(\mathbf{c}|\mathbf{z})$. We can then integrate (3.4) with respect to this density to obtain $f(\mathbf{y}_0,\mathbf{y}_1,\ldots,\mathbf{y}_T|\mathbf{z})$. This approach requires specifying a model for $f(\mathbf{y}_0|\mathbf{z},\mathbf{c})$ and $f(\mathbf{c}|\mathbf{z})$, and can be computationally demanding. Plus, sample selection on the basis of the initial condition $\mathbf{y}_{i0}$ generally leads to inconsistency of the MLE (see Section 3.2).

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 (3.2) -- we need only specify the density of $\mathbf{c}_i$ conditional on $(\mathbf{y}_{i0},\mathbf{z}_i)$. Because this density is not restricted in any way by the specification in Assumption A.2, we can choose it for convenience, or flexibility, or, hopefully, both. (Even if $f(\mathbf{y}_0|\mathbf{z},\mathbf{c})$ is restricted by Assumption A.2 -- for example, by our desire to have the steady-state distribution -- $f(\mathbf{c}|\mathbf{y}_0,\mathbf{z})$ is not restricted because $f(\mathbf{c}|\mathbf{z})$ is not restricted.) 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 $\boldsymbol{\theta}_o$. Specifying a model for $f(\mathbf{c}|\mathbf{y}_0,\mathbf{z})$ seems no worse than having to specify models, which themselves can only be approximate, for $f(\mathbf{y}_0|\mathbf{z},\mathbf{c})$. Further, as we will see in Section 4, we are also able to estimate a variety of average partial effects.

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

Technically, we need to introduce the σ-finite measure, η, in Assumption A.3. In practice, the measure would be either Lebesgue measure -- when $\mathbf{c}_i$ is assumed to have a continuous distribution, and so integrals involving $\mathbf{c}_i$ are the usual Riemann integrals -- or η would be the counting measure if $\mathbf{c}_i$ is discrete, in which case the integrals are weighted averages.

Assumption A.3 is much more controversial than Assumptions A.1 and A.2. Ideally, we would not have to specify anything about the relationship between $\mathbf{c}_i$ and $(\mathbf{y}_{i0},\mathbf{z})$, whereas A.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 A.3. We mentioned several of these in the introduction and in Section 2. But these estimators are complicated and need not have particularly good statistical properties (although they are consistent without Assumption A.3). Another problem with semiparametric estimators often goes unnoticed: in nonlinear models where unobserved effects are correlated with explanatory variables, semiparametric methods, essentially by construction, do not allow us to recover the partial effects

of interest, because these depend on the distribution of the unobserved heterogeneity. Therefore, while we can often estimate the directions of the effect of a policy under weaker assumptions, we cannot estimate the size of the effect. As we will see in Section 4, by imposing Assumption A.3, we are able to identify and estimate average partial effects.

To make the asymptotics straightforward, we assume the density in A.3 depends on a parameter vector, $\boldsymbol{\delta}$, with fixed dimension. This makes our analysis traditionally parametric. Alternatively, we could use a seminonparametric approach, as in Gallant and Nychka (1987). Unfortunately, the limiting distribution results when the dimension of $\boldsymbol{\delta}$ is allowed to increase with N are not generally available for nonlinear models. Plus, generaly identifiability of $\boldsymbol{\theta}_o$ would become an issue. In practice, researchers applying seminonparametric methods choose flexible forms for the auxiliary densities -- in this case, $h(\mathbf{c}|\mathbf{y}_0,\mathbf{z};\boldsymbol{\delta})$ -- but, for inference, use the usual parametric asymptotics.

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

$$\int_{\mathbb{R}^J}\left(\prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta}_o)\right)h(\mathbf{c}|\mathbf{y}_0,\mathbf{z};\boldsymbol{\delta}_o)\eta(d\mathbf{c}), \tag{3.5}$$

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

$$\ell_i(\boldsymbol{\theta},\boldsymbol{\delta}) = \log\left[\int_{\mathbb{R}^J}\left(\prod_{t=1}^{T} f_t(\mathbf{y}_{it}|\mathbf{z}_{it},\mathbf{y}_{i,t-1},\mathbf{c};\boldsymbol{\theta})\right)h(\mathbf{c}|\mathbf{y}_{i0},\mathbf{z}_i;\boldsymbol{\delta})\eta(d\mathbf{c})\right]. \tag{3.6}$$

To estimate $\boldsymbol{\theta}_o$ and $\boldsymbol{\delta}_o$, we sum the log likelihoods in (3.6) across $i = 1,\ldots,N$ and maximize with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$. The resulting conditional MLE (CMLE) is $\sqrt{N}$-consistent and asymptotically normal under standard regularity conditions. (One set of conditions is covered in Wooldridge (2002, Chapter

13).) 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 $D(\mathbf{c}_i|\mathbf{y}_{i0},\mathbf{z}_i)$ 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 always include time-constant explanatory variables in $\mathbf{z}_{it}$, but, unless we make specific exclusion restrictions, we will not be able to identify separate the partial effect of the time-constant variable from its correlation with $\mathbf{c}_i$.

If $\mathbf{z}_{it}$ contains only contemporaneous variables, then the time-zero value, $\mathbf{z}_{i0}$, does not appear in $D(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT}|\mathbf{y}_{i0},\mathbf{z}_i,\mathbf{z}_{i0},\mathbf{c}_i)$. Nevertheless, we could model the density for $D(\mathbf{c}_i|\mathbf{y}_{i0},\mathbf{z}_i,\mathbf{z}_{i0})$ by including $\mathbf{z}_{i0}$, in which case the likelihood function is conditional on $(\mathbf{y}_{i0},\mathbf{z}_i,\mathbf{z}_{i0})$. If the structural model is orignally specified for $D(\mathbf{y}_{it}|\mathbf{z}_{it},\mathbf{y}_{i,t-1},\mathbf{z}_{i,t-1},\mathbf{c}_i)$ -- so that a lag of $\mathbf{z}_{it}$ is included along with a lag of $\mathbf{y}_{it}$ -- then $\mathbf{z}_{i0}$ must appear in the final conditioning set.

If we want to expand the structural model in (3.1) to allow, say, $\mathbf{y}_{i,t-2}$ in the conditional distribution, then the density in Assumption A.3 would be for $D(\mathbf{c}_i|\mathbf{y}_{i0},\mathbf{y}_{i,-1},\mathbf{z}_i)$, where $\mathbf{y}_{i0}$ and $\mathbf{y}_{i,-1}$ are the first two initial values. This increases the data requirements. With larger T we can afford to be more flexible in the dynamics in the structural model.

3.2. Sample Selection and Attrition

We derived the log-likelihood in Section 3.1 under the assumption that we

observe data on all cross-sectional units in all time periods. 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 observing $\mathbf{y}_{i0}$), 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 because the log-likelihood on the restricted sample is simply $\sum_{i=1}^{N} s_i \ell_i(\boldsymbol{\theta},\boldsymbol{\delta})$. Now, for each i, $E[s_i\ell_i(\boldsymbol{\theta},\boldsymbol{\delta})] = E\{E[s_i\ell_i(\boldsymbol{\theta},\boldsymbol{\delta})|\mathbf{y}_{i0},\mathbf{z}_i]\} = E\{E(s_i|\mathbf{y}_{i0},\mathbf{z}_i)\cdot E[\ell_i(\boldsymbol{\theta},\boldsymbol{\delta})|\mathbf{y}_{i0},\mathbf{z}_i]\}$, where the first equality follows by iterated expectations and the second follows by the conditional independence assumption between $(\mathbf{y}_{i1},\ldots,\mathbf{y}_{iT})$ and s_i. But $(\boldsymbol{\theta}_o,\boldsymbol{\delta}_o)$ maximizes $E[\ell_i(\boldsymbol{\theta},\boldsymbol{\delta})|\mathbf{y}_{i0},\mathbf{z}_i]$ for all $(\mathbf{y}_{i0},\mathbf{z}_i)$. Therefore, $(\boldsymbol{\theta}_o,\boldsymbol{\delta}_o)$ maximizes $E[s_i\ell_i(\boldsymbol{\theta},\boldsymbol{\delta})]$; provided $P(s_i = 1|\mathbf{y}_{i0},\mathbf{z}_i)$ is sufficiently bounded from zero, and standard identification conditions hold on the original model, $(\boldsymbol{\theta}_o,\boldsymbol{\delta}_o)$ is still identified by the log likelihood using the balanced panel. See Wooldridge (2002, Chapter 17) for further discussion.

When sample selection and attrition are 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 selection and attrition to depend on the initial condition, $\mathbf{y}_{i0}$. For example, if y_{i0} is annual hours worked, an MLE analysis based on the conditional log-likelihood (3.6) allows attrition to differ across initial hours worked; in particular workers who were initially unemployed are allowed to having missing data probabilities different from

full- or part-time workers. In the traditional approach, one would have to explicitly model selection or attrition as a function of $\mathbf{y}_{i0}$, and do a complicated Heckit-type analysis.

Of course, reducing the data set to a balanced panel potentially discards a lot of information. But the 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.

Similar comments apply to stratified sampling. Any stratification that is a function of $(\mathbf{y}_{i0},\mathbf{z}_i)$ can be ignored in the conditional MLE analysis. In fact, it is more efficient not to use any sampling weights. See Wooldridge (1999, 2001) for a general treatment of stratification based on endogenous and conditioning variables. By contrast, the usual approach of finding a joint density of $(\mathbf{y}_{iT},...,\mathbf{y}_{i1},\mathbf{y}_{i0})$ given $\mathbf{z}_i$ requires estimation using sampling weights if stratification depends on $\mathbf{y}_{i0}$.

4. ESTIMATING AVERAGE PARTIAL EFFECTS

As mentioned in the introduction and in Section 2, in nonlinear models it is often insufficient to have consistent, asymptotically normal estimators of the parameters, $\boldsymbol{\theta}_o$. For example, even in a standard binary response model for cross section data, without unobserved heterogeneity, the sizes of the coefficients do not allow us to determine the effects of the covariates on the response probabilities. Instead, in models such as logit and probit, the partial effects on the response probability -- evaluated at interesting

values of the covariates, or averaged across the covariates -- are usually reported. The same is true of Tobit models applied to corner solution outcomes, where the effects of the covariates on mean responses are of primary interest.

In dynamic panel data models with unobserved effects, estimating partial effects is even more complicated, and the semiparametric literature has been mostly silent on this issue. Typically, we would like the effect on the mean response after averaging the unobserved heterogeneity across the population. Essentially by construction, semiparametric approaches do not allow for estimation of average partial effects: the main goal of semiparametric methods in panel data contexts is to estimate parameters without making distributional assumptions on the unobserved effects. In this section, I show that average partial effects (APEs) are generally identified, and I propose consistent, $\sqrt{N}$-asymptotically normal estimators. When we apply the estimators to specific examples in Section 5, we obtain some particularly simple estimators. Estimating average partial effects allows us to determine the importance of any dynamics in the model, as opposed to just testing whether there are dynamics.

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. In other words, we are interested in

$$m(\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta}) = \mathrm{E}[q(\mathbf{y}_{it})|\mathbf{z}_{it} = \mathbf{z}_t,\mathbf{y}_{i,t-1} = \mathbf{y}_{t-1},\mathbf{c}_i = \mathbf{c})]$$
$$= \int_{\mathbb{R}^G} q(\mathbf{y}_t) f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta}_o)\nu(\mathrm{d}\mathbf{y}_t), \qquad (4.1)$$

where $\nu(\mathrm{d}\mathbf{y}_t)$ is the measure for the density f_t (see Assumption A.2) and $\mathbf{z}_t$, $\mathbf{y}_{t-1}$, and $\mathbf{c}$ are values that we must choose. Unfortunately, since the

unobserved heterogeneity rarely, if ever, has natural units of measurements, it is unclear which values we should plug in for $\mathbf{c}$. One possibility is the population mean value, $\boldsymbol{\alpha}_o = E(\mathbf{c}_i)$. Under Assumptions, A.1, A.2, and A.3, $\boldsymbol{\alpha}_o$ is identified. To see this, by iterated expectations we have $\boldsymbol{\alpha}_o = E[E(\mathbf{c}_i|\mathbf{y}_{i0},\mathbf{z}_i)] = E[\mathbf{a}(\mathbf{y}_{i0},\mathbf{z}_i;\boldsymbol{\delta}_o)]$, where

$$\mathbf{a}(\mathbf{y}_{i0},\mathbf{z}_i;\boldsymbol{\delta}_o) = \int_{\mathbb{R}^J} \mathbf{c}h(\mathbf{c}|\mathbf{y}_{i0},\mathbf{z}_i;\boldsymbol{\delta}_o)\eta(d\mathbf{c}). \tag{4.2}$$

Equation (4.2) is simply the expectation of $\mathbf{c}_i$ conditional on $(\mathbf{y}_{i0},\mathbf{z}_i)$, and it can be found by Assumption A.3. Because the CMLE $\hat{\boldsymbol{\delta}}$ is a consistent, $\sqrt{N}$-asymptotically normal estimator of $\boldsymbol{\delta}_o$, a consistent, $\sqrt{N}$-asymptotically normal estimator of $\boldsymbol{\alpha}_o$ is

$$\hat{\boldsymbol{\alpha}} = N^{-1}\sum_{i=1}^{N} \mathbf{a}(\mathbf{y}_{i0},\mathbf{z}_i;\hat{\boldsymbol{\delta}}). \tag{4.3}$$

A consistent estimator of $m(\mathbf{z}_t,\mathbf{y}_{t-1},\boldsymbol{\alpha}_o;\boldsymbol{\theta}_o)$ is then

$$m(\mathbf{z}_t,\mathbf{y}_{t-1},\hat{\boldsymbol{\alpha}};\hat{\boldsymbol{\theta}}) = \int_{\mathbb{R}^G} q(\mathbf{y}_t)f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\hat{\boldsymbol{\alpha}};\hat{\boldsymbol{\theta}})\nu(d\mathbf{y}_t). \tag{4.4}$$

We can estimate partial effects by computing derivatives of (4.4) with respect to elements of $(\mathbf{z}_t,\mathbf{y}_{t-1})$ or computing differences with respect to elements of $(\mathbf{z}_t,\mathbf{y}_{t-1})$. As we will see in Section 5, (4.4) is straightforward to compute when $q(y_t) = y_t$ for probit and Tobit models. (On the other hand, obtaining standard errors is more challenging. We can use the delta method or, perhaps, bootstrapping.)

One problem with evaluating $m(\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\hat{\boldsymbol{\theta}})$ at $\mathbf{c} = \hat{\boldsymbol{\alpha}}$ is that it estimates partial effects at the population unit with the average heterogeneity, and this may apply to only a small fraction of the population. If $\mathbf{c}_i$ has a continuous distribution, $m(\mathbf{z}_t,\mathbf{y}_{t-1},\hat{\boldsymbol{\alpha}};\hat{\boldsymbol{\theta}})$ technically represents none of the population (because $P(\mathbf{c}_i = \boldsymbol{\alpha}_o) = 0$).

An alternative is to average $m(\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta}_o)$ across the distribution of

$\mathbf{c}_i$. That is, we estimate

$$\mu(\mathbf{z}_t,\mathbf{y}_{t-1}) = \mathrm{E}[m(\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c}_i;\boldsymbol{\theta}_o)], \tag{4.5}$$

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 A.1, A.2, and A.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 (4.5). Fortunately, this is not the case. We can obtain a consistent estimator of (4.5) using iterated expectations:

$$\begin{aligned}\mathrm{E}[m(\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c}_i;\boldsymbol{\theta}_o)] &= \mathrm{E}\{\mathrm{E}[m(\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c}_i;\boldsymbol{\theta}_o)|\mathbf{y}_{i0},\mathbf{z}_i]\} \\ &= \mathrm{E}\left[\left(\int_{\mathbb{R}^G} q(\mathbf{y}_t) f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta}_o)\nu(\mathrm{d}\mathbf{y}_t)\right) h(\mathbf{c}|\mathbf{y}_{i0},\mathbf{z}_i;\boldsymbol{\delta}_o)\eta(\mathrm{d}\mathbf{c})\right], \end{aligned} \tag{4.6}$$

where the outside expectation is with respect to the distribution of $(\mathbf{y}_{i0},\mathbf{z}_i)$. While (4.6) 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 $f_t(\mathbf{y}_t|\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c};\boldsymbol{\theta}_o)$; often the expectation is available in closed form. Typically, the hard part is integrating $m(\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{c})$ with respect to $h(\mathbf{c}|\mathbf{y}_{i0},\mathbf{z}_i;\boldsymbol{\delta}_o)$.

One point worth emphasizing about (4.6) is that $\boldsymbol{\delta}_o$ appears explicitly. In other words, while $\boldsymbol{\delta}_o$ may be properly viewed as a nuisance parameter for estimating $\boldsymbol{\theta}_o$, it is not a nuisance parameter for estimating APEs. Because the semiparametric literature treats $\boldsymbol{\delta}_o$ as a nuisance parameter -- more generally, $h(\mathbf{c}|\mathbf{y}_0,\mathbf{z})$ is a nuisance function -- there seems little hope that semiparametric approaches will generally deliver consistent, let alone $\sqrt{N}$-asymptotically normal, estimates of APEs in dynamic, unobserved effects

panel data models.

Given (4.6), a consistent estimator of $q(\mathbf{z}_t,\mathbf{y}_{t-1})$ follows immediately:

$$N^{-1}\sum_{i=1}^{N} r(\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{y}_{i0},\mathbf{z}_i;\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\delta}}), \tag{4.7}$$

where $r(\mathbf{z}_t,\mathbf{y}_{t-1},\mathbf{y}_{i0},\mathbf{z}_i;\boldsymbol{\theta}_o,\boldsymbol{\delta}_o)$ is the function inside the expectation in (4.6). This is a $\sqrt{N}$-asymptotically normal estimator of $\mu(\mathbf{z}_t,\mathbf{y}_{t-1})$. Note that this estimator is not in any way conditional on the initial conditions, $\mathbf{y}_{i0}$, or the exogenous variables, $\mathbf{z}_i$: we are averaging these out over a large cross section, which gives us a consistent estimator of the mean in the population.

In order for (4.7) to be consistent for $q(\mathbf{z}_t,\mathbf{y}_{t-1})$, we assume a random sample from the population. If the sample has been stratified on the basis of $(\mathbf{y}_{i0},\mathbf{z}_i)$ then we would replace (4.7) with a weighted average, where the weights are the inverse probability sampling weights. If the sample is selected on the basis of $(\mathbf{y}_{i0},\mathbf{z}_i)$, we would generally have to model the selection probability, $P(s_i = 1|\mathbf{y}_{i0},\mathbf{z}_i)$, in order to consistently estimate the APE. See Wooldridge (2002b).

5. THE EXAMPLES REVISITED

We now reconsider the examples from Section 2, showing how we can apply the results from Sections 3 and 4. We emphasize that, for certain choices of the density $h(\mathbf{c}|\mathbf{y}_0,\mathbf{z};\boldsymbol{\delta})$ in Assumption A.3, very convenient simplications exist for many leading cases. For notational simplicity, we drop the "o" subscript on the true values of the parameters.

5.1. Dynamic Binary and Ordered Response Models

In addition to (2.1), assume that

$$c_i|y_{i0},\mathbf{z}_i \sim \text{Normal}(\alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2,\sigma_a^2), \tag{5.1}$$

where $\mathbf{z}_i$ is the row vector of all (nonredundant) explanatory variables in all time periods. If, as occurs in many applications, $\mathbf{z}_{it}$ 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 (5.1) 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 (5.1).

Given (2.1) and (5.1), we can write

$$f(y_1,y_2,\ldots,y_T|y_0,\mathbf{z},c;\boldsymbol{\beta}) = \prod_{t=1}^{T}\{\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + c)^{y_t} \cdot [1 - \Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + c)]^{1-y_t}\}, \tag{5.2}$$

where $\boldsymbol{\beta} = (\boldsymbol{\gamma}',\rho)'$. When we integrate this with respect to the normal distribution in (5.1), we obtain the density of $(y_{i1},\ldots,y_{iT}|y_{i0},\mathbf{z}_i)$.

Interestingly, we can specify the integrated density in such a way that standard random effects probit software can be used for estimation. If we write

$$c_i = \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i, \tag{5.3}$$

where a_i is independent of $(y_{i0},\mathbf{z}_i)$ and distributed as 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

$$\Phi(\mathbf{z}_{it}\boldsymbol{\gamma} + \rho y_{i,t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i). \tag{5.4}$$

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 (5.3):

$$y_{it}^* = \mathbf{z}_{it}\boldsymbol{\gamma} + \rho y_{i,t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i + u_{it}. \tag{5.5}$$

Equation (5.4) follows from (5.5) by noting that u_{it} given $(\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) \sim \text{Normal}(0,1)$. It follows that the density of $(y_{i1}, \ldots, y_{iT})$ given $(y_{i0} = y_0, \mathbf{z}_i = \mathbf{z}, a_i = a)$ is

$$\prod_{t=1}^{T} \{\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_1 y_0 + \mathbf{z}\boldsymbol{\alpha}_2 + a)^{y_t} \cdot [1 - \Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_1 y_0 + \mathbf{z}\boldsymbol{\alpha}_2 + a)]^{1-y_t}\}. \tag{5.6}$$

Therefore, the density of $(y_{i1}, \ldots, y_{iT})$ given $(y_{i0} = y_0, \mathbf{z}_i = \mathbf{z})$ is obtained by integrating (5.6) against the $\text{Normal}(0, \sigma_a^2)$ density:

$$\int_{\mathbb{R}} \left(\prod_{t=1}^{T} \{\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_1 y_0 + \mathbf{z}\boldsymbol{\alpha}_2 + a)^{y_t} \cdot [1 - \Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_1 y_0 + \mathbf{z}\boldsymbol{\alpha}_2 + a)]^{1-y_t}\} (1/\sigma_a)\phi(a/\sigma_a) da. \right. \tag{5.7}$$

Interestingly, the likelihood in (5.7) has exactly the same structure as the standard random effects probit model, except that the explanatory variables at time period t are

$$\mathbf{x}_{it} \equiv (1, \mathbf{z}_{it}, y_{i,t-1}, y_{i0}, \mathbf{z}_i). \tag{5.8}$$

Importantly, we are not saying that a_i is independent of $y_{i,t-1}$, which is clearly impossible. Further, the density in (5.7) is clearly 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 precisely what (5.7) 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 $\boldsymbol{\gamma}$, ρ, α_0, α_1, $\boldsymbol{\alpha}_2$, and σ_a^2. (We might want to conserve on degrees of freedom by, say, using the time average, $\bar{\mathbf{z}}_i$, in place of $\mathbf{z}_i$.) The estimate of α_1 is of interest in its own right, as it tells us the direction of the relationship between c_i and y_{i0}. (Incidentally, for applications that

include aggregate time effects -- which is usually warranted -- (5.8) makes it clear that aggregate time dummies should appear only once in $\mathbf{x}_{it}$, as these do not vary across i.)

Unlike in semiparametric approaches to these models, we can easily obtain estimated partial effects at interesting values of the explanatory variables. As discussed in Section 4, there are two possibilities. First, we can evaluate the standard normal cdf at an estimate of $E(c_i)$. But $E(c_i) = \alpha_0 + \alpha_1 E(y_{i0}) + E(\mathbf{z}_i)\boldsymbol{\alpha}_2$, which is consistently estimated by $\hat{\alpha}_0 + \hat{\alpha}_1 \bar{y}_0 + \bar{\mathbf{z}}\hat{\boldsymbol{\alpha}}_2$, where the estimates are the conditional MLEs and the overbars represent averages in the cross section. For example, $\bar{y}_0$ is the proportion of people with $y_{i0} = 1$. Then, we can compute derivatives or differences of

$$\Phi(\mathbf{z}_t\hat{\boldsymbol{\gamma}} + \hat{\rho}y_{t-1} + \hat{\alpha}_0 + \hat{\alpha}_1\bar{y}_0 + \bar{\mathbf{z}}\hat{\boldsymbol{\alpha}}_2) \tag{5.9}$$

with respect to elements of $\mathbf{z}_t$ or y_{t-1}, and evaluate these at interesting values of $\mathbf{z}_t$ and y_{t-1}.

We can also estimate the average partial effects on the response probabilities which, in this model, are based on

$$E[\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + c_i)], \tag{5.10}$$

where the expectation is with respect to the distribution of c_i. The general formula in (4.7) 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 (5.8) is

$$E[\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i)], \tag{5.11}$$

where the expectation is over the distribution of $(y_{i0}, \mathbf{z}_i, a_i)$. Now, just as in Section 4, we use iterated expectations:

$$E[\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i)]$$
$$= E\{E[\Phi(\mathbf{z}_t\boldsymbol{\gamma} + \rho y_{t-1} + \alpha_0 + \alpha_1 y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i) | y_{i0}, \mathbf{z}_i]\}. \tag{5.12}$$

The conditional expectation inside (5.12) is easily shown to be

$$\Phi(\mathbf{z}_t\boldsymbol{\gamma}_a + \rho_a y_{t-1} + \alpha_{a0} + \alpha_{a1} y_{i0} + \mathbf{z}_i\boldsymbol{\alpha}_{a2}), \tag{5.13}$$

where the *a* subscript denotes the original parameter multiplied by $(1 + \sigma_a^2)^{-1/2}$. Now, we want to estimate the expected value of (5.11) with respect to the distribution of $(y_{i0}, \mathbf{z}_i)$. A consistent estimator is

$$N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{z}_t\hat{\boldsymbol{\gamma}}_a + \hat{\rho}_a y_{t-1} + \hat{\alpha}_{a0} + \hat{\alpha}_{a1} y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\alpha}}_{a2}), \tag{5.14}$$

where the *a* subscript now denotes multiplication by $(1 + \hat{\sigma}_a^2)^{-1/2}$, and $\hat{\boldsymbol{\gamma}}$, $\hat{\rho}$, $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\boldsymbol{\alpha}}_2$, and $\hat{\sigma}_a^2$ are the conditional MLEs. We can compute changes or derivatives of equation (5.14) with respect to $\mathbf{z}_t$ or y_{t-1} to obtain average partial effects.

Equation (5.14) extends Chamberlain's (1984, equation (3.4)) method of computing partial effects in the probit model with strictly exogenous explanatory variables. The delta method can be used to obtain asymptotic standard errors for these average effects. See, for example, Newey and McFadden (1994).

The importance of an estimate such as (5.14) is that it allows us to determine the magnitudes of partial effects, including the importance of any state dependence. While semiparametric approaches allow us to test for state dependence, we cannot generally conclude whether state dependence is economically important. The same can be said of Chamberlain's (1978) test for state dependence: while it can be made robust to structural serial correlation, it does not provide an estimate of the importance of the state dependence.

It is straightforward to allow a more flexible conditional mean in (5.1), provided it is linear in parameters. For examples, including interactions between y_{i0} and $\mathbf{z}_i$ is simple. Allowing for heteroskedasticity is more complicated and would probably require special programming.

Specification testing is relatively easy. 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. Score tests for, say, exponential heteroskedasticity in $\text{Var}(c_i|y_{i0},\mathbf{z}_i)$, or nonnormality in $\text{D}(c_i|y_{i0},\mathbf{z}_i)$, would be valuable.

The same kind of derivation goes through if we replace $\Phi(\cdot)$ in (5.7) with the logit function. The parameters can be estimated using standard random effects logit software where the explanatory variables are as in (5.8) and the unobserved effect has a normal distribution.

A dynamic ordered probit (or ordered logit) model would also be fairly straightforward to estimate using the current approach. Suppose that y_{it} takes on one of the values in $\{0,1,...,J\}$. Then, we can specify y_{it} as following an ordered probit model with J lagged indicators, $1[y_{i,t-1} = j]$, $j = 1,...,J$, and strictly exogenous explanatory variables, $\mathbf{z}_{it}$. So, 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 end points or cut points must be estimated. If we specify $c_i|y_{i0},\mathbf{z}_i$ as having a homoskedastic normal distribution, standard random effects ordered probit (or random effects ordered logit) software can be used. Probably we would allow $h(c|y_0,\mathbf{z};\boldsymbol{\delta})$ to depend on a full set of indicators, $1[y_{i0} = j]$, $j = 1,...,J$, which describe all of the possible states for the initial outcome.

Certainly there are some criticisms that one can make about the conditional MLE approach in this example. First, suppose that there are no covariates, so that (5.1) reduces to $c_i|y_{i0} \sim \text{Normal}(\alpha_0 + \alpha_1 y_{i0},\sigma_a^2)$. Unless

$\alpha_1 = 0$, this assumption implies that c_i has a mixture of normals distribution [with mixing probability $P(y_{i0} = 1)$], 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. In fact, 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 (though this can be relaxed). But in many applications we assume that unobserved heterogeneity has a conditional normal distribution rather than an unconditional normal distribution.

A related 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 (5.1). Plus, Chamberlain's static model does not allow estimation of either ρ or the amount of state dependence, as measured by the average partial effect. (In an application of an early version of this paper, Erdem and Sun (2001) find evidence for $\rho \neq 0$ in the choice dynamics for five different products. Interestingly, they cannot reject $\alpha_1 = 0$ in any case.)

Another criticism of an assumption like (5.1) is the same criticism that has been aimed at 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 (5.1) imposes distributional restrictions on the $\mathbf{z}_{it}$. For example, suppose that $\alpha_1 = 0$. Then, for (5.1) to hold for both T and $T - 1$, $\mathbf{z}_{iT}\boldsymbol{\alpha}_{2T}$ given $(\mathbf{z}_{i1},...,\mathbf{z}_{i,T-1})$ would have to have a normal distribution. While theoretically this is a valid criticism, it is hardly unique to this setting. For example, suppose we specify a probit model for employment status, based on a set of characteristics that exclude health status. Later, a binary indicator indicating bad health becomes available. If we add the health indicator to the covariates, the correct model can no longer be probit. In fact, every time an explanatory variable is added to a probit or Tobit analysis, the probit or Tobit model can no longer hold unless the new variable is normally distributed. It seems counterproductive to worry about the logical inconsistencies that arise when estimating nonlinear models with different sets of explanatory variables. Such considerations make a strong theoretical case for semiparametric methods, but when semiparametric metods are difficult, inefficient, and do not estimate the quantities of interest, we must look to parametric methods.

Criticisms of assumptions like (5.1) have more bite if we have unbalanced panel data. Then, we would have to specify a different conditional distribution of $\mathbf{c}_i$ for each configuration of missing data. Currently, the only solution to this problem is the one described in Section 3.2: if sample selection is exogenous conditional on $(\mathbf{y}_{i0},\mathbf{z}_i)$, we can use the balanced subpanel. As discussed in Sections 1 and 3.2, semiparametric methods also must exclude data when the panel is not balanced.

5.2. Dynamic Corner Solution Models

For the Tobit model with functions of the lagged dependent variable, the density in Assumption A.2 is

$$f_t(y_t|\mathbf{z}_t,y_{t-1},c,\boldsymbol{\theta}) = 1 - \Phi[(\mathbf{z}_t\boldsymbol{\gamma} + \mathbf{g}(y_{t-1})\boldsymbol{\rho} + c)/\sigma_u], \; y_t = 0$$

$$= (1/\sigma_u)\phi[(y_t - \mathbf{z}_t\boldsymbol{\gamma} - \mathbf{g}(y_{t-1})\boldsymbol{\rho} - c)/\sigma_u], \; y_t > 0.$$

To implement the conditional MLE, we need to specify a density in Assumption A.3. Again, it is convenient for this to be normal, as in (5.1). 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$(0,\sigma_u^2)$ 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. We are thinking of cases where y_{it} is not the result of true data censoring (such as top coding) but rather is a corner solution response (such as labor supply, charitable contributions, amount of life insurance, and so on).

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

$$m(\mathbf{w}_t\boldsymbol{\beta} + c,\sigma_u^2) = \Phi[(\mathbf{w}_t\boldsymbol{\beta} + c)/\sigma_u](\mathbf{w}_t\boldsymbol{\beta} + c) + \sigma_u\phi[(\mathbf{w}_t\boldsymbol{\beta} + c)/\sigma_u], \qquad (5.15)$$

where $\mathbf{w}_t = (\mathbf{z}_t,\mathbf{g}_{t-1})$. A consistent estimator of $E(c_i)$ is $\hat{\alpha}_0 + \bar{\mathbf{r}}_0\hat{\boldsymbol{\alpha}}_1 + \bar{\mathbf{z}}\hat{\boldsymbol{\alpha}}_2$, where the estimates are the conditional MLEs and the overbars denote sample averages. Even better, we can estimate the average partial effects. As in

the probit case, it is convenient to rewrite the APE in terms of a_i:

$$\begin{aligned} E[m(\mathbf{w}_t\boldsymbol{\beta} + c_i,\sigma_u^2)] &= E[m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i,\sigma_u^2)] \\ &= E\{E[m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i,\sigma_u^2)|\mathbf{r}_{i0},\mathbf{z}_i]\}, \end{aligned} \tag{5.16}$$

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 $a_i \sim \text{Normal}(0,\sigma_a^2)$, the conditional expectation in (5.16) is obtained by integrating $m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i,\sigma_u^2)$ over a_i with respect to the $\text{Normal}(0,\sigma_a^2)$ distribution. Since $m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2 + a_i,\sigma_u^2)$ 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 $\text{Normal}(0,\sigma_u^2)$ distribution, it is easily seen that the conditional expectation in (5.16) is

$$m(\mathbf{w}_t\boldsymbol{\beta} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2,\sigma_a^2 + \sigma_u^2). \tag{5.17}$$

A consistent estimator of the expected value of (5.17) (with respect to the distribution of $(\mathbf{r}_{i0},\mathbf{z}_i)$) is simply

$$N^{-1}\sum_{i=1}^{N} m(\mathbf{w}_t\hat{\boldsymbol{\beta}} + \hat{\alpha}_0 + \mathbf{r}_{i0}\hat{\boldsymbol{\alpha}}_1 + \mathbf{z}_i\hat{\boldsymbol{\alpha}}_2,\hat{\sigma}_a^2 + \hat{\sigma}_u^2). \tag{5.18}$$

The same kind of argument can be used to estimate averaged partial effects conditional on y_{it} being positive, that is, on $E(y_{it}|y_{it} > 0,\mathbf{w}_{it} = \mathbf{w}_t,c_i = c)$.

Other corner solution responses can be handled in a similar manner. For example, suppose y_{it} is a fractional variable that can take on the values zero and one with positive probability (for example, fraction of pension assets in the stock market). Then we can define y_{it} in terms of the latent variable y_{it}^* introduced earlier. The practical issues are how the lagged dependent variable should appear and how the initial value y_{i0} should appear in the distribution for c_i.

The discussion of the merits and drawbacks of the conditional normality assumption for c_i given $(\mathbf{y}_{i0},\mathbf{z}_i)$ are essentially the same as in the probit case.

5.3. Dynamic Poisson Model

As in Section 2, we assume that y_{it} given $(y_{i,t-1},...,y_{i0},\mathbf{z}_i,c_i)$ has a Poisson distribution with mean given in (2.4). As in the previous cases, there exists a choice of the conditional density in Assumption A.3 that simplifies the analysis. Write

$$c_i = a_i\exp(\alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2), \tag{5.19}$$

where $\mathbf{r}_{i0}$ is a vector of functions of y_{i0}. Assume that that a_i is independent of $(\mathbf{z}_i,y_{i0})$ and $a_i \sim \text{Gamma}(\eta,\eta)$, which is analogous to Hausman, Hall, and Griliches (1984). (This implies the normalization restriction $E(a_i) = 1$.) Then, for each t, y_{it} given $(y_{i,t-1},...,y_{i0},\mathbf{z}_i,a_i)$ has a Poisson distribution with mean

$$a_i\exp(\mathbf{z}_{it}\boldsymbol{\delta} + \mathbf{g}_{i,t-1}\boldsymbol{\rho} + \alpha_0 + \mathbf{r}_{i0}\boldsymbol{\alpha}_1 + \mathbf{z}_i\boldsymbol{\alpha}_2), \tag{5.20}$$

where $\mathbf{r}_{i0}$ denotes a vector function of y_{i0}. Call the mean in (5.20) a_im_{it}. Then the density of $(y_{i1},...,y_{iT})$ given $(\mathbf{z}_i,y_{i0},a_i)$ is obtained, as usual, by the product rule:

$$\prod_{t=1}^{T}\exp(-a_im_{it})(a_im_{it})^{y_t}/y_t! = \left(\prod_{t=1}^{T}m_{it}^{y_t}/y_t!\right)\exp\left(-a_i\sum_{t=1}^{T}m_{it}\right)a_i^n, \tag{5.21}$$

where $n = y_1 + ... + y_T$. When we integrate out a_i with respect to the Gamma(η,η) density, we obtain a density which has the usual random effects Poisson form with Gamma(η,η) heterogeneity, as in Hausman, Hall, and Griliches (1984, equation (2.3)). The difference is that the explanatory

variables are $(\mathbf{z}_{it}, \mathbf{g}_{i,t-1}, \mathbf{r}_{i0}, \mathbf{z}_i)$. These are obviously not strictly exogenous due to the presence of $\mathbf{g}_{i,t-1}$. But the log likelihood for each i has the same form as if they are. This makes estimation especially convenient in a software package such as Stata, which estimates random effects Poisson models with Gamma heterogeneity. We could instead assume that a_i has a lognormal distribution with mean unity.

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 interesting 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

$$\begin{aligned} P(union_{it} = 1 | union_{i,t-1}, \ldots, union_{i0}, marr_{i1}, \ldots, marr_{iT}, c_i) \qquad (6.1) \\ = \Phi(\eta_t + \gamma_1 marr_t + \rho_1 union_{i,t-1} + c_i), \; t = 1, \ldots, T, \end{aligned}$$

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 (5.1), 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.

The first column in Table 1 contains the conditional maximum likelihood estimates. These were obtained simply by using the Stata® 7.0 "xtprobit" command, where a full set of time dummies, 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. (The coefficients on the year dummies are not reported.) Asymptotic standard errors are given in parentheses.

(Table 1 about here.)

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. It seems practically large, too (.884), although we hold off discussing magnitudes until we have estimated the partial effect on the response probability with the unobserved hetegerogeneity averaged out. 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 $union_0$ (1.499) is much larger than the coefficent on the lag, $union_{t-1}$.

Getting married is estimated to have a marginally significant effect on belonging to a union, with a t statistic of about 1.61. Recall that the variables $marr_1$, ..., $marr_7$ are included to allow for partial correlation between c_i and marital status in all time periods. Interestingly, 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 two 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 always include them in the model for unobserved heterogeneity in (5.1), which means we just list them as additional explanatory variables. The coefficient on *educ* is very 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 $union_0$, $marr_1$, ..., $marr_7$, *educ*, and *black*: $\hat{\sigma}_a = 1.086$ (the estimate of the conditional standard deviation of c_i), and it is statistically different from zero. This means that the unobserved effect $a_i = c_i - E(c_i|union_{i0}, marr_{i1}, ..., marr_{i7}, educ_i, black_i)$ accounts for about 54.1% of the unexplained variance of the composite error, $a_i + u_{it}$, where u_{it} has a conditional standard normal distribution.

As emphasized in Section 4, it is often important to obtain an estimated partial effects with respect to the lagged dependent variable (and perhaps other explanatory variables). Here, 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 the distribution of c_i using equation (5.14), and we compute the effect for married and single men separately. Specifically, Table 2 reports

$$N^{-1} \sum_{i=1}^{N} \Phi[(-1.800 - .0083 + .178\, marr_t + .884\, union_{t-1} + \hat{\alpha}_1 y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\alpha}}_2)/(1 + 1.248)^{1/2}],$$

for $union_{t-1} = 0$ or 1 and $marr_t = 0$ or 1, where -.0083 is the coefficient on the 1987 year dummy and $\hat{\sigma}_a^2 = 1.248$. The $\hat{\alpha}_j$ are reported in column one of Table 1.

(Table 2 about here.)

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 .415. For a married man not belonging to a union in

1986, the estimated probability is .227. The difference, .188, is an estimate of the state dependence of union membership. The magnitude for unmarried men, .176, is similar.

7. CONCLUSIONS

I have suggested a general method for handling the initial conditions problem in 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. Because we model the density of the unobserved effect conditional on the observed initial condition (and exogenous variables), this is not the same as treating the initial condition as fixed. Conditional MLE can be used and has its standard asymptotic properties as the cross section sample size increases.

The auxiliary conditional density can be modeled in a very flexible way, but perhaps the most important contribution of the paper is that it shows how to obtain remarkably simple estimators in dynamic probit, Tobit, and Poisson unobserved effects models for specific choices of the auxiliary density. We have considered the important problems of estimating the partial effects at the average value of the unobserved heterogeneity and the partial effects averaged across the distribution of the unobserved heterogeneity. The APEs are generally identified under Assumptions A.1, A.2, and A.3. For some leading cases, the APEs are easy to estimate; hopefully, the availability of simple estimates will make reporting them routine in empirical work, where the current focus is on parameter estimates.

Many issues can be studied in future research. For example, because we might choose the model for $D(\mathbf{c}_i|\mathbf{y}_{i0},\mathbf{z}_i)$ for convenience -- as with the examples in Section 5 -- it is important to know the consequences of misspecifying the density in Assumption A.3. Intuitively, as the size of the cross section increases, we can make the density $h(\mathbf{c}|\mathbf{y}_0,\mathbf{z};\boldsymbol{\delta})$ more and more flexible (as in the so-called seminonparametric literature). In all of the examples in Section 5, as N gets large we can easily let the conditional mean function $E(c_i|y_{i0},\mathbf{z}_i)$ be very flexible in $(y_{i0},\mathbf{z}_i)$ -- so, for example, we can add interactions and various powers. If we want to be flexible along other dimensions -- for example, in the probit and Tobit cases allowing $\text{Var}(c_i|y_{i0},\mathbf{z}_i)$ to be heteroskedastic -- computation becomes more of an issue.

Unless nonlinearities in the model are caused by true data censoring, any study to evaluate the impact of various choices in Assumption A.3 on the robustness of the estimators should focus on estimates of average partial effects. As is well known, it frequently makes no sense to compare parameter estimates across different nonlinear models. (An example is probit and logit, where the scale factors entering the partial effects differ by the multiple .625.)

The approach proposed in Section 3 can be modified when some of the explanatory variables fail the strict exogeneity requirement. When the $\mathbf{z}_{it}$ contain policy variables or individual-choice variables, these can respond to past movements in y_{it}, and this can invalidate Assumption A.1. (For example, if marital status is included in an employment probit, future marital status may depend on lagged employment status.) Wooldridge (2000) lays out a framework for handling models with feedback. Finally, the idea of specifying a conditional distribution for the unobserved effect given the initial

conditions should prove useful for analyzing dynamic unobserved effects models with attrition or sample selection. Wooldridge (1995) covers the case of linear models with strictly exogenous explanatory variables, but allowing for a lagged dependent variable in the structural equation is nontrivial.

REFERENCES

Ahn, S.C. and P. Schmidt (1995), "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics* 68, 5-27.

Anderson, T.W. and C. Hsiao (1982), "Formulation and Estimation of Dynamic Models Using Panel Data," *Journal of Econometrics* 18, 67-82.

Arellano, M. and S.R. Bond (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. and O. Bover (1995), "Another Look at the Instrumental Variables Estimation of Error-Component Models," *Journal of Econometrics* 68, 29-51.

Arellano, M. and R. Carrasco (2002), "Binary Choice Panel Data Models with Predetermined Variables," CEMFI Working Paper No. 9618.

Bhargava, A. and J.D. Sargan (1983), "Estimating Dynamic Random Effects Models From Panel Data Covering Short Time Periods," *Econometrica* 51, 1635-1659.

Blundell, R. and S. Bond (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics* 87, 115-143.

Blundell, R.W. and R.J. Smith (1991), "Initial Conditions and Efficient Estimation in Dynamic Panel Data Models," *Annales d'Economie et de Statistique* 20/21, 109-123.

Chamberlain, G. (1978), "On the Use of Panel Data," unpublished manuscript.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," *Review of Economic Studies* 47, 225-238.

Chamberlain, G. (1984), "Panel Data," in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics*, Volume 2. Amsterdam: North Holland, 1247-1318.

Chamberlain, G. (1992), "Comment: Sequential Moment Restrictions in Panel Data," *Journal of Business and Economic Statistics* 10, 20-26.

Erdem, T. and B. Sun (2001), "Testing for Choice Dynamics in Panel Data," forthcoming, *Journal of Business and Economic Statistics*.

Gallant, A.R. and D.W. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica* 55, 363-390.

Geweke, J. and M. Keane (1999), "Mixture of Normals Probit Models," in C. Hsaio, K. Lahiri, L.-F. Lee, and M.H. Pesaran (eds.), *Analysis of Panels and Limited Dependent Variable Models*. Cambridge: Cambridge University Press,

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, J.A., B.H. Hall, and Z. Griliches (1984), "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica* 52, 909-938.

Heckman, J.J. (1981), "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," in C.F. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, 179-195. MIT Press: Cambridge, MA.

Honoré, B.E. (1993), "Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables," *Journal of Econometrics* 59, 35-61.

Honoré, B.E., E. Kyriazidou (2000), "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica* 68, 839-874.

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

Newey, W.K. and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in R.F. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Volume 4. Amsterdam: North Holland, 2111-2245.

Vella, F. and M. Verbeek (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, J.M. (1995), "Selection Corrections for Panel Data Models Under Conditional Mean Independence Assumptions," *Journal of Econometrics* 68, 115-132.

Wooldridge, J.M. (1997), "Multiplicative Panel Data Models without the Strict Exogeneity Assumption," *Econometric Theory* 13, 667-678.

Wooldridge, J.M. (2000), "A Framework for Estimating Dynamic, Unobserved Effects Panel Data Models with Possible Feedback to Future Explanatory Variables," *Economics Letters* 68, 245-250.

Wooldridge, J.M. (2001), "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples," *Econometric Theory* 17, 451-470.

Wooldridge, J.M. (2002a), *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, J.M. (2002b), "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification," *Portuguese Economic Journal*1, 117-139.

Table 1

Dependent Variable: $union_t$		
Explanatory Variable	(1)	(2)
$marr_t$	.179 (.111)	.180 (.111)
$union_{t-1}$	.884 (.094)	.895 (.094)
$union_0$	1.499 (.165)	1.461 (0.172)
$marr_1$	.155 (.210)	.145 (.212)
$marr_2$	-.194 (.210)	-.182 (.208)
$marr_3$	-.087 (.231)	-.090 (.236)
$marr_4$	.215 (.244)	.265 (.247)
$marr_5$	.00004 (.00007)	.00003 (.00007)
$marr_6$	.292 (.253)	.273 (.255)
$marr_7$	-.429 (.215)	-.390 (.220)
educ	———	-.013 (.036)
black	———	.526 (.192)
constant	-1.800 (0.148)	-1.731 (0.445)
$\hat{\sigma}_a$	1.117 (0.097)	1.086 (0.093)
Log Likelihood Value	-1,288.28	-1,284.40

Table 2

Estimated Probability of Being in a Union, 1987		
	In Union, 1986	Not in Union, 1986
Married, 1987	.415	.227
Not Married, 1987	.373	.197