An Exponential Class of Dynamic Binary Choice Panel Data Models with Fixed Effects*

Majid M. Al-Sadoon
Universitat Pompeu Fabra and Barcelona GSE

Tong Li
Vanderbilt University

M. Hashem Pesaran
Department of Economics & USC Dornsife INET, University of Southern California, USA
and Trinity College, Cambridge, UK

August 28, 2016

Abstract

This paper proposes an exponential class of dynamic binary choice panel data models for the analysis of short T(time dimension) large N (cross section dimension) panel data sets that allows for unobserved heterogeneity (fixed effects) to be arbitrarily correlated with the covariates. The paper derives moment conditions that are invariant to the fixed effects which are then used to identify and estimate the parameters of the model. Accordingly, GMM estimators are proposed that are consistent and asymptotically normally distributed at the root-N rate. We also study the conditional likelihood approach, and show that under exponential specification it can identify the effect of state dependence but not the effects of other covariates. Monte Carlo experiments show satisfactory finite sample performance for the proposed estimators, and investigate their robustness to miss-specification.

JEL Classification: C23, C25

Keywords: Dynamic Discrete Choice; Fixed Effects; Panel Data; GMM; CMLE

[*] We are grateful to the Co-Editor and two referees for their constructive comments that have greatly improved the paper. Al-Sadoon acknowledges financial assistance from the Spanish Ministry of Economy and Competitiveness (project ECO2012-33247). Li gratefully acknowledges support from National Science Foundation (SES-0922109). Pesaran acknowledges support from ESRC grant no. ES/I031626/1.

1 Introduction

This paper considers estimation and inference in the case of short T (time dimension) and large N (cross section dimension) dynamic binary choice panel data models with unobserved heterogeneity that is allowed to be arbitrarily correlated with the covariates. Such models are of particular interest in many applications since they can be used to distinguish between the presence of state dependence and the effect of unobserved heterogeneity, as discussed in Heckman (1981a and 1981b). These models are usually specified in terms of the distribution of the dependent variable conditional on the lagged dependent variable, a set of (possibly time-varying) covariates, and an individual specific term that represents unobserved heterogeneity.

As is well known, for dynamic panel data models with unobserved effects, an important issue is the treatment of the initial observations. While in some cases initial observations can be viewed as fixed constants if the actual start of the dynamic process coincides with the first time period in the data, in general, if the dynamic model under consideration has been in effect before the first period of the sample under consideration, there is an intrinsic and complex relationship between the initial observations and the unobserved heterogeneity. Therefore, in general, it is important that initial observations are allowed to depend on the unobserved individual effects in a model-coherent manner, in the sense that the dynamic model assumed to generate the observations is compatible with the processes assumed for the initial observations.

For linear models with an additive unobserved effect, appropriate transformations such as differencing have been used to eliminate the unobserved effect, and GMM type estimators have been proposed to estimate the transformed model. For example, see Anderson and Hsiao (1982), Arellano and Bover (1995), Arellano and Carrasco (2003), Ahn and Schmidt (1995), Blundell and Bond (1998), Hahn (1999), and Hsiao, Pesaran, and Tahmiscioglu (2002), Hayakawa and Pesaran (2015), and among others surveyed in Arellano and Honoré (2001), Hsiao (2014), and Pesaran (2015, Ch. 27). However, for nonlinear panel data models in general and binary choice models in particular the treatment becomes more complicated. When the unobserved effect is assumed to be a random effect in that it is not correlated with the strictly exogenous variables, Heckman (1981b) suggests to approximate the conditional distribution of the initial values given the exogenous variables and the unobserved individual effects, and use maximum likelihood to estimate the model parameters. Alternatively, Wooldridge (2005) proposes to specify an auxiliary distribution for the unobserved individual effects conditional on the initial values and the exogenous variables leading to a simple conditional maximum likelihood estimation. Both methods, while useful in addressing the initial value problem, can at best be viewed as approximations of the true (conditional) distribution of the initial values, and the unobserved heterogeneity, respectively. As discussed in Honoré (2002), because of the complicated relationship that exist between initial values, unobserved heterogeneity and the exogenous variables, it is almost unavoidable that modeling these two conditional distributions could be inconsistent with the original model specification, and there could even be some potential incoherency problems in the case of unbalanced panel data models.

Analysis of dynamic nonlinear panel data models with fixed effects, on the other hand, is further complicated by the so-called incidental parameters problem, in addition to the initial value problem. The incidental parameters problem arises because the number of unobserved effects increases with N, the number of the individuals in the panel. As a result, the maximum likelihood estimator of the structural parameters, while consistent when both N and T tend to infinity, it is inconsistent with large N and fixed T.

There are two approaches to dealing with the short T problem in non-linear dynamic panels. One strand of the literature has focussed on developing modified maximum likelihood estimators that attain bias reductions when T is fixed. Examples include the papers by Arellano (2003) for static binary choice panel data models, and by Carro (2007), Bartolucci, Bellio, Salvan, and Sartori (2012), and Lee and Phillips (2015) for dynamic binary choice panel data models. This approach still requires T to be relatively large to attain significant bias reductions, as demonstrated in a number of Monte Carlo studies reported in the literature, even in the simplest case where the initial values are taken to be fixed constants.

Another approach in the literature is to eliminate the fixed effects as done in the linear models. This approach, solves the incidental parameters problem, although the initial values problem remains. So far, however, there are only a few papers following this approach. Honoré and Kyriazidou (2000) consider the dynamic logit model and derive a set of conditions under which the parameters of the model are identified. They also propose consistent estimators of the model based on the identification results, albeit the rate of convergence of the estimators is slower than the usual $\sqrt{N}$-rate. In more recent papers, Bartolucci and Nigro (2010, 2012) consider a version of the quadratic exponential model that closely mimics the dynamic logit model and propose a conditional maximum likelihood estimator conditioning on the sufficient statistics for the individual specific effects. However, with this specification the strict exogeneity assumption usually made on the covariates in the standard dynamic panel data models is not met.[1] Also there could be some potential incoherency problems arising from the separate model specification for the last period from the other periods if one conducts sequential estimation, or if one deals with an unbalanced panel. Arellano and Bonhomme (2011) provide a review of recent developments in the econometric analysis of nonlinear panel data models.

In this paper we consider a dynamic binary choice panel data model with fixed effects, where the error term follows an exponential distribution. We show that the model can be written as an inhomogeneous Markov chain and, using a result from Pesaran and Timmermann (2009), we convert the non-linear model into a linear first-order autoregressive process in the indicator variables, and derive moment conditions that are free from incidental parameters, and allow us to identify the structural parameters (the state dependent parameter as well as the coefficients of the exogenous covariates). Based on these moment conditions we propose GMM estimators that are consistent and asymptotically normally distributed at the $\sqrt{N}$-rate. Compared with the existing approaches, our method identifies all the parameters of the model and yields simple-to-implement estimators that have standard asymptotic properties. It turns out that the exponential specification we entertain, as well as the moment conditions we employ, are variants of those proposed in Wooldridge (1997).[2]

In addition to the GMM estimators, since the conditional maximum likelihood approach has been adopted in the literature in the case of the logistic distribution or the quadratic exponential distribution in order to eliminate the fixed effects, we also study the conditional likelihood approach, which can only identify the effect of state dependence under exponentially distributed errors. Since our GMM estimators are general and simple to implement, we study their finite sample performance through a comprehensive simulation study and the results indicate that our estimators perform quite well in relatively small size samples.

Given that we are the first to propose the use of an exponential error distribution in a binary choice setting, it is important that this choice is motivated and further discussed. The first point to

[1] Strict exogeneity typically allows us to specify the likelihood of y_{it} conditional on c_i, $\mathbf{x}_{it}$ and y_{it-1}. But in the Bartolucci and Nigro (2010) specification, all periods observations of $\mathbf{x}_{it}$ must be taken into account. On the strict exogeneity assumption and the other approaches in the literature, see Wooldridge (2002) for a survey.

[2] See Remark 1 for more details.

bear in mind is that in the case of fixed effects binary choice models, the choice of the distribution is in fact secondary; fixed effects (which are free of any restrictions) can be used to match probability outcomes based on exponential and any other error distribution, including the logistic ones used in the literature. In the case of models without any covariates ($\mathbf{x}_{it}$'s), the match can be performed perfectly for all distributional specifications. When the model contains covariates, the match between the exponential and other distributions, including the logistic, can be done for specific values of $\mathbf{x}_{it}$, (at some t) or at the mean of $\mathbf{x}_{it}$, namely at $\overline{\mathbf{x}}_i$, as we demonstrate in Section 4.3. Therefore, at least in a binary choice setting the choice of the distribution is more a matter of analytical and estimation convenience. Moreover, since in analyzing a nonlinear model such as a binary choice model, a key quantity of interest is the average partial effect (APE), we will investigate through Monte Carlo simulations how well the APEs are estimated with the exponential model if the true underlying model happen to be logistic. Our results show that the exponential model yields sensible estimates for the APEs even with a misspecified error distribution.

The rest of the paper is organized as follows. Section 2 sets out the model. Section 3 considers the pure dynamic case without any covariates. Section 4 generalizes and extends the analysis of Section 3 and allows for time-varying covariates. Section 5 presents the Monte Carlo results, and Section 6 provides some concluding remarks. Technical proofs are provided in an appendix.

2 A General Dynamic Binary Choice Model and its Markov Chain Representation

Suppose that y_{it} takes the values of *zero* and *unity*, for $i = 1, 2, ..., N$, and $t = 1, 2, ..., T$, and $\mathbf{x}_{it}$ is a $k \times 1$ vector of strictly exogenous, time-varying regressors; common time-varying regressors, such as a time dummy, can also be included in $\mathbf{x}_{it}$. The standard dynamic binary panel data model with fixed effects assumes that

$$\begin{aligned} y_{it} &= I\left(y_{it}^* \geq 0\right), \\ y_{it}^* &= \rho y_{i,t-1} + \boldsymbol{\beta}'\mathbf{x}_{it} + c_i + u_{it}. \end{aligned} \tag{1}$$

where y_{it}^* is a latent variable that is not observed by the econometrician, $I(A)$ is an indicator that takes the value of 1 if A holds and 0 otherwise, u_{it} is the random error term assumed to be identically, independently distributed (i.i.d.) with mean zeros, and c_i represents the individual unobserved effect that can be arbitrarily correlated with $\mathbf{x}_{it}$ and $y_{i,t-1}$. We suppose that T is fixed and N is sufficiently large. We are interested in ρ, the state dependence parameter, and $\boldsymbol{\beta}$, the coefficients of the $k \times 1$ vector of the covariates, $\mathbf{x}_{it}$, where k is fixed. We refer to ρ and $\boldsymbol{\beta}$ as the *structural parameters*, and refer to $\{c_i,\ i = 1, 2, \ldots, N\}$, as the *incidental parameters*.

Denote the distribution of $-u_{it}$ by $F(\cdot)$. Then we have

$$\begin{aligned} \Pr(y_{it} &= 1 \left| y_{1,t-1}, y_{2,t-1}, ..., y_{N,t-1}; c_1, c_2, ..., c_N; \mathbf{x}_{1t}, \mathbf{x}_{2t}, ..., \mathbf{x}_{Nt} \right.) \\ &= \Pr(y_{it} = 1 \left| y_{i,t-1}, c_i, \mathbf{x}_{it} \right.) = F(\rho y_{i,t-1} + \boldsymbol{\beta}' \mathbf{x}_{it} + c_i), \qquad (2) \end{aligned}$$

where the first equation follows from the strict exogeneity assumption on $\mathbf{x}_{it}$. The commonly used probit or logit models correspond to $F(\cdot)$ being either the standard normal distribution or the logistic distribution, respectively.

The model can also be characterized as an inhomogeneous Markov chain with transition probabilities

$$\begin{array}{ccc} & y_{it} = & 0 & 1 \\ y_{i,t-1} = & 0 & 1 - F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i) & F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i) \\ & 1 & 1 - F(\rho + \boldsymbol{\beta}'\mathbf{x}_{it} + c_i) & F(\rho + \boldsymbol{\beta}'\mathbf{x}_{it} + c_i) \end{array}$$

The distribution of y_{i1} conditional on c_i and $\mathbf{x}_{i1}$ is complicated and in general depends on the past (unknown) values of $\mathbf{x}_{it}$ for $t \leq 1$. For given values of c_i and $\mathbf{x}_{it}$, unconditional probability of $y_{it} = 1$, is given by $\pi_{it} = \Pr\left(y_{it} = 1 \left| c_i, \mathbf{x}_{it}, \mathbf{x}_{i,t-1}, ...\mathbf{x}_{i1}, \mathbf{x}_{i0}, \mathbf{x}_{i,-1}, ... \right.\right)$. Then from the structure of the Markov chain we have

$$\pi_{it} = F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i + \rho)\pi_{i,t-1} + F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i)(1 - \pi_{i,t-1}),$$

or

$$\pi_{it} = \lambda_{it}\pi_{i,t-1} + F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i), \qquad (3)$$

where

$$\lambda_{it} = F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i + \rho) - F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i).$$

The above difference equation has a stable solution if $|\lambda_{it}| < 1$. To avoid absorbing states we assume that $|c_i| < K < \infty$, $\left|\boldsymbol{\beta}'\mathbf{x}_{it}\right| < K < \infty$, and $|\rho| < K < \infty$, and then note that $0 < F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i + \rho) - F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i) < 1$, if $\rho > 0$, and $0 < F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i) - F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i + \rho) < 1$, if $\rho < 0$.

Recall that $F(z)$ is a non-decreasing positive function of z. Therefore, the distribution of the initial observation, π_{i1}, converges to a well defined limit on $(0,1)$. In general, the expression for π_{i1} is a complicated function of c_i, ρ, and all values of $\boldsymbol{\beta}'\mathbf{x}_{i,\tau}$, for $\tau \leq 1$. An explicit expression for π_{i1} can be found when $\boldsymbol{\beta} = \mathbf{0}$. However, the GMM estimators that we develop in this paper do not require modelling the initial conditions, so long π_{i1} does not de-generate to 0 or 1, which is satisfied under the bounding conditions given above.

3 The Case of $\boldsymbol{\beta} = \mathbf{0}$

3.1 The Likelihood Function

In the case where $\boldsymbol{\beta} = \mathbf{0}$, the Markov chain has a *time-invariant stationary distribution*

$$\Pr\left(y_{it} = 1 \left| c_i \right.\right) = \frac{F(c_i)}{1 - F(c_i + \rho) + F(c_i)} = \pi_i^*, \tag{4}$$

$$\Pr\left(y_{it} = 0 \left| c_i \right.\right) = \frac{1 - F(c_i + \rho)}{1 - F(c_i + \rho) + F(c_i)} = 1 - \pi_i^*. \tag{5}$$

We restrict ρ and c_i so that c_i and $c_i + \rho$ lie in the domain of $F(\cdot)$ and the above probabilities are well defined. Note that unlike in the linear case, this does not necessarily restrict ρ to be bounded above by 1. It is only required that c_i and ρ are bounded.

The joint probability distribution of $c_i, y_{i1}, y_{i2}, ..., y_{iT}$ can now be derived using the familiar decomposition

$$\Pr\left(c_i, y_{i1}, y_{i2}, ..., y_{iT}\right) = \Pr(c_i) \Pr(y_{i1} \left| c_i \right.) \Pr(y_{i2} \left| y_{i1}, c_i \right.) .... \Pr(y_{iT} \left| y_{i,T-1}, c_i \right.).$$

Consider now the observations y_{it} for $t = 1, 2, ..., T$, and note that, under stationarity, the likelihood function for the i^{th} unit at time $t = 1$ is given by

$$\Pr(y_{i1} \left| c_i, \rho \right.) = (\pi_i^*)^{y_{i1}} (1 - \pi_i^*)^{1 - y_{i1}}, \tag{6}$$

and for time $t = 2, 3, ..., T$, by

$$\begin{aligned} & \Pr(y_{it} \left| y_{i,t-1}, c_i, \rho \right.) & (7) \\ = \; & [F(c_i + \rho)]^{y_{it} y_{i,t-1}} [1 - F(c_i + \rho)]^{(1 - y_{it}) y_{i,t-1}} [F(c_i)]^{y_{it}(1 - y_{i,t-1})} [1 - F(c_i)]^{(1 - y_{it})(1 - y_{i,t-1})}. \end{aligned}$$

Setting $\mathbf{Y}=(y_{it},\, i=1,...,N; t=1,2,...,T)$, the log likelihood function for the panel (assuming independence across i) is given by

$$\begin{aligned} l(\rho\,|\mathbf{Y},\mathbf{c}\,) &= \sum_{i=1}^{N}\left[y_{i1}\ln(\pi_i^*)+(1-y_{i1})\ln(1-\pi_i^*)\right]+ \\ & \sum_{i=1}^{N}\sum_{t=2}^{T}y_{it}y_{i,t-1}\ln\left[F(c_i+\rho)\right]+\sum_{i=1}^{N}\sum_{t=2}^{T}(1-y_{it})y_{i,t-1}\ln\left[1-F(c_i+\rho)\right]+ \\ & \sum_{i=1}^{N}\sum_{t=2}^{T}y_{it}(1-y_{i,t-1})\ln\left[F(c_i)\right]+\sum_{i=1}^{N}\sum_{t=2}^{T}(1-y_{it})(1-y_{i,t-1})\ln\left[1-F(c_i)\right]. \end{aligned}$$

It is clear that there is an incidental parameter problem here that cannot be resolved without a specification of $\Pr(c_i)$. This can be accomplished by specifying a distribution in terms of the observables. Note, however, that $\Pr(c_i)$ can be specified independently of the initial value, y_{i1}, or the other observations. The assumption that c_i are independent across i can also be relaxed to allow for simple patterns of cross-sectional dependence across i (i.e. using more general specifications of $\Pr(\mathbf{c})$) although we do not pursue this here.

3.2 Exponential Dynamic Binary Choice Models

The literature on estimation of binary choice panel data models with fixed effects has focussed on a logit specification for $F(\cdot)$. In this paper we consider an alternative specification. We consider first the case where $\boldsymbol{\beta}=\mathbf{0}$ and equations (4) and (5) hold, and focus on consistent estimation of ρ. Pesaran and Timmermann (2009) show that a Markov chain can be written as a vector autoregressive (VAR) model in the indicator variables. In our context it can be easily established that the implied error term, ε_{it}, defined by

$$\varepsilon_{it}=y_{it}-F(c_i)-\left[F(c_i+\rho)-F(c_i)\right]y_{i,t-1},$$

is a martingale difference process with respect to $y_{i,t-1}, y_{i,t-2}, ...$, namely $E\left(\varepsilon_{it}\,|y_{i,t-1},y_{i,t-2},..\right)=0$. This result can be established explicitly by noting that for each i and t, ε_{it} is a discrete random variable that takes only 4 distinct values, namely $-F(c_i)$, $1-F(c_i)$, $1-F(c_i+\rho)$, and $-F(c_i+\rho)$, with probabilities given by the Markov chain.

The above representation of the dynamic binary choice model suggests the following linear binary AR(1) regression with reduced form parameters that are non-linear functions of the parameters of the underlying model:

$$y_{it} = F(c_i) + \left[F(c_i + \rho) - F(c_i)\right] y_{i,t-1} + \varepsilon_{it}. \tag{8}$$

This representation holds for all choices of $F(\cdot)$, but the fixed effect, c_i, is not readily separated from ρ using equation (8) without further assumptions. One such assumption considers whether it is possible to factorize $F(c_i + \rho) - F(c_i)$ into a product form such as $G(\rho)H(c_i)$, since this would allow us to isolate $F(c_i)$ from the structural parameter, ρ. Such a factorization is indeed possible when $F(z) = 1 - \exp(-z)$ as it satisfies

$$F(c_i + \rho) - F(c_i) = \exp(-c_i)\left[1 - \exp(-\rho)\right]. \tag{9}$$

In Appendix 7.1 we prove that the exponential is the only non–constant, differentiable, distribution function that satisfies this condition.[3]

Consistent estimation of ρ can now be achieved using the conditional maximum likelihood or the GMM methods.

3.3 Conditional Maximum Likelihood Estimation

Building on an early work by Cox (1958), Chamberlain (1985) shows that it is possible to estimate ρ consistently using a conditional maximum likelihood estimator (CMLE) approach if $F(\cdot)$ is logistic, $\boldsymbol{\beta} = \mathbf{0}$ and $T \geq 4$.[4] Honoré and Kyriazidou (2000) extend this analysis to the case where $\boldsymbol{\beta} \neq \mathbf{0}$, under certain restrictions on the distribution of the covariates, $\mathbf{x}_{it}$, over time. In this sub-section we show similar results hold if $F(\cdot)$ is exponential, $\boldsymbol{\beta} = \mathbf{0}$ and $T \geq 3$.

Using (6) and (7) the likelihood function (conditional on c_i) for the i^{th} unit can be written as

$$\begin{aligned}\left[1 - F(c_i + \rho) + F(c_i)\right] \Pr\left(\mathbf{y}_{iT} \left| c_i, \rho\right.\right) &= \left[F(c_i + \rho)\right]^{\sum_{t=2}^{T} y_{it} y_{i,t-1}} \left[1 - F(c_i + \rho)\right]^{1 - y_{i1} + \sum_{t=2}^{T}(1 - y_{it}) y_{i,t-1}} \\ &\quad \times \left[F(c_i)\right]^{y_{i1} + \sum_{t=2}^{T} y_{it}(1 - y_{i,t-1})} \left[1 - F(c_i)\right]^{\sum_{t=2}^{T}(1 - y_{it})(1 - y_{i,t-1})}.\end{aligned}$$

[3] To be more precise, we prove that the general form of a function F that satisfies the factorization is given by $F(z) = 1 - C\exp(-Dz)$, for C and $D > 0$. Since these two parameters are not identifiable, we set them both equal to 1. Similar rescaling and normalization is also used for the standard logit and probit models.

[4] See Chamberlain (2010) for identification in a two-period case and Magnac (2004) for more general identification results with the conditional likelihood approach, and also Magnac (2001) for an empirical application.

Let $s_{iT} = \sum_{t=1}^{T} y_{it}$ and $p_{iT} = \sum_{t=2}^{T} y_{it}y_{i,t-1}$, and write the above likelihood function as

$$\begin{aligned} \Pr\left(\mathbf{y}_{iT}\left|c_i, \rho\right.\right) &= \Pr\left(s_{iT}, p_{iT}, y_{i1}, y_{iT}\left|c_i, \rho\right.\right) \\ &= \frac{\begin{array}{c}[F(c_i+\rho)]^{p_{iT}}\left[1-F(c_i+\rho)\right]^{1-y_{i1}-y_{iT}+s_{iT}-p_{iT}} \\ \left[F(c_i)\right]^{s_{iT}-p_{iT}}\left[1-F(c_i)\right]^{(T-1)+y_{i1}+y_{iT}-2s_{iT}+p_{iT}}\end{array}}{\left[1-F(c_i+\rho)+F(c_i)\right]}. \end{aligned}$$

It is clear that s_{iT}, p_{iT}, y_{i1}, and y_{iT} are minimal sufficient statistics for c_i and ρ. Following Andersen (1970), we consider the likelihood function of ρ conditional on given values of $s_{iT} = s^0$ and $p_{iT} = p^0$ for all i. Let $\mathcal{B}_{iT}(s^0, p^0)$ be the set of all sequences $y_{i1}, y_{i2}, ..., y_{iT}$ that satisfy $\sum_{t=1}^{T} y_{it} = s^0$ and $\sum_{t=2}^{T} y_{it}y_{i,t-1} = p^0$, for $s^0 = 1, ..., T-1$ and $p^0 = 0, 1, .., T-1$ ($s^0 > p^0$). There is no point considering the values of $s^0 = 0, T$, since for these values the conditional likelihood function does not depend on ρ.

In general we have

$$\Pr\left(y_{i1}, y_{iT}\left|s_{iT} = s^0, p_{iT} = p^0, c_i, \rho\right.\right) = \frac{\Pr\left(s_{iT} = s^0, p_{iT} = p^0, y_{i1}, y_{iT}\left|c_i, \rho\right.\right)}{\Pr\left(s_{iT} = s^0, p_{iT} = p^0\left|c_i, \rho\right.\right)},$$

where

$$\Pr\left(s_{iT} = s^0, p_{iT} = p^0, y_{i1}, y_{iT}\left|c_i, \rho\right.\right) = \frac{A_i(s^0, p^0)\left[1-F(c_i)\right]^{y_{i1}+y_{iT}}\left[1-F(c_i+\rho)\right]^{-y_{i1}-y_{iT}}}{\left[1-F(c_i+\rho)+F(c_i)\right]},$$

and

$$\Pr\left(s_{iT} = s^0, p_{iT} = p^0\left|c_i, \rho\right.\right) = \frac{A_i(s^0, p^0)\sum_{y_{i1}, y_{iT} \in \mathcal{B}_{iT}(s^0, p^0)}\left[1-F(c_i)\right]^{y_{i1}+y_{iT}}\left[1-F(c_i+\rho)\right]^{-y_{i1}-y_{iT}}}{\left[1-F(c_i+\rho)+F(c_i)\right]},$$

in which

$$A_i(s^0, p^0) = \left[F(c_i+\rho)\right]^{p^0}\left[F(c_i)\right]^{1+s^0-p^0}\left[1-F(c_i)\right]^{(T-1)-2s^0+p^0}\left[1-F(c_i+\rho)\right]^{1+s^0-p^0.}$$

Therefore

$$\Pr\left(y_{i1}, y_{iT}\left|s_{iT} = s^0, p_{iT} = p^0, c_i, \rho\right.\right) = \frac{\left[1-F(c_i)\right]^{y_{i1}+y_{iT}}\left[1-F(c_i+\rho)\right]^{-y_{i1}-y_{iT}}}{\sum_{y_{i1}, y_{iT} \in \mathcal{B}_{iT}(s^0, p^0)}\left[1-F(c_i)\right]^{y_{i1}+y_{iT}}\left[1-F(c_i+\rho)\right]^{-y_{i1}-y_{iT}}}.$$

It is clear that for a general specification of $F(\cdot)$, the conditional distribution of y_{i1} and y_{iT} still depends on the incidental parameters, c_i. But in the case of the exponential distribution we have

$$\Pr\left(y_{i1}, y_{iT}\left|s_{iT} = s^0, p_{iT} = p^0, c_i, \rho\right.\right) = \frac{\exp\left[\rho(y_{i1}+y_{iT})\right]}{\sum_{y_{i1}, y_{iT} \in \mathcal{B}_{iT}(s^0, p^0)}\exp\left[\rho(y_{i1}+y_{iT})\right]},$$

which does not depend on $c_i^{\prime}s$.

The conditional likelihood for the cross section observations $i = 1, 2, .., N$ is now given by

$$L_c(\rho) = \prod_{i=1}^{N} \prod_{p^0=0}^{T-2} \prod_{s^0=1}^{T-1} \frac{\exp\left[\rho(y_{i1} + y_{iT})\right]}{\sum_{y_{i1}, y_{iT} \in \mathcal{B}_{iT}(s^0, p^0)} \exp\left[\rho(y_{i1} + y_{iT})\right]}. \tag{10}$$

Not all the components of this conditional likelihood function will depend on ρ. For example, in the case where $T = 3$, which is derived in detail in the appendix, the only component that depends on ρ is for values of $s^0 = 1$ and $p^0 = 0$. When $T = 3$ we exclude observation where $s^0 = 3$ and $p^0 = 2$. The remaining values are $(s^0, p^0) = (2, 0)$ and $(s^0, p^0) = (2, 1)$. Under the former we must have $y_{i1} = 1, y_{i2} = 0$ and $y_{i3} = 1$ and

$$\frac{\exp\left[\rho(y_{i1} + y_{i3})\right]}{\sum_{y_{i1}, y_{i3} \in \mathcal{B}_{i3}(2,0)} \exp\left[\rho(y_{i1} + y_{i3})\right]} = 1.$$

Under $(s^0, p^0) = (2, 1)$ the only admissible sequences are (110) and (011), and we have

$$\frac{\exp\left[\rho(y_{i1} + y_{i3})\right]}{\sum_{y_{i1}, y_{i3} \in \mathcal{B}_{i3}(2,1)} \exp\left[\rho(y_{i1} + y_{i3})\right]} = \frac{\exp(\rho)}{2\exp(\rho)} = \frac{1}{2}.$$

The only set of observations for which the conditional likelihood depends on ρ is given by

$$\frac{\exp\left[\rho(y_{i1} + y_{i3})\right]}{\sum_{y_{i1}, y_{i3} \in \mathcal{B}_{i3}(1,0)} \exp\left[\rho(y_{i1} + y_{i3})\right]} = \begin{cases} \frac{\exp(\rho)}{2\exp(\rho)+1}, \text{for } (100) \\ \frac{1}{2\exp(\rho)+1}, \text{for } (010) \\ \frac{\exp(\rho)}{2\exp(\rho)+1}, \text{for } (001) \end{cases}$$

Hence, the conditional log-likelihood function for the case where $T = 3$ can be written as

$$\ell_c(\rho) = \rho \sum_{i=1}^{N} (y_{i1} + y_{i3}) I(s_{i3} = 1) I(p_{i3} = 0) - \log\left[2\exp(\rho) + 1\right] \sum_{i=1}^{N} I(s_{i3} = 1) I(p_{i3} = 0).$$

It is easily verified that this is the same as (25) obtained in the appendix. Following Andersen (1970), consistency and $\sqrt{N}$-asymptotic normality of the resulting conditional maximum likelihood estimator can be established.

3.4 GMM Estimation

Under the exponential distribution, the binary AR(1) model (8) can be written as

$$y_{it} = \alpha_i + (1 - \alpha_i)\gamma y_{i,t-1} + \varepsilon_{it}, \tag{11}$$

where $\alpha_i = 1 - \exp(-c_i)$, and $\gamma = 1 - \exp(-\rho)$. The stability of the above AR(1) is ensured for all values of c_i and ρ for which π_i^* defined by (4) strictly lies inside the range $(0, 1)$. Note that for the exponential distribution and using (9) we have

$$\begin{aligned} \pi_i^* &= \frac{F(c_i)}{1 - F(c_i + \rho) + F(c_i)} = \frac{1 - \exp(-c_i)}{1 - \exp(-c_i)\left[1 - \exp(-\rho)\right]} \qquad (12) \\ &= \frac{\alpha_i}{1 - (1 - \alpha_i)\gamma}, \end{aligned}$$

and condition $0 < \pi_i^* < 1$ implies that $(1 - \alpha_i)\gamma < 1$, $1 - \alpha_i = \exp(-c_i) > 0$, and $1 - \gamma > \exp(-\rho) > 0$. The latter two conditions are met for all bounded values of c_i and ρ. Further, since $F(c_i) = 1 - \exp(-c_i) = \alpha_i > 0$, then condition $(1 - \alpha_i)\gamma < 1$ must also be satisfied since $\pi_i^* > 0$.

The AR(1) formulation considerably simplifies the estimation problem, but it is still subject to the incidental parameter problem. First–differencing will not eliminate α_i, the incidental parameters either, since the coefficient of $y_{i,t-1}$ also depends on α_i. But, instead of first-differencing we can equate two solutions of α_i obtained for for two successive periods[5]

$$\alpha_i = \frac{y_{i,t} - \gamma y_{i,t-1}}{1 - \gamma y_{i,t-1}} - \frac{\varepsilon_{i,t}}{1 - \gamma y_{i,t-1}}, \text{ for } t,$$

and

$$\alpha_i = \frac{y_{i,t-1} - \gamma y_{i,t-2}}{1 - \gamma y_{i,t-2}} - \frac{\varepsilon_{i,t-1}}{1 - \gamma y_{i,t-2}}, \text{ for } t - 1.$$

Equating the above two solutions of α_i, now yields the following non-linear difference equation

$$y_{it} = \gamma y_{i,t-1} + \left(\frac{1 - \gamma y_{i,t-1}}{1 - \gamma y_{i,t-2}}\right)(y_{i,t-1} - \gamma y_{i,t-2}) + v_{it}, \qquad (13)$$

where

$$v_{it} = \varepsilon_{it} - \left(\frac{1 - \gamma y_{i,t-1}}{1 - \gamma y_{i,t-2}}\right)\varepsilon_{i,t-1}.$$

Unfortunately, v_{it} does not satisfy any obvious orthogonality condition with respect to the lags of y_{it}. For example, $E(v_{it} \,|y_{i,t-2}) = \gamma E\left(y_{i,t-1}\varepsilon_{i,t-1} \,|y_{i,t-2}\right) / \left(1 - \gamma y_{i,t-2}\right)$, which is not generally equal to zero due to the contemporaneous dependence of $y_{i,t-1}$ on $\varepsilon_{i,t-1}$. However, the alternative formulation

$$\begin{aligned} e_{it} &= \left(\frac{1 - \gamma y_{i,t-2}}{1 - \gamma y_{i,t-1}}\right)\varepsilon_{it} - \varepsilon_{i,t-1} = \frac{(y_{it} - \gamma y_{i,t-1})\left(1 - \gamma y_{i,t-2}\right)}{(1 - \gamma y_{i,t-1})} - (y_{i,t-1} - \gamma y_{i,t-2}), \qquad (14) \\ &= (1 - y_{i,t-1}) - (1 - y_{it})\left(\frac{1 - \gamma y_{i,t-2}}{1 - \gamma y_{i,t-1}}\right), \qquad (15) \end{aligned}$$

[5] Note that since $1 - \gamma > 0$, then $1 - \gamma y_{i,t-1} \neq 0$, noting that $y_{i,t-1}$ can only take the values of 0 and 1.

which is obtained by multiplying both sides of (13) by $(1-\gamma y_{i,t-2})\,/\,(1-\gamma y_{i,t-1})$, does satisfy usable orthogonality conditions. To see this, note that

$$E(e_{it}\,|y_{i,t-1},y_{i,t-2}\,)=\left(\frac{1-\gamma y_{i,t-2}}{1-\gamma y_{i,t-1}}\right)E(\varepsilon_{it}\,|y_{i,t-1},y_{i,t-2}\,)-E(\varepsilon_{i,t-1}\,|y_{i,t-1},y_{i,t-2}\,).$$

But $E(\varepsilon_{it}\,|y_{i,t-1},y_{i,t-2}\,)=0$ by the Markov property as established in Pesaran and Timmermann (2009). Hence $E(e_{it}\,|y_{i,t-1},y_{i,t-2}\,)=-E(\varepsilon_{i,t-1}\,|y_{i,t-1},y_{i,t-2}\,)$. Now by chain rule of conditional expectations

$$\begin{aligned}E(e_{it}\,|y_{i,t-2}\,) &= E\left[E(e_{it}\,|y_{i,t-1},y_{i,t-2}\,)\,|y_{i,t-2}\,\right]\\ &= -E\left[E(\varepsilon_{i,t-1}\,|y_{i,t-1},y_{i,t-2}\,)\,|y_{i,t-2}\,\right]\\ &= -E(\varepsilon_{it}\,|y_{i,t-2}\,)=0,\end{aligned}$$

as required. In fact we have, more generally,

$$E(e_{it}\,|y_{i,t-s}\,)=0,\text{ for } s=2,3,... \tag{16}$$

These moment conditions can be used to estimate γ by GMM using $y_{i,t-2},y_{i,t-3},...$, as well as the constant, as instruments, very much as when GMM is applied to the first-differenced version in the linear case.

Note that the constant (i.e. the sequence of 1's) should be used as an instrument with caution. It is easy to show that $E(e_{it})=0$ whenever $\gamma=0$ or $\gamma=\gamma_0$. Thus the constant instrument fails to uniquely pin down γ_0. However, the other instruments do not suffer from this anomaly. Therefore, there is no danger in using the constant as an instrument if it is augmented by one or more lagged values $y_{i,t-2},y_{i,t-3},\ldots$.

Remark 1 *Wooldridge (1997) considers multiplicative panel data models of the form* $\tau(y_{it},\boldsymbol{\lambda}_0)=\phi_i\mu(\mathbf{x}_{it},\boldsymbol{\beta}_0)u_{it}$*, and shows that with sequential moment conditions on* u_{it} *as specified in Chamberlain (1992), the transformation*

$$r_{it}(\boldsymbol{\theta})\equiv\tau(y_{it},\boldsymbol{\lambda})-[\mu(\mathbf{x}_{it},\boldsymbol{\beta}_0)/\mu(\mathbf{x}_{it+1},\boldsymbol{\beta}_0)]\tau(y_{it},\boldsymbol{\lambda}),t=1,...,T-1,$$

satisfies the conditional moment condition

$$E[r_{it}(\boldsymbol{\theta}_0)|\phi_i,\mathbf{x}_{i1},...\mathbf{x}_{it}]=0,t=1,...,T-1.$$

The sequential nature of the moment conditions allows $y_{i,t-1}$ to be included in $\mathbf{x}_{it}$. In our case, we can rewrite our model as 1− $y_{it} = \phi_i \mu(\mathbf{x}_{it}, \boldsymbol{\beta}_0) u_{it}$, where $\mu(\mathbf{x}_{it}, \boldsymbol{\beta}_0) = \exp(-\rho y_{i,t-1})$ and $\phi_i = \exp(-c_i)$. Noting that $\exp(\rho \Delta y_{it}) = (1 - \gamma y_{i,t-1})/(1 - \gamma \rho y_{it})$, *it can be shown that $r_{it}(\boldsymbol{\theta}_0)$ in our case is identical to $e_{i,t+1}$. As a result, the conditional moment conditions in (16) can also be derived following the set up by Wooldridge (1997).*

Notice that since $\rho = -\ln(1-\gamma)$, to estimate ρ consistently we must have $\gamma < 1$. Alternatively, one could consider the GMM estimation problem directly in terms of ρ, namely by considering the moment conditions in terms of

$$e_{it}(\rho) = \frac{[\Delta y_{it} + y_{i,t-1}\exp(-\rho)]\,[1 - y_{i,t-2} + y_{i,t-2}\exp(-\rho)]}{[1 - y_{i,t-1} + y_{i,t-1}\exp(-\rho)]} - ([\Delta y_{i,t-1} + y_{i,t-2}\exp(-\rho)]\,. \quad (17)$$

Let $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iT})'$ and let $m_k(\mathbf{y}_i, \gamma)$ be an enumeration of $e_{it}(\rho)$ for $2 \leq t \leq T$ and $e_{it}(\rho) y_{it-s}$ for $2 \leq s \leq t \leq T$.

$$E\left[m_k(\mathbf{y}_i, \gamma)\right] = 0,\ k = 1, 2, ..., (T+1)(T-2)/2.$$

When $T = 3$, there are two moment conditions:[6]

$$\begin{aligned} E\left[m_1(\mathbf{y}_i, \gamma)\right] &= E\left(e_{i3}\right) = E\left[\frac{(y_{i3} - \gamma y_{i2})\,(1 - \gamma y_{i1})}{(1 - \gamma y_{i2})} - (y_{i2} - \gamma y_{i1})\right] = 0, \\ E\left[m_2(\mathbf{y}_i, \gamma)\right] &= E\left(y_{i1} e_{i3}\right) = E\left\{y_{i1}\left[\frac{(y_{i3} - \gamma y_{i2})\,(1 - \gamma y_{i1})}{(1 - \gamma y_{i2})} - (y_{i2} - \gamma y_{i1})\right]\right\} = 0. \end{aligned}$$

For $T > 3$, further moment conditions can be considered. Let $\mathbf{m}(\mathbf{y}_i, \gamma) = (m_1(\mathbf{y}_i, \gamma), m_2(\mathbf{y}_i, \gamma), ..., m_K(\mathbf{y}_i, \gamma))'$, and write the $K = (T+1)(T-2)/2$ moment conditions as $E\left[\mathbf{m}(\mathbf{y}_i, \gamma)\right] = 0$. Using the familiar results on GMM estimation we have

$$\hat{\gamma}_{GMM} = \arg\min_{\gamma}\left[\mathbf{M}_N'(\gamma)\mathbf{A}_N'\mathbf{A}_N\mathbf{M}_N(\gamma)\right],$$

where

$$\mathbf{M}_N(\gamma) = N^{-1}\sum_{i=1}^{N}\mathbf{m}(\mathbf{y}_i, \gamma),$$

and $\mathbf{A}_N$ is a $1 \times K$ weight vector. An optimal choice for $lim_{N\to\infty}\mathbf{A}_N = \mathbf{A}(\gamma_0)$ is given by

$$\mathbf{A}(\gamma_0) = \mathbf{D}'(\gamma_0)\mathbf{S}^{-1}(\gamma_0),$$

[6] In the appendix, we considered in detail the case of $T = 3$ and the single moment $E(e_{i3}y_{i1}) = 0$. In this case, the GMM estimator has a closed-form solution.

where γ_0 is the true value of γ, and

$$\begin{aligned}\mathbf{S}(\gamma_0) &= E\left[N\mathbf{M}_N(\gamma_0)\mathbf{M}_N'(\gamma_0)\right],\\ \mathbf{D}(\gamma_0) &= E\left[N^{-1}\sum_{i=1}^{N}\frac{\partial\mathbf{m}(\mathbf{y}_i,\gamma_0)}{\partial\gamma}\right] = N^{-1}\sum_{i=1}^{N}E\left(\frac{\partial\mathbf{m}(\mathbf{y}_i,\gamma_0)}{\partial\gamma}\right).\end{aligned}$$

But

$$E\left[N\mathbf{M}_N(\gamma_0)\mathbf{M}_N'(\gamma_0)\right] = E\left[N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}E\left[\mathbf{m}(\mathbf{y}_i,\gamma)\mathbf{m}'(\mathbf{y}_j,\gamma)\left|\mathbf{c}\right.\right]\right].$$

where $\mathbf{c} = (c_1, c_2, ..., c_N)'$. Note that conditional on $\mathbf{c}$, $\mathbf{y}_i$ and $\mathbf{y}_j$ are independently distributed, which establishes that $\mathbf{m}(\mathbf{y}_i,\gamma)$ and $\mathbf{m}(\mathbf{y}_j,\gamma)$ are also conditionally independent (since range of variations of $\mathbf{y}_i$ does not depend on γ). Hence, recalling that $E\left[\mathbf{m}(\mathbf{y}_i,\gamma)\right] = 0$, we have

$$E\left[N\mathbf{M}_N(\gamma_0)\mathbf{M}_N'(\gamma_0)\right] = N^{-1}\sum_{i=1}^{N}E\left[\mathbf{m}(\mathbf{y}_i,\gamma)\mathbf{m}'(\mathbf{y}_i,\gamma)\right].$$

In general, analytical expressions for $E\left[\frac{\partial\mathbf{m}(\mathbf{y}_i,\gamma_0)}{\partial\gamma}\right]$ and $E\left[\mathbf{m}(\mathbf{y}_i,\gamma)\mathbf{m}'(\mathbf{y}_i,\gamma)\right]$ will be a complicated function of $\mathbf{c}$. However, for a given initial consistent estimate of γ, say $\hat{\gamma}$, $\mathbf{A}_N$ can be consistently estimated as

$$\hat{\mathbf{A}}_N = \mathbf{A}_N(\hat{\gamma}) = \left[N^{-1}\sum_{i=1}^{N}\frac{\partial\mathbf{m}'(\mathbf{y}_i,\hat{\gamma})}{\partial\gamma}\right]\left[N^{-1}\sum_{i=1}^{N}\mathbf{m}(\mathbf{y}_i,\hat{\gamma})\mathbf{m}'(\mathbf{y}_i,\hat{\gamma})\right]^{-1}. \tag{18}$$

The asymptotic variance of $\hat{\gamma}_{GMM}$ is given by

$$AsyVar\left[\sqrt{N}(\hat{\gamma}_{GMM}-\gamma_0)\right] = \left[\mathbf{D}'(\gamma_0)\mathbf{S}^{-1}(\gamma_0)\mathbf{D}(\gamma_0)\right]^{-1},$$

which can be consistently estimated as

$$\widehat{Var}\left(\hat{\gamma}_{GMM}\right) = \frac{1}{N}\left[\hat{\mathbf{D}}'(\hat{\gamma}_{GMM})\hat{\mathbf{S}}^{-1}(\hat{\gamma}_{GMM})\hat{\mathbf{D}}(\hat{\gamma}_{GMM})\right]^{-1},$$

where

$$\hat{\mathbf{D}}(\hat{\gamma}_{GMM}) = N^{-1}\sum_{i=1}^{N}\frac{\partial\mathbf{m}'(\mathbf{y}_i,\hat{\gamma}_{GMM})}{\partial\gamma},$$

and

$$\hat{\mathbf{S}}(\hat{\gamma}_{GMM}) = N^{-1}\sum_{i=1}^{N}\mathbf{m}(\mathbf{y}_i,\hat{\gamma}_{GMM})\mathbf{m}'(\mathbf{y}_i,\hat{\gamma}_{GMM}).$$

The initial estimate of γ, say $\hat{\gamma}_{INI}$ can be obtained, for example, by imposing equal weights on the K moment conditions, namely

$$\hat{\gamma}_{INI} = \arg\min_{\gamma} \left[\mathbf{M}'_N(\gamma)\mathbf{M}_N(\gamma)\right].$$

This initial estimate can then be used to compute

$$\hat{\mathbf{A}}_N(\hat{\gamma}_{INI}) = \left[N^{-1}\sum_{i=1}^{N}\frac{\partial \mathbf{m}'(\mathbf{y}_i, \hat{\gamma}_{INI})}{\partial \gamma}\right]\left[N^{-1}\sum_{i=1}^{N}\mathbf{m}(\mathbf{y}_i, \hat{\gamma}_{INI})\mathbf{m}'(\mathbf{y}_i, \hat{\gamma}_{INI})\right]^{-1},$$

with $\hat{\gamma}_{GMM}$ is computed as

$$\hat{\gamma}_{GMM} = \arg\min_{\gamma} \left[\mathbf{M}'_N(\gamma)\hat{\mathbf{A}}'_N(\hat{\gamma}_{INI})\hat{\mathbf{A}}_N(\hat{\gamma}_{INI})\mathbf{M}_N(\gamma)\right], \tag{19}$$

An iterated GMM estimator can also be considered, where in computation of $\hat{\mathbf{A}}_N(\hat{\gamma}_{INI})$, $\hat{\gamma}_{INI}$ is replaced by $\hat{\gamma}_{GMM}$, and a new GMM estimator is computed using $\hat{\mathbf{A}}_N(\hat{\gamma}_{GMM})$, and so on.

The variance of $\hat{\rho}_{GMM} = -\ln(1-\hat{\gamma}_{GMM})$ can now be obtained using the delta method as

$$\widehat{Var}\left(\hat{\rho}_{GMM}\right) = \left(\frac{1}{1-\hat{\gamma}_{GMM}}\right)^2 \widehat{Var}\left(\hat{\gamma}_{GMM}\right).$$

The following theorem illustrates the issues involved in proving the asymptotic properties of the GMM estimator when only a single instrument, namely $y_{i,t-2}$, is used. The general case where additional instruments are considered can be established along similar lines.

Theorem 1. Suppose $y_{it} = 1(c_i + \rho_0 y_{i,t-1} + u_{it} \geq 0)$ for $i = 1, \ldots, N, t = 1, \ldots, T$ and the following conditions hold

(A1) $\Pr(c_i + \rho_0 > 0) = 1$, $\Pr(c_i > 0) = 1$, and $\Pr(c_i < \infty) = 1$ for $i = 1, 2, \ldots, N$.

(A2) $\{u_{it} : i = 1, 2, \ldots, N, t = 1, 2, \ldots, T\}$ is an independent array of random variables. u_{i1} is uniformly distributed on $[0, 1]$, while for $t > 1$, $-u_{it}$ is geometrically distributed with mean 1. $\{u_{it}\}$ is distributed independently of $\{c_i\}$.

(A3) $y_{i1} = 1\left(u_{i1} \leq \frac{1-e^{-c_i}}{1-e^{-c_i}(1-e^{-\rho_0})}\right)$, for $i = 1, \ldots, N$.

(A4) R is a compact subset of $\mathbb{R}$ containing ρ_0 in its interior.

(A5) For all $\rho \in R$, $N^{-1}\sum_{i=1}^{N} e_{it}(\rho)y_{i,t-2} \rightarrow_p E\left[e_{it}(\rho)y_{i,t-2}\right]$.

(A6) For all $\rho \in R$, $N^{-1}\sum_{i=1}^{N}\left(\partial e_{it}(\rho)/\partial\rho\right)y_{i,t-2} \rightarrow_p E\left[\left(\partial e_{it}(\rho)/\partial\rho\right)y_{i,t-2}\right]$.

(A7) $N^{-1/2}\sum_{i=1}^{N} e_{it}(\rho_0)y_{i,t-2} \rightarrow_d N(0, v^2)$, where $v^2 = \lim_{N\rightarrow\infty} N^{-1}\sum_{i=1}^{N} E\left[e_{it}^2(\rho_0)y_{i,t-2}^2\right] > 0$.

Then $N^{1/2}(\widehat{\rho}_{GMM} - \rho_0) \to_d N\left(0, \frac{v}{E[e_{it}(\rho_0)y_{i,t-2}]^2}\right)$, where $\hat{\rho}_{GMM} = -\ln(1 - \hat{\gamma}_{GMM})$, and $\widehat{\rho}_{GMM}$ is the GMM estimator defined (19) using $y_{i,t-2}$ as the instrument.

The positivity of c_i and $c_i + \rho_0$ in assumption (A1) allows us to circumvent the positivity constraint on geometrically distributed random variables. Without it, $\Pr(y_{it} = 1|c_i, y_{i,t-1}) = 1 - \exp(-\max\{0, c_i + \rho_0 y_{i,t-1}\})$, which greatly complicates the analysis. Assumption (A1) also requires c_i to be finite almost surely; clearly if $c_i = \infty$, then ρ_0 is not identified.

Assumptions (A2) and (A3) provide the probabilistic structure of the model conditional on c_i. Note that the uniform distribution of u_{i1} allows y_{i1} to have the correct stationary distribution given by (6). Together, assumptions (A2) and (A3) allows the distribution of y_{it} conditional on c_i to be stationary. This makes it possible to find analytic expressions for the unconditional moments of functions of the data.

Assumptions (A4) is standard in the GMM literature.

Assumptions (A5)-(A7) are high–level asymptotic conditions that hold under a variety of weak–dependence assumptions on the fixed effects. They hold when c_i are cross-sectionally independent but they may also allow for weak cross-sectional dependence, including weak spatial dependence.[7]

4 The Case of $\boldsymbol{\beta} \neq 0$

In contrast to the logit model studied in Honoré and Kyriazidou (2000, HK), it does not seem possible to identify $\boldsymbol{\beta}$ using the CMLE approach in the case of the exponential model considered in this paper. A key difference is that under exponential specification $\Pr(y_{it} = 1|c_i, y_{i,t-1}, \mathbf{x}_{it}) = 1 - \exp(-\rho y_{i,t-1} - \boldsymbol{\beta}'\mathbf{x}_{it} - c_i)$, and c_i does not get cancelled out from the numerator and the denominator of conditional probabilities. In contrast HK use a logistic specification, which is not subject to this problem, although to cancel the incidental parameters in the context of dynamic

[7] The assumptions we lay out here demonstrate the fact that while the asymptotic properties of GMM estimators such as consistency and asymptotic normality are established under high level regularity conditions as in Hansen (1982), whether they are satisfied in a specific nonlinear model is often technically involved and has to be examined case by case. It is worth noting that in the literature where GMM estimators are proposed, the conventional approach has been to derive moment conditions of the model and then claim the GMM estimators based on these moment conditions are consistent and asymptotically normally distributed implicitly assuming that the required regularity conditions are satisfied.

logit models we must have $T \geq 4$. See Section 2.1 of HK. But the GMM procedure is still applicable and can be used to identify both γ and $\boldsymbol{\beta}$ under the exponential model. The GMM approach has the added advantage that it does not require strong conditions on the covariates. Recall that in the case of the logistic model with a single exogenous regressor and $T = 4$, as shown by HK, identification of β requires $x_{i2} = x_{i3}$ with $x_{i1} \neq x_{i2}$, for all i.

4.1 GMM Estimation in the General Case

In the case where $\boldsymbol{\beta} \neq \mathbf{0}$, the dynamic non-linear autoregressive model, (8), associated to the binary choice model generalizes to

$$y_{it} = F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i) + \left[F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i + \rho) - F(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i)\right] y_{i,t-1} + \varepsilon_{it},$$

and we continue to have $E\left(\varepsilon_{it} \left| y_{i,t-1}, y_{i,t-2}, ...; \mathbf{x}_{it}, \mathbf{x}_{i,t-1}, ... \right.\right) = 0$. In the exponential case under consideration, the non-linear AR(1) formulation can be written as

$$y_{it} - 1 = \exp(-\boldsymbol{\beta}'\mathbf{x}_{it} - c_i) + \exp(-\boldsymbol{\beta}'\mathbf{x}_{it} - c_i)(1 - \exp(-\rho))y_{i,t-1} + \varepsilon_{it}.$$

Setting $\gamma = 1 - \exp(-\rho)$ and solving for the fixed effect as before,

$$\exp(-c_i) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\,(1 - y_{it})}{(1 - \gamma y_{i,t-1})} + \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\varepsilon_{it}}{(1 - \gamma y_{i,t-1})}.$$

Now first differencing to eliminate c_i yields

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\,(1 - y_{it})}{(1 - \gamma y_{i,t-1})} - \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{i,t-1})\,(1 - y_{i,t-1})}{(1 - \gamma y_{i,t-2})} = -\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\varepsilon_{it}}{(1 - \gamma y_{i,t-1})} - \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{i,t-1})\varepsilon_{i,t-1}}{(1 - \gamma y_{i,t-2})},$$

which after some algebra simplifies to

$$\begin{aligned} e_{it} &= \exp(\boldsymbol{\beta}'\Delta\mathbf{x}_{it})\left(\frac{1 - \gamma y_{i,t-2}}{1 - \gamma y_{i,t-1}}\right)\varepsilon_{it} - \varepsilon_{i,t-1} \\ &= (1 - y_{i,t-1}) - (1 - y_{it})\left(\frac{1 - \gamma y_{i,t-2}}{1 - \gamma y_{i,t-1}}\right)\exp(\boldsymbol{\beta}'\Delta\mathbf{x}_{it}). \end{aligned} \tag{20}$$

It is easily seen that e_{it} given above reduces to (15) if we set $\boldsymbol{\beta} = \mathbf{0}$, as to be expected. Also as before, $y_{i,t-2}$,$y_{i,t-3}$,.... and the constant can be used as instruments.[8] Additional instruments are

[8] The same caveat as mentioned earlier continues to hold. $E(e_{it}) = 0$ for $(\gamma, \boldsymbol{\beta}) = (0, \mathbf{0})$ and for $(\gamma, \boldsymbol{\beta}) = (\gamma_0, \boldsymbol{\beta}_0)$. Therefore, the constant should never be used as an instrument unless accompanied by at least one lagged variable as an additional instrument.

also available depending on the nature of the covariates. In the case where $\mathbf{x}_{it}$ is exogenous, then the regressors $\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}$ can also be used as instruments.

In empirical applications of the GMM approach the choice of instruments can play an important role for the small sample properties of the estimators. The problem becomes particularly serious in panel data models where the number of instruments can rise quite rapidly with T. The pitfalls in using too many instruments in the case of linear dynamic panel data models is investigated in Roodman (2009). In the case of non-linear specifications, the use of additional instruments that involve powers of $y_{i,t-s,}$for $s \geq 2$, or powers of lagged exogenous variables, such as $y_{i,t-2}y_{i,t-3}$, $\mathbf{x}_{i,t-s} \otimes \mathbf{x}_{i,t-s}$, and $y_{i,t-2}\mathbf{x}_{i,t-s}$, can also be justified which could lead to even a larger set of instruments to be used in GMM estimation. A number of procedures have been proposed to deal with this problem. Carrasco (2012) proposes using regularization techniques to invert the covariance matrix of the instruments. Mehrhoff (2009) proposes factorizing the instrument set whereby the full set of instruments is replaced by a few principal components of the instrument set. Both approaches rely on related choice parameters such as the extent of regularization/shrinkage in the case of Carrasco's approach and the number of principle components to be used as instruments. The application of these basically linear techniques to the non-linear specification that we consider could also be problematic as they need not be optimal in non-linear settings. In view of these difficulties we do not recommend the use of GMM approach developed in this paper for applications where T is relatively large, say more than 6. In case of non-linear panels with moderate to large T samples the ML approach combined with bias correction (as proposed by Carro, 2007) might be more appropriate.

4.2 Discussion on Robustness of the Exponential Specification

As discussed in Section 1, various specifications of dynamic binary choice panel data models have been used in the literature depending on their convenience and/or whether they enable the researcher to resolve the incidental parameter problem. In the same vein, we propose to use the exponential specification and construct GMM estimators that are consistent and asymptotically normally distributed. As for any specification in the parametric approach, a natural question is how robust it is with regard to misspecification. More specifically, suppose that for a realization of $y_{i,t-1} = \{0, 1\}$ and $\mathbf{x}_{it} = \mathbf{x}_i^0$, the true distribution function is given by

$$\Pr(y_{it} = 1 \left| y_{i,t-1}, c_i, \mathbf{x}_i^0 \right.) = F(\rho y_{i,t-1} + \boldsymbol{\beta}' \mathbf{x}_i^0 + c_i),$$

but consider an investigator that uses the exponential specification and obtains

$$\Pr(y_{it}=1\left|y_{i,t-1},c_{i,e},\mathbf{x}_i^0;M_e\right.)=1-\exp(-\rho_e y_{i,t-1}-\boldsymbol{\beta}_e'\mathbf{x}_i^0-c_{i,e}),$$

where the symbol M_e denotes an exponential distribution to distinguish it from the true distribution function. In the case where the Markov chain underlying the true process is stationary we have $0<F(\rho y_{i,t-1}+\boldsymbol{\beta}'\mathbf{x}_i^0+c_i)<1$, for all finite values of ρ, $\boldsymbol{\beta}'\mathbf{x}_i^0$, and c_i, and hence there exists c_{ie} such that $\Pr(y_{it}=1\left|y_{i,t-1},c_{i,e},\mathbf{x}_i^0;M_e\right.)=F(\rho y_{i,t-1}+\boldsymbol{\beta}'\mathbf{x}_i^0+c_i)$, namely

$$\begin{aligned}-c_{ie} &= \ln\left[1-F(\rho+\boldsymbol{\beta}'\mathbf{x}_i^0+c_i)\right]+\rho_e+\boldsymbol{\beta}_e'\mathbf{x}_i^0,\text{ if }y_{i,t-1}=1,\\ &= \ln\left[1-F\left(\boldsymbol{\beta}'\mathbf{x}_i^0+c_i\right)\right]+\boldsymbol{\beta}_e'\mathbf{x}_i^0,\text{ if }y_{i,t-1}=0.\end{aligned}$$

Since, under the exponential distribution c_{ie}'s are treated as fixed effects and are allowed to have an arbitrary degree of correlations with x_{it}, then it is possible to match any distribution function, $F(.)$, that satisfy the stationary condition $0<F(.)<1$ for a given realization of $\mathbf{x}_{it}$. It is important to emphasise that this match is local and not global, and holds approximately in the neighborhood of $\mathbf{x}_i^0$, which can be taken as the sample mean, $\bar{\mathbf{x}}_i$.[9] This does not seem to be an important limitation since in most empirical applications the investigator is concerned with 'average' effects and as we shall see from the Monte Carlo results reported in the sub-section 5.4, the average partial effects from logistic distribution tend to be well approximated if the estimates are incorrectly based on an exponential distribution.

5 Simulation Studies

In order to investigate the performance of the GMM and CMLE estimators we conduct a series of Monte Carlo studies, which we summarize here. We have endeavored where possible to match the Monte Carlo design employed by Honoré and Kyriazidou (2000).[10]

5.1 The GMM Estimator

To investigate the small sample properties of the proposed GMM estimator, we generate data from the exponential dynamic binary choice model, with $\rho=0.5$, and include a single exogenous

[9] We thank a referee for drawing our attention to this point.

[10] The full set of Monte Carlo results is available from the authors on request.

regressor in the model. We draw $c_i \sim |N(0, \sigma_c^2)|$ and $x_{it} \sim |N(0,1)|$, independently over i and t. We then set $\sigma_c = \beta$ so that the fixed effects and exogenous regressors each contribute an equal amount of variation. The two parameters are solved numerically for a proportion of 1s in the population of $\bar{\pi} = 50\%$, which gives $\sigma_c = \beta = 0.318815$. The distribution of y_{i1} is set to the stationary distribution conditional on c_i and x_{i1}. We generate data sets of sizes $T = 3, 4, 6, 8$ and $N = 250, 500, 1,000, 2,500, 5,000, 10,000$ and consider the mean, variance, bias, and RMSE of the estimates for ρ and β in $2,000$ replications for each experiment. The estimates are obtained using the moment conditions

$$\begin{aligned}
E(e_{it}) &= 0, \qquad && t = 3, 4, \ldots, T, && \\
E(x_{is}e_{it}) &= 0, && t = 3, 4, \ldots, T, && s = 1, 2, \ldots, T, \\
E(y_{is}e_{it}) &= 0, && t = 3, 4, \ldots, T, && s = 1, 2, \ldots, t-2,
\end{aligned}$$

and using an estimate for the optimal choice of the GMM weight matrix. There are a total of $\frac{1}{2}(3T+1)(T-2)$ moment conditions. We also consider the size of the tests $H_0 : \rho = 0$ and power for $H_a : \rho = 0.6$ and $H_b : \rho = 0.4$ as well as the size of the tests $H_0 : \beta = 0$ and power for $H_a : \beta = 0.418815$ and $H_b : \beta = 0.218815$, all at 5% significance. Henceforth, this setting will be referred to as the benchmark specification.[11]

Tables 1 and 2 give results for variance, bias, and RMSE in the benchmark simulations. Variance, bias, and RMSE improve with larger N. RMSE and variance improve with increased T. However, the bias of the GMM estimator of ρ increases with T.

[11] To simplify the computations we first estimated γ and then estimated ρ as -ln$(1-\gamma)$. See (11). This approach requires $\gamma < 1$. In a number of experiments we encountered estimates for γ that were inadmissible (namely they were larger than 1). This was particularly the case for small values of N. However, the likelihood of obtaining an inadmissible estimate decreased sharply with N. As a check, in the case of a few experiments we also estimated ρ directly and without any restrictions and overall found the results to be very similar to the ones based on the indirect approach.

Table 1. Benchmark Small Samples Results for Variance, Bias, and RMSE of $\widehat{\rho}_{GMM}$.[⋆]

$T\backslash N$		250	500	1,000	2,500	5,000	10,000
3	Variance	0.0571	0.0326	0.0166	0.0065	0.0031	0.0016
	Bias	0.0032	-0.0014	0.0027	0.0009	-0.0007	0.0004
	RMSE	0.2239	0.1767	0.1282	0.0806	0.0556	0.0394
4	Variance	0.0240	0.0123	0.0066	0.0025	0.0012	0.0006
	Bias	-0.0446	-0.0253	-0.0104	-0.0041	-0.0020	-0.0011
	RMSE	0.1514	0.1110	0.0815	0.0503	0.0349	0.0248
6	Variance	0.0105	0.0060	0.0026	0.0010	0.0005	0.0003
	Bias	-0.0889	-0.0442	-0.0209	-0.0057	-0.0026	-0.0011
	RMSE	0.1252	0.0879	0.0554	0.0328	0.0226	0.0159
8	Variance	0.0075	0.0042	0.0018	0.0006	0.0003	0.0002
	Bias	-0.1557	-0.0774	-0.0309	-0.0081	-0.0032	-0.0014
	RMSE	0.1613	0.0992	0.0528	0.0267	0.0181	0.0128

[⋆] $\rho = 0.5$, $\beta = 0.32$, $x_{it} = |N(0,1)|$, $c_i \sim |N(0, 0.32^2)|$.

Table 2. Benchmark Small Samples Results for Variance, Bias, and RMSE of $\widehat{\beta}_{GMM}$.[⋆]

$T\backslash N$		250	500	1,000	2,500	5,000	10,000
3	Variance	0.0192	0.0078	0.0035	0.0015	0.0007	0.0004
	Bias	0.0100	0.0073	0.0024	0.0012	0.0006	0.0007
	RMSE	0.1300	0.0869	0.0591	0.0384	0.0274	0.0195
4	Variance	0.0101	0.0039	0.0019	0.0008	0.0004	0.0002
	Bias	0.0024	0.0016	-0.0012	0.0006	0.0000	0.0003
	RMSE	0.0942	0.0609	0.0430	0.0277	0.0198	0.0137
6	Variance	0.0047	0.0021	0.0010	0.0004	0.0002	0.0001
	Bias	-0.0172	-0.0040	-0.0002	0.0006	0.0003	0.0005
	RMSE	0.0653	0.0448	0.0323	0.0206	0.0140	0.0099
8	Variance	0.0035	0.0016	0.0008	0.0003	0.0001	0.0001
	Bias	-0.0323	-0.0128	-0.0008	0.0005	0.0003	0.0001
	RMSE	0.0607	0.0406	0.0279	0.0175	0.0122	0.0085

[⋆] $\rho = 0.5$, $\beta = 0.32$, $x_{it} = |N(0,1)|$, $c_i \sim |N(0, 0.32^2)|$.

Tables 3 and 4 give the results for size and power. For $T = 3$ and 4, size is satisfactory even for a relatively small N. However, there are large size distortions for $T = 6$ and 8, most likely owing to the rapidly (quadratically) growing number of instruments. For these cases, one needs large N to reduce the percentage of over-rejection. Notably, size for the β tests improves more rapidly than the size for the ρ tests with increased N. We need $N \geq 2,500$ to bring down the size to below 10% for ρ and $N \geq 1,000$ for β.

Table 3. Benchmark Small Samples Results for Size and Power of Tests Based on $\hat{\rho}_{GMM}$.[⋆]

$T\backslash N$		250	500	1,000	2,500	5,000	10,000
3	Size H_0^*	0.0536	0.0636	0.0627	0.0600	0.0515	0.0545
	Power $H_a^\dagger$	0.1157	0.1382	0.1728	0.2811	0.4595	0.7115
	Power $H_b^\ddagger$	0.0433	0.0683	0.1102	0.2331	0.4255	0.7380
4	Size H_0	0.0817	0.0728	0.0697	0.0540	0.0545	0.0505
	Power H_a	0.2240	0.2619	0.3180	0.5560	0.8315	0.9780
	Power H_b	0.0618	0.0781	0.1976	0.5045	0.8205	0.9875
6	Size H_0	0.2478	0.1508	0.0901	0.0625	0.0560	0.0530
	Power H_a	0.5937	0.5780	0.6855	0.9045	0.9955	1.0000
	Power H_b	0.0986	0.1549	0.3540	0.8525	0.9935	1.0000
8	Size H_0	0.7072	0.3977	0.1816	0.0750	0.0530	0.0605
	Power H_a	0.9309	0.8785	0.9020	0.9875	1.0000	1.0000
	Power H_b	0.3026	0.1433	0.4667	0.9630	1.0000	1.0000

[⋆] $\rho = 0.5$, $\beta = 0.32$, $x_{it} = |N(0,1)|$, $c_i \sim |N(0, 0.32^2)|$.

[*] $H_0 : \rho = 0.5$. [†] $H_a : \rho = 0.6$. [‡] $H_b : \rho = 0.4$ (5% level).

Table 4. Benchmark Small Samples Results for Size and Power of Tests Based on $\hat{\beta}_{GMM}$.[⋆]

$T \backslash N$		250	500	1,000	2,500	5,000	10,000
3	Size H_0^*	0.0604	0.0511	0.0541	0.0490	0.0610	0.0540
	Power $H_a^\dagger$	0.1608	0.2457	0.4184	0.7274	0.9430	0.9990
	Power $H_b^\ddagger$	0.1140	0.2102	0.4149	0.7654	0.9705	0.9995
4	Size H_0	0.0800	0.0660	0.0522	0.0545	0.0505	0.0445
	Power H_a	0.2564	0.4081	0.6670	0.9400	0.9990	1.0000
	Power H_b	0.1940	0.4023	0.6354	0.9675	1.0000	1.0000
6	Size H_0	0.1450	0.0875	0.0641	0.0620	0.0450	0.0485
	Power H_a	0.5737	0.7185	0.8848	0.9975	1.0000	1.0000
	Power H_b	0.3658	0.6500	0.9049	0.9990	1.0000	1.0000
8	Size H_0	0.2732	0.1376	0.0950	0.0660	0.0565	0.0590
	Power H_a	0.8258	0.8842	0.9630	1.0000	1.0000	1.0000
	Power H_b	0.4664	0.7399	0.9750	1.0000	1.0000	1.0000

[⋆] $\rho = 0.5$, $\beta = 0.32$, $x_{it} = |N(0,1)|$, $c_i \sim |N(0, 0.32^2)|$.

[*] $H_0 : \beta = 0.3188$. [†] $H_a : \beta = 0.4188$. [‡] $H_b : \beta = 0.2188$ (5% level).

We next modify the benchmark DGP of y_{it}, x_{it} and c_i in various ways and look at the behavior of our estimators. A selection of the results of these alternative specifications is given in Table 5 for $T = 3$ and $N = 500$.

First, we look at the effect of changing the variance of the fixed effects. We increase σ_c so that $\bar{\pi} = 0.75$ and then further so that $\bar{\pi} = 0.95$. As to be expected, increasing σ_c causes a deterioration of the estimates, increasing the percentage of γ's falling out of bounds, along with variance, bias, and RMSE, a rise in size and decrease in power. However, the empirical size is still generally close to the nominal size for $N \geq 5,000$.

Next, we vary ρ and β individually in the benchmark simulation, choosing $\rho = \rho^{\text{bm}} \pm 0.4$ and $\beta = \beta^{\text{bm}} \pm 0.2$, where ρ^{bm} and β^{bm} denote the benchmark values. These variations impart little change to the results of the benchmark. The higher value of ρ causes a fall in the percentage of γ falling out of bounds.

We then modify the benchmark to allow the fixed effect to be correlated with the exogenous variables. We set $c_i = b_{\omega,T}(\omega \bar{x}_i^{\text{bm}} + (1-\omega) c_i^{\text{bm}})$, where $\bar{x}_i^{\text{bm}} = \frac{1}{T}\sum_{t=1}^{T} x_{it}^{\text{bm}}$ and c_i^{bm} is the benchmark value of the fixed effect, for $\omega = 0.25, 0.50, 0.75$. $b_{\omega,T}$ is chosen so that $\bar{\pi}$ is equal to the benchmark value. This has little or no effect on the results.

We also consider the effect of cross-sectional heterogeneity in x_{it} by modifying the benchmark exogenous process to, $x_{it} = h(\mu_i + \sigma_i|\varepsilon_{it}|)$, where $\mu_i \sim U(0,1)$, $\sigma_i^2 \sim \chi_2^2$, and $\varepsilon_{it} \sim N(0,1)$. We set $h = 0.52444$ to match the value of $\bar{\pi}$ in the benchmark model. We find that the results for the estimates of ρ are not much affected by the heterogeneity in the x_{it} processes. The results for β, on the other hand, have higher variance, bias, and RMSE than the results obtained under the benchmark model. The same also applies to size and power where under heterogeneity we observe a deterioration in size and power as compared to the benchmark case.

We then consider the effect of autocorrelation in the exogenous variables on the results. In this case we modify the benchmark exogenous process to $x_{it} = |0.1\zeta_{it} + d_T + 0.2t|$, where for each i, ζ_{it} is a Gaussian AR(1) with autoregressive coefficient 0.5, variance 1, and independently distributed across i. c_i are generated as in the benchmark case. The parameters are calibrated by simulation to produce an expected proportion of 1's of $\bar{\pi}^{\text{bm}}$ in populations of size $N = 10,000$. We find that autocrrelation in the covariates has no significant effect on the results for the estimates of ρ. However, the variance, bias, and RMSE of the estimate of β are all higher than in the benchmark. Size also deteriorates with autocorrelation, with power being significantly lower than under the benchmark case.

Table 5. Sample of Small Sample Results for GMM Estimation Under Alternative Specifications ($T = 3$ and $N = 500$).

Specification[♭]	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
ρ Results												
% of γs ≥ 1	4.15	15	33.5	4.3	3.5	0.85	31.3	2.7	2.35	2.5	3.7	0
Average % of 1s	49.82	74.4	94.38	57.79	38.93	59.68	40.08	49.79	49.74	49.69	49.67	50.33
Variance	0.0326	0.0706	0.3837	0.0440	0.0257	0.0741	0.0098	0.0316	0.0312	0.0314	0.0299	0.0314
Bias	-0.0014	-0.0072	0.0050	-0.0026	-0.0043	0.0000	0.0354	-0.0012	-0.0017	0.0024	-0.0070	-0.0139
RMSE	0.1767	0.2450	0.5052	0.2051	0.1576	0.2711	0.0873	0.1755	0.1746	0.1749	0.1698	0.1777
Size H_0^*	0.0636	0.0600	0.0534	0.0601	0.0617	0.0746	0.0189	0.0612	0.0538	0.0590	0.0571	0.0680
Power $H_a^\dagger$	0.1382	0.1024	0.0729	0.1306	0.1560	0.1140	0.0633	0.1408	0.1413	0.1277	0.1350	0.1555
Power $H_b^\ddagger$	0.0683	0.0312	0.0286	0.0601	0.0622	0.0711	0.1426	0.0612	0.0681	0.0708	0.0675	0.0455
β Results												
Variance	0.0078	0.0206	13.1853	0.0141	0.0040	0.0125	0.0052	0.0084	0.0086	0.0086	0.0196	0.0965
Bias	0.0073	0.0116	0.1297	0.0166	0.0029	0.0114	0.0088	0.0084	0.0077	0.0094	0.0130	0.0111
RMSE	0.0869	0.1328	2.9630	0.1173	0.0623	0.1118	0.0599	0.0909	0.0918	0.0922	0.1378	0.3109
Size H_0^{**}	0.0511	0.0524	0.0774	0.0559	0.0534	0.0575	0.0488	0.0462	0.0471	0.0508	0.0696	0.0450
Power $H_a^{\dagger\dagger}$	0.2457	0.1276	0.1053	0.1594	0.3767	0.1599	0.2897	0.2199	0.2161	0.2154	0.1672	0.0685
Power $H_b^{\ddagger\ddagger}$	0.2102	0.1000	0.0624	0.1411	0.3876	0.1704	0.2933	0.2163	0.2110	0.2113	0.1345	0.0555

[♭] (1) benchmark, (2) mediums σ_c, (3) high σ_c, (4) high β, (5) low β, (6) high ρ, (7) low ρ, (8) $\omega = 0.25$, (9) $\omega = 0.50$, (10) $\omega = 0.75$, (11) heterogenous x_{it}, (12) autocorrelated x_{it}.

[*] $H_0 : \rho = \rho_0$, where ρ_0 is the value of ρ used in the DGP of the particular specification. [†] $H_a : \rho = \rho_0 + 0.1$. [‡] $H_b : \rho = \rho_0 - 0.1$. [**] $H_0 : \beta = \beta_0$, where β_0 is the value of β used in the DGP of the particular specification.. [††] $H_a : \beta = \beta_0 + 0.2$. [‡‡] $H_b : \beta = \beta_0 - 0.2$ (5% level).

5.2 GMM versus CMLE

In this subsection we report comparative results for GMM and CMLE estimation methods for ρ with $\boldsymbol{\beta} = \mathbf{0}$. Recall that CMLE method is not applicable to the exponential model if $\boldsymbol{\beta} \neq \mathbf{0}$. GMM estimation uses the following moment conditions,

$$\begin{aligned} E(e_{it}) &= 0, & t &= 3, \ldots, T, & & \\ E(y_{is}e_{it}) &= 0, & t &= 3, \ldots, T, & s &= 1, \ldots, t-2. \end{aligned}$$

The CMLE procedure is described in Section 3.3.

The results for bias and RMSE are summarized in Tables 6 and 7, and for size and power in Tables 8 and 9. In terms of RMSE, GMM outperforms CMLE for all values of T under consideration ($T = 3, 4,$ 6,8), although for $T = 6$ and 8 GMM shows a higher degree of bias than CMLE. In terms of size, CMLE does better than GMM, and matches the nominal size for all values of T, whilst GMM tends to over-reject when $T > 6$. But generally GMM outperforms CMLE in terms of power when the sizes are comparable.

Table 6. Small Samples Results for CMLE Estimates of ρ when $\boldsymbol{\beta} = \mathbf{0}$.*

$T\backslash N$		250	500	1,000	2,500	5,000	10,000
3	Variance	0.1000	0.0484	0.0237	0.0093	0.0044	0.0024
	Bias	0.0300	0.0150	0.0107	0.0031	0.0025	0.0006
	RMSE	0.3176	0.2205	0.1543	0.0966	0.0666	0.0487
4	Variance	0.0477	0.0230	0.0116	0.0050	0.0022	0.0011
	Bias	0.0078	0.0034	0.0052	0.0017	-0.0009	-0.0008
	RMSE	0.2186	0.1518	0.1077	0.0706	0.0474	0.0336
6	Variance	0.0300	0.0130	0.0064	0.0026	0.0013	0.0006
	Bias	-0.0100	-0.0031	-0.0039	-0.0007	0.0003	-0.0005
	RMSE	0.1600	0.1141	0.0804	0.0512	0.0357	0.0255
8	Variance	0.0203	0.0105	0.0055	0.0020	0.0010	0.0005
	Bias	-0.0019	0.0009	-0.0008	-0.0004	-0.0006	0.0001
	RMSE	0.1427	0.1026	0.0745	0.0447	0.0318	0.0230

* $\rho = 0.5$, $\beta = 0$, $c_i \sim |N(0, 0.32^2)|$.

Table 7. Small Samples Results for GMM Estimates of ρ when $\boldsymbol{\beta} = \mathbf{0}$.*

$T \backslash N$		250	500	1,000	2,500	5,000	10,000
3	Variance	0.0640	0.0325	0.0170	0.0069	0.0032	0.0017
	Bias	0.0301	0.0130	0.0055	0.0005	0.0001	0.0007
	RMSE	0.2427	0.1774	0.1301	0.0828	0.0567	0.0412
4	Variance	0.0264	0.0135	0.0067	0.0027	0.0012	0.0006
	Bias	-0.0121	-0.0045	-0.0022	-0.0015	-0.0017	-0.0001
	RMSE	0.1599	0.1161	0.0818	0.0522	0.0353	0.0249
6	Variance	0.0115	0.0054	0.0026	0.0010	0.0005	0.0002
	Bias	-0.0288	-0.0121	-0.0057	-0.0014	-0.0005	-0.0006
	RMSE	0.1105	0.0747	0.0515	0.0318	0.0217	0.0156
8	Variance	0.0080	0.0036	0.0015	0.0006	0.0003	0.0002
	Bias	-0.0514	-0.0174	-0.0052	-0.0018	-0.0005	0.0000
	RMSE	0.1030	0.0622	0.0394	0.0249	0.0177	0.0127

* $\rho = 0.5$, $\beta = 0$, $c_i \sim |N(0, 0.32^2)|$.

Table 8. Small Sample Size and Power Results for CMLE Estimation of ρ when $\boldsymbol{\beta} = \mathbf{0}$.*

$T \backslash N$		250	500	1000	2,500	5,000	10,000
3	Size H_0^*	0.0445	0.0440	0.0520	0.0410	0.0430	0.0540
	Power $H_a^\dagger$	0.0640	0.0730	0.0915	0.1750	0.2895	0.5475
	Power $H_b^\ddagger$	0.0455	0.0600	0.0900	0.1715	0.3100	0.5320
4	Size H_0	0.0525	0.0510	0.0560	0.0600	0.0540	0.0510
	Power H_a	0.0800	0.0970	0.1490	0.3155	0.5650	0.8450
	Power H_b	0.0625	0.0900	0.1545	0.3265	0.5365	0.8430
6	Size H_0	0.0500	0.0475	0.0500	0.0525	0.0475	0.0535
	Power H_a	0.1000	0.1530	0.2640	0.4995	0.7935	0.9765
	Power H_b	0.0900	0.1415	0.2230	0.4990	0.7990	0.9725
8	Size H_0	0.0455	0.0520	0.0615	0.0485	0.0445	0.0540
	Power H_a	0.1090	0.1600	0.3000	0.6010	0.8710	0.9920
	Power H_b	0.1050	0.1790	0.2890	0.6025	0.8810	0.9905

* $\rho = 0.5$, $\beta = 0$, $c_i \sim |N(0, 0.32^2)|$.

* $H_0 : \rho = 0.5$. † $H_a : \rho = 0.6$. ‡ $H_b : \rho = 0.4$ (5% level).

Table 9. Small Sample Size and Power Results for GMM Estimation of ρ when $\boldsymbol{\beta} = \mathbf{0}$.*

$T\backslash N$		250	500	1,000	2,500	5,000	10,000
3	Size H_0^*	0.0496	0.0509	0.0533	0.0510	0.0485	0.0515
	Power $H_a^\dagger$	0.0926	0.1081	0.1523	0.2646	0.4405	0.6915
	Power $H_b^\ddagger$	0.0380	0.0566	0.1016	0.2216	0.3895	0.7025
4	Size H_0	0.0712	0.0607	0.0595	0.0650	0.0545	0.0480
	Power H_a	0.1761	0.2007	0.2890	0.5305	0.8090	0.9755
	Power H_b	0.0577	0.1069	0.2150	0.4840	0.8130	0.9815
6	Size H_0	0.1097	0.0795	0.0690	0.0545	0.0425	0.0420
	Power H_a	0.3179	0.3865	0.5750	0.8795	0.9930	1.0000
	Power H_b	0.1268	0.2310	0.4640	0.8840	0.9960	1.0000
8	Size H_0	0.1989	0.1055	0.0615	0.0490	0.0580	0.0540
	Power H_a	0.5746	0.5915	0.7735	0.9785	1.0000	1.0000
	Power H_b	0.1643	0.3490	0.7035	0.9840	1.0000	1.0000

* $\rho = 0.5$, $\beta = 0$, $c_i \sim |N(0, 0.32^2)|$.

* $H_0 : \rho = 0.5$. † $H_a : \rho = 0.6$. ‡ $H_b : \rho = 0.4$ (5% level).

5.3 Reducing the Number of Instruments

In order to address the issue of the large number of instruments, we us the benchmark DGP and limit the number of instruments adopting five different procedures. (1) The first (benchmark) procedure uses all available linear instruments as detailed in subsection 5.1. Procedure (2) restricts the set of instruments, following the method proposed by Mehrhoff (2009), by utilizing only the few largest principal components (PC) of the instruments in estimation. The number of principal components is selected so that at least 95% of the total variation of the instruments under consideration is explained by the PC's.[12] Procedure (3) reduces the number of instruments to two lags of y_{it} and x_{it}, as well as the constant. That is, it utilizes the following $5T - 11$ moment conditions,

$$E(e_{it}) = 0,\ E(x_{it}e_{it}) = 0,\ E(x_{i,t-1}e_{it}) = 0,\ \text{for } t = 3, 4, ..., T;$$

$$E(y_{i,t-2}e_{it}) = 0,\ \text{for } t = 3, 4, ..., T;$$

$$E(y_{i,t-3}e_{it}) = 0, \text{ for } t = 4, 5, ..., T.$$

[12] We also tried setting the threshold at 90%. This gets rid of too much information when T is small and does not help much for large T so it does not substantively change the main results of our experiments.

Procedure (4) applies Mehrhoff's method to the reduced set of instruments under (3). Finally, procedure (5) reduces the number of instruments further by using two lags of y_{it}, and only one lag of x_{it}, as well as the constant, bringing the total number of instruments to $4T-9$.

Tables 10 and 11 report the results for $T = 4, 6, 8$ and $N = 250, 500, 2, 500$, as these were the sample sizes for which the GMM estimator performed worse. Reducing the number of instruments typically improves bias and size at a small cost to variance and RMSE. The benefit of the reduction in the number of instruments is most pronounced for $T = 6, 8$, where bias and size are significantly improved. In terms of variance, procedure (1) is optimal. Procedures (4) and (5) have the lowest bias. Procedure (2) is best for the RMSE of $\widehat{\beta}$. For the RMSE of $\widehat{\rho}$, there is no clear winner among the alternative instrument selection procedures, although procedure (5) performs best in terms of RMSE for $T = 8$. Procedures (4) and (5) have the best size properties. We conclude that the GMM estimator performs well for large T when the number of instruments is reduced by one of the methods employed here.

5.4 Average Partial Effects

To provide additional support for our choice of the exponential specification, here we present evidence of its ability to reproduce the average partial effects of a dynamic logistic model. Suppose the DGP is given by the logistic specification

$$\Pr(y_{it} = 1 \,|y_{i,t-1}, c_{il}, x_{it}) = \frac{e^{\rho_l y_{i,t-1}+\beta_l x_{it}+c_{il}}}{1+e^{\rho_l y_{i,t-1}+\beta_l x_{it}+c_{il}}}.$$

Then the marginal effect for continuous x_{it} is

$$\frac{\partial \Pr\left(y_{it} = 1|y_{i,t-1}, c_{il}, x_{it}\right)}{\partial x_{it}} = \frac{\beta_l e^{\rho_l y_{i,t-1}+\beta_l x_{it}+c_{il}}}{(1+e^{\rho_l y_{i,t-1}+\beta_l x_{it}+c_{il}})^2}.$$

On the other hand, the marginal effect of $y_{i,t-1}$ is given as

$$\Pr\left(y_{it} = 1 \,|y_{i,t-1} = 1, c_{il}, x_{it}\right) - \Pr\left(y_{it} = 1 \,|y_{i,t-1} = 0, c_{il}, x_{it}\right) = \frac{e^{\rho_l+\beta_l x_{it}+c_{il}}}{1+e^{\rho_l+\beta_l x_{it}+c_{il}}} - \frac{e^{\beta_l x_{it}+c_{il}}}{1+e^{\beta_l x_{it}+c_{il}}}.$$

For a particular x_{it}, say the average $\bar{x} = \frac{1}{NT}\sum_{i,t} x_{it}$, we may be interested in the average marginal effect over the entire population (i.e. averaging over the fixed effects). These quantities may be

Table 10. Small Sample Results for GMM Estimation with Reduced Number of Instruments (ρ results).

$N \backslash T$		4					6					8				
	Est. Method[♭]	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
250	% of γs ≥ 1	11.85	12.35	7.75	6.65	4.5	14.85	19.65	3.4	3.05	1.65	18.2	25.3	2.4	2.4	1.05
	Ave. # Inst's[♯]	13	10	9	7	7	38	28	19	14	15	75	55	29	21	23
	Variance	0.0240	0.0262	0.0247	0.0288	0.0261	0.0105	0.0132	0.0119	0.0134	0.0120	0.0075	0.0089	0.0079	0.0087	0.0081
	Bias	-0.0446	-0.0328	-0.0206	-0.0084	-0.0203	-0.0889	-0.1010	-0.0316	-0.0233	-0.0205	-0.1557	-0.1535	-0.0411	-0.0326	-0.0248
	RMSE	0.1514	0.1546	0.1521	0.1641	0.1591	0.1252	0.1371	0.1116	0.1162	0.1105	0.1613	0.1558	0.0969	0.0975	0.0926
	Size H_0^*	0.0817	0.0759	0.0672	0.0616	0.0691	0.2478	0.2091	0.0916	0.0774	0.0859	0.7072	0.5562	0.1414	0.1050	0.1046
	Power $H_a^\dagger$	0.2240	0.1934	0.1702	0.1527	0.1623	0.5937	0.5016	0.3370	0.2666	0.2832	0.9309	0.8534	0.4734	0.3847	0.3926
	Power $H_b^\ddagger$	0.0618	0.0531	0.0650	0.0621	0.0639	0.0986	0.0747	0.1211	0.0970	0.1281	0.3026	0.1961	0.1450	0.1260	0.1789
500	% of γs ≥ 1	4.55	4.9	2.2	1.35	0.7	2.85	7.05	0.25	0.25	0.05	3.7	10.15	0	0	0
	Ave. # Inst's	13	10	9	7	7	38	28	19	14	15	75	55	29	21	23
	Variance	0.0123	0.0139	0.0128	0.0146	0.0129	0.0060	0.0090	0.0061	0.0071	0.0061	0.0042	0.0069	0.0041	0.0047	0.0040
	Bias	-0.0253	-0.0207	-0.0131	-0.0104	-0.0134	-0.0442	-0.0631	-0.0142	-0.0098	-0.0098	-0.0774	-0.0984	-0.0153	-0.0113	-0.0089
	RMSE	0.1110	0.1169	0.1126	0.1204	0.1142	0.0879	0.1100	0.0795	0.0846	0.0789	0.0992	0.1219	0.0656	0.0693	0.0639
	Size H_0	0.0728	0.0736	0.0649	0.0669	0.0564	0.1508	0.1651	0.0842	0.0727	0.0760	0.3977	0.3829	0.0970	0.0745	0.0810
	Power H_a	0.2619	0.2376	0.2316	0.2048	0.2200	0.5780	0.5062	0.3880	0.3208	0.3597	0.8785	0.7969	0.5275	0.4440	0.4815
	Power H_b	0.0781	0.0857	0.0992	0.0882	0.0962	0.1549	0.0979	0.2261	0.1955	0.2281	0.1433	0.1297	0.3200	0.2685	0.3490
2500	% of γs ≥ 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Ave. # Inst's	13	10	9	7	7	38	28	19	14	15	75	56	29	21	23
	Variance	0.0025	0.0028	0.0025	0.0029	0.0026	0.0010	0.0018	0.0011	0.0013	0.0011	0.0006	0.0011	0.0007	0.0008	0.0007
	Bias	-0.0041	-0.0033	-0.0017	-0.0009	-0.0023	-0.0057	-0.0128	-0.0015	0.0003	-0.0014	-0.0081	-0.0155	-0.0013	0.0002	-0.0009
	RMSE	0.0503	0.0534	0.0500	0.0541	0.0506	0.0328	0.0442	0.0333	0.0363	0.0338	0.0267	0.0360	0.0264	0.0291	0.0264
	Size H_0	0.0540	0.0550	0.0525	0.0515	0.0525	0.0625	0.0805	0.0590	0.0570	0.0600	0.0750	0.0945	0.0525	0.0500	0.0470
	Power H_a	0.5560	0.5105	0.5405	0.4935	0.5455	0.9045	0.7865	0.8570	0.7815	0.8505	0.9875	0.9415	0.9635	0.9220	0.9630
	Power H_b	0.5045	0.4730	0.5190	0.4500	0.4990	0.8525	0.6070	0.8690	0.8025	0.8550	0.9630	0.7875	0.9740	0.9425	0.9725

[♭] Estimation methods: (1) all linear instruments, (2) all linear instruments + Mehrhoff's method utilizing 95% of the variation of the instruments, (3) two lags of y_{it}, x_{it}, and the constant, (4) two lags of y_{it}, x_{it}, and the constant + Mehrhoff's method utilizing 95% of the variation of the instruments, (5) two lags of y_{it}, one lag of x_{it}, and the constant. [♯] rounded to the nearest integer. [*] $H_0 : \rho = 0.5$. [†] $H_a : \rho = 0.6$. [‡] $H_b : \rho = 0.4$ (5% level).

Table 11. Small Sample Results for GMM Estimation with Reduced Number of Instruments (β results).

β		4					6					8				
	Est. Method$^\flat$	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
250	Ave. # Inst's$^\sharp$	13	10	9	7	7	38	28	19	14	15	75	55	29	21	23
	Variance	0.0101	0.0102	0.0110	0.0116	0.0193	0.0047	0.0045	0.0053	0.0054	0.0076	0.0035	0.0032	0.0037	0.0037	0.0051
	Bias	0.0024	0.0037	0.0054	0.0072	0.0140	-0.0172	-0.0160	-0.0041	-0.0028	-0.0086	-0.0323	-0.0277	-0.0028	-0.0012	-0.0094
	RMSE	0.0942	0.0945	0.1011	0.1044	0.1363	0.0653	0.0618	0.0714	0.0725	0.0867	0.0607	0.0543	0.0600	0.0598	0.0716
	Size H_0^*	0.0800	0.0787	0.0732	0.0702	0.0623	0.1450	0.1232	0.0839	0.0748	0.0778	0.2732	0.2108	0.0984	0.0845	0.0960
	Power $H_a^\dagger$	0.2564	0.2470	0.2130	0.2041	0.1293	0.5737	0.5464	0.3980	0.3853	0.3122	0.8258	0.7825	0.5379	0.5015	0.4426
	Power $H_b^\ddagger$	0.1940	0.1922	0.1751	0.1741	0.1387	0.3658	0.3379	0.3199	0.3079	0.2272	0.4664	0.4357	0.4575	0.4370	0.3300
500	Ave. # Inst's	13	10	9	7	7	38	28	19	14	15	75	55	29	21	23
	Variance	0.0039	0.0039	0.0043	0.0044	0.0063	0.0021	0.0020	0.0023	0.0023	0.0032	0.0016	0.0015	0.0017	0.0017	0.0023
	Bias	0.0016	0.0021	0.0033	0.0033	0.0045	-0.0040	-0.0045	0.0017	0.0019	-0.0001	-0.0128	-0.0133	-0.0008	-0.0006	-0.0041
	RMSE	0.0609	0.0606	0.0648	0.0658	0.0789	0.0448	0.0432	0.0477	0.0480	0.0565	0.0406	0.0384	0.0414	0.0417	0.0478
	Size H_0	0.0660	0.0631	0.0608	0.0558	0.0564	0.0875	0.0715	0.0662	0.0612	0.0625	0.1376	0.1196	0.0750	0.0720	0.0740
	Power H_a	0.4081	0.4027	0.3650	0.3568	0.2553	0.7185	0.7171	0.5880	0.5789	0.4712	0.8842	0.8904	0.7385	0.7195	0.6480
	Power H_b	0.4023	0.4012	0.3671	0.3553	0.2618	0.6500	0.6401	0.6075	0.5930	0.4572	0.7399	0.7251	0.7300	0.7215	0.5680
2500	Ave. # Inst's	13	10	9	7	7	38	28	19	14	15	75	56	29	21	23
	Variance	0.0008	0.0008	0.0009	0.0009	0.0013	0.0004	0.0004	0.0005	0.0005	0.0007	0.0003	0.0003	0.0004	0.0004	0.0005
	Bias	0.0006	0.0006	0.0008	0.0008	0.0018	0.0006	0.0002	0.0010	0.0012	0.0008	0.0005	0.0001	0.0013	0.0014	0.0006
	RMSE	0.0277	0.0276	0.0294	0.0294	0.0365	0.0206	0.0205	0.0222	0.0222	0.0263	0.0175	0.0174	0.0190	0.0191	0.0220
	Size H_0	0.0545	0.0515	0.0565	0.0540	0.0490	0.0620	0.0610	0.0595	0.0585	0.0565	0.0660	0.0585	0.0640	0.0680	0.0605
	Power H_a	0.9400	0.9415	0.9145	0.9065	0.7485	0.9975	0.9975	0.9915	0.9905	0.9620	1.0000	1.0000	0.9995	0.9995	0.9955
	Power H_b	0.9675	0.9665	0.9400	0.9385	0.7930	0.9990	0.9985	0.9950	0.9960	0.9685	1.0000	1.0000	1.0000	1.0000	0.9965

$\flat$ Estimation methods: (1) all linear instruments, (2) all linear instruments + Mehrhoff's method utilizing 95% of the variation of the instruments, (3) two lags of y_{it}, x_{it}, and the constant, (4) two lags of y_{it}, x_{it}, and the constant + Mehrhoff's method utilizing 95% of the variation of the instruments, (5) two lags of y_{it}, one lag of x_{it}, and the constant. $^\sharp$ rounded to the nearest integer. * $H_0 : \beta = 0.3188$. † $H_a : \beta = 0.4188$. ‡ $H_b : \beta = 0.2188$ (5% level).

calculated as,

$$
\begin{aligned}
APEX\left(y_{i,t-1}=1,\, x_{it}=\bar{x}\right) &= e^{\beta_l' \bar{x}+\rho_l}\beta_l \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{e^{c_{il}}}{\left(1+e^{c_{il}+\beta_l\bar{x}+\rho_l}\right)^2},\\
APEX\left(y_{i,t-1}=0,\, x_{it}=\bar{x}\right) &= e^{\beta_l \bar{x}}\beta_l \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{e^{c_{il}}}{\left(1+e^{c_{il}+\beta_l\bar{x}}\right)^2},\\
APEY\left(x_{it}=\bar{x}\right) &= e^{\beta_l \bar{x}}\left(e^{\rho_l}-1\right) \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left[\frac{e^{c_{il}}}{\left(1+e^{c_{il}+\beta_l\bar{x}}\right)\left(1+e^{c_{il}+\rho_l+\beta_l\bar{x}}\right)}\right],
\end{aligned}
$$

where the averages over i are obtained by drawing from the distribution of c_{il}. That is, the average partial effects are obtained by stochastic integration over c_{il}.

Now suppose that data from this logistic DGP are used to estimate ρ_e and β_e using the GMM procedure we have outlined above (i.e. based on the exponential specification). The question is, how well do these estimates reproduce the (true) average partial effects given above for the logistic specification? To answer this question, we must first specify how the fixed effects of the exponential specification are to be computed. We do this by deriving fixed effects under exponential specification, c_{ie}, in terms of the fixed effects of the true logistic specification, c_{il}, by matching the transitions from 0 to 1 given $x_{it}=\bar{x}_i=\frac{1}{T}\sum_t x_{it}$ across the two specifications, namely[13]

$$
1-e^{-c_{ie}-\beta_e\bar{x}_i}=\frac{e^{c_{il}+\beta_l\bar{x}_i}}{1+e^{c_{il}+\beta_l\bar{x}_i}},
$$

which yields

$$
e^{-c_{ie}}=\frac{e^{\beta_e\bar{x}_i}}{1+e^{c_{il}+\beta_l\bar{x}_i}}.
$$

We may then estimate the average partial effects as

$$
\begin{aligned}
\widehat{APEX}\left(y_{i,t-1}=1,\, x_{it}=\bar{x}\right) =& \widehat{\beta_e} e^{-\widehat{\rho}_e-\widehat{\beta}_e\bar{x}} \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\hat{\beta}_e\bar{x}_i}}{1+e^{c_{il}+\beta_l\bar{x}_i}},\\
\widehat{APEX}\left(y_{i,t-1}=0,\, x_{it}=\bar{x}\right) =& \widehat{\beta_e} e^{-\widehat{\beta}_e\bar{x}} \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\hat{\beta}_e\bar{x}_i}}{1+e^{c_{il}+\beta_l\bar{x}_i}},\\
\widehat{APEY}\left(x_{it}=\bar{x}\right) =& e^{-\hat{\beta}_e\bar{x}}(1-e^{-\hat{\rho}_e}) \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\hat{\beta}_e\bar{x}_i}}{1+e^{c_{il}+\beta_l\bar{x}_i}}.
\end{aligned}
$$

[13] It is also possible to match the transitions from 1 to 1 given $x_{it}=\bar{x}_i$. This gives slightly different exponential fixed effects. But it does not change the general conclusion of this section. The results that condition on $\bar{x}_i$ are available from the authors on request.

The benchmark APE results are computed under the logistics model employed by Honoré and Kyriazidou (2000), where $\rho_l = 0.5$, $\beta_l = 1$, $x_{it} \sim N(0, \pi^2/3)$, and $c_{il} \sim N(0,1)$. To avoid any complications with initial conditions, the data are burned in for the first 100 periods in each replication, while being careful to keep x_{it} fixed across replications. The simulations are based on $N = 1,000$, $T = 3$, and each experiment is repeated $2,000$ times to obtain the mean, variance, bias, and RMSE of the APEs. We vary the DGP and the data sets in a variety of ways (see Table 12).

The results indicate that the average partial effects obtained using the exponential specification, with matched fixed effects as explained above, are close to the true average partial effects. In particular, the $\widehat{APEY}$ is typically quite close to $APEY$. This provides further evidence of the robustness of the exponential specification in that it yields sensible estimates for the average partial effects even when the exponential distribution is misspecified. In fact, the same exercise was conducted using a probit distribution. The results were similar and in all cases matched the sign of the true partial effects, although they showed a greater degree of bias.[14]

6 Conclusion

In this paper we consider identification and estimation of dynamic binary response panel data models. We develop an exponential class of models and derive CML and GMM estimators that enable us to eliminate the unobserved heterogeneity and at the same time to identify the model parameters. We show that for the exponential family of distributions that we consider the GMM approach is more generally applicable and yields consistent and $\sqrt{N}$ asymptotically normal estimators for dynamic models with and without covariates. But in the case of exponentially distributed errors the CML approach can only identify the state dependence parameter and cannot identify the parameters of the covariates. The GMM approach proposed here is simple, general, and offers several advantages over the existing estimators that will be particularly appealing for analyzing microeconomic panel data from a dynamic perspective.

As is well known, it is important to use a dynamic binary choice specification to model the state dependence in a panel setting because of the model's ability to distinguish the state dependence from the unobserved heterogeneity among other useful features. The dynamic binary choice models,

[14] To save space the results for the probit distribution are available in an online supplement.

Table 12. Logistic vs. Implied Exponential Average Partial Effects.

Experiment	1	2	3	4	5	6	7	8	9	10	11
$APEX_1$	0.1987	0.1891	0.2050	0.2983	0.0993	0.1721	0.2238	0.1988	0.1987	0.2009	0.1984
Mean $\widehat{APEX}_1$	0.1349	0.1271	0.1425	0.1543	0.0860	0.1305	0.1360	0.1158	0.1564	0.1375	0.1380
Variance $\widehat{APEX}_1$	0.0003	0.0002	0.0003	0.0004	0.0001	0.0002	0.0003	0.0002	0.0004	0.0000	0.0001
Bias $\widehat{APEX}_1$	-0.0639	-0.0620	-0.0625	-0.1441	-0.0134	-0.0417	-0.0877	-0.0831	-0.0423	-0.0634	-0.0605
RMSE $\widehat{APEX}_1$	0.0659	0.0640	0.0647	0.1453	0.0177	0.0444	0.0893	0.0842	0.0464	0.0637	0.0610
$APEX_0$	0.2075	0.2075	0.2075	0.3112	0.1037	0.1780	0.2364	0.2075	0.2075	0.2078	0.2068
Mean $\widehat{APEX}_0$	0.1630	0.1657	0.1602	0.1906	0.1031	0.1483	0.1751	0.1421	0.1869	0.1645	0.1643
Variance $\widehat{APEX}_0$	0.0002	0.0002	0.0002	0.0003	0.0001	0.0001	0.0003	0.0002	0.0003	0.0000	0.0001
Bias $\widehat{APEX}_0$	-0.0445	-0.0418	-0.0473	-0.1206	-0.0007	-0.0297	-0.0613	-0.0654	-0.0206	-0.0433	-0.0425
RMSE $\widehat{APEX}_0$	0.0467	0.0443	0.0492	0.1219	0.0112	0.0321	0.0634	0.0667	0.0268	0.0437	0.0431
$APEY$	0.1022	0.1507	0.0516	0.1022	0.1021	0.0879	0.1160	0.1022	0.1022	0.1028	0.1019
Mean $\widehat{APEY}$	0.0777	0.1052	0.0494	0.0802	0.0791	0.0546	0.0994	0.0802	0.0760	0.0768	0.0765
Variance $\widehat{APEY}$	0.0019	0.0019	0.0019	0.0017	0.0020	0.0021	0.0018	0.0018	0.0020	0.0002	0.0003
Bias $\widehat{APEY}$	-0.0245	-0.0456	-0.0023	-0.0220	-0.0231	-0.0333	-0.0167	-0.0220	-0.0262	-0.0261	-0.0255
RMSE $\widehat{APEY}$	0.0500	0.0633	0.0440	0.0467	0.0506	0.0563	0.0457	0.0477	0.0517	0.0303	0.0315

The average partial effects are $APEX_1 = \int \frac{\partial P[y_{it}=1|y_{i,t-1}=1,c_{il},x_{it}]}{\partial x_{it}}\Big|_{x_{it}=\bar{x}} dF_c(c_i)$, $APEX_0 = \int \frac{\partial P[y_{it}=1|y_{i,t-1}=0,c_{il},x_{it}]}{\partial x_{it}}\Big|_{x_{it}=\bar{x}} dF_c(c_i)$, and $APEY = \int \left(P[y_{it} = 1\,|y_{i,t-1} = 1, c_{il}, x_{it} = \bar{x}] - P[y_{it} = 1\,|y_{i,t-1} = 0, c_{il}, x_{it} = \bar{x}]\right) dF_c(c_i)$, where F_c is the distribution function of the fixed effects. See the discussion above for the calculation and estimation of these quantities.

The simulations are as follows: (1) the benchmark, (2) ρ increased to 0.75, (3) ρ decreased to 0.25, (4) β increased to 1.5, (5) β decreased to 0.5, (6) σ_c increased to 1.5, (7) σ_c decreased to 0.5, (8) σ_x increased by 0.5, (9) σ_x decreased by 0.5, (10) N increased to 10,000, (11) T increased to 8. Parameters are estimated using the full set of linear instruments.

however, have been rarely used in analyzing microeconomic data, mainly due to the problems associated with the initial condition in combination with the incidental parameter problems. Our approach based on the exponential specification resolves the incidental parameter problem and the resulting estimators can be readily implemented, and also have good asymptotic properties.

Both the GMM and the CML estimators perform well under a variety of scenarios. Our results show that the estimators are robust to changes in the variance of the fixed effects, different values of ρ and $\boldsymbol{\beta}$, correlation between the fixed effects and the regressors, heterogeneity in the regressors across the different units, and autocorrelation in the regressors. In each of the experiments, we considered bias, variance, RMSE, size, and power of the GMM estimators. GMM worked quite well for relatively small sample sizes. We also tested the CMLE and compared its performance to the GMM estimator. Interestingly, GMM emerges as a better estimator than CMLE for small values of T (when $\beta = 0$ and both estimators are applicable). In the case of large T we experimented with the moment reduction techniques of Mehrhoff (2009) finding significant improvements in performance in small samples. We also presented evidence of the ability of the exponential specification to match the average partial effects from a logistic dynamic binary choice model.

7 Appendix

7.1 Proof of the Uniqueness of the Exponential Distribution

Proposition A1: Suppose F is a differentiable cumulative distribution function. If there exist functions G and H such that $F(x+y) - F(x) = G(y)H(x)$ then $F = 1 - C\exp(-Dx)$ for some positive constants C and D.

Proof: Assume without loss of generality that $\text{sgn}(G(y)) = \text{sgn}(y)$ and H is non–negative. Now take the limit as $y \to \infty$. Then $A = \lim_{y\to\infty} G(y)$ exists and $1 - F(x) = AH(x)$. Since F is a cumulative distribution function, it is non–constant and so $A \neq 0$. In particular, the non–negativity of G over positive real numbers implies that $A > 0$. This now implies that $F(x+y) - F(x) = A^{-1}(1 - F(x))G(y)$. Divide both sides by y and take the limit as $y \to 0$. The differentiability of F implies that $B = \lim_{y\to 0} G(y)/y$ exists and $F'(x) = \frac{B}{A}(1 - F(x))$. Since F is non–decreasing and bounded by 0 and 1, the sign of B cannot be negative. Since F is also non–constant $B \neq 0$ so we must have $B > 0$. The final step is to note that we have arrived at a differential equation in x that can be solved as, $F(x) = 1 - C\exp(-\frac{B}{A}x)$ for some constant C. Again, since F is a cumulative distribution function, we must have $C > 0$.

7.2 GMM in the case where $\beta = 0$ and $T = 3$

In the case where $T = 3$ we only have one moment condition with which to estimate γ (or ρ), namely

$$\sum_{i=1}^{N} e_{i3}(\gamma) y_{i1} = \sum_{i=1}^{N} y_{i1} \left[\frac{(y_{i3} - \gamma y_{i2})(1 - \gamma y_{i1})}{(1 - \gamma y_{i2})} - (y_{i2} - \gamma y_{i1}) \right] = 0. \tag{21}$$

Note that $e_{i3}(\gamma)$ does not depend on γ if $y_{i1} + y_{i2} + y_{i3} = 0$ or $= 3$. Consider now the case where $y_{i1} + y_{i2} + y_{i3} = 2$, and note further that observations where $y_{i1} = 0$ and $y_{i2} = y_{i3} = 1$ can be dropped since $y_{i1} e_{i3}(\gamma) = 0$. The other remaining cases are $(y_{i1}, y_{i2}, y_{i3}) = (1, 0, 0), (1, 1, 0)$, and $(1, 0, 1)$. Denote the number of cross section units associated with these patterns of observations over time by n_{100}, n_{110} and n_{101}, respectively. Then the moment condition in γ can be written as

$$n_{100} \hat{\gamma}_{GMM,1} - n_{110} + n_{101} = 0.$$

Hence, if $n_{100} \neq 0$

$$\hat{\gamma}_{GMM,1} = \frac{n_{110} - n_{101}}{n_{100}}.$$

An estimate for ρ can be obtained if $n_{110} < n_{100} + n_{101}$.

In the case where $n_{100} = 0$, the above GMM estimator is not valid. But since $E(e_{it} \left| y_{i,t-s} \right.) = 0$, we also have unconditionally that $E(e_{it}) = 0$. This suggests the following sample moment condition

$$\sum_{i=1}^{N} \left[\frac{(y_{i3} - \gamma y_{i2})(1 - \gamma y_{i1})}{(1 - \gamma y_{i2})} - (y_{i2} - \gamma y_{i1}) \right] = 0. \tag{22}$$

Once again we only need to consider observations where $y_{i1} + y_{i2} + y_{i3} = 1$ or $y_{i1} + y_{i2} + y_{i3} = 2$. Then we have

$$n_{100}\gamma - \frac{1}{1-\gamma} n_{010} + n_{001} + n_{101} - n_{110} = 0, \tag{23}$$

$$-n_{100}\gamma^2 + (n_{100} + n_{110} - n_{001} - n_{101})\gamma + n_{001} + n_{101} - n_{110} - n_{010} = 0. \tag{24}$$

Preliminary analysis suggests that the solutions to (24) could be complex, and when real could fall outside the range $[0, 1)$, and hence might not yield sensible estimates for ρ. It is, therefore, more meaningful to use the unconditional moment condition only when $n_{100} = 0$. In this case the solution to the unconditional moment condition is unique and is given by (obtained by setting n_{100} in (23) zero)

$$\hat{\gamma}_{GMM,2} = 1 - \frac{n_{101}}{n_{001} + n_{101} - n_{110}}.$$

Hence, in general we could estimate γ by

$$\begin{aligned} \hat{\gamma}_{GMM} &= \frac{n_{110} - n_{101}}{n_{100}}, \text{ if } n_{100} \neq 0, \\ &= 1 - \frac{n_{101}}{n_{001} + n_{101} - n_{110}}, \text{ if } n_{100} = 0. \end{aligned}$$

7.3 CMLE in the Case where $\beta = 0$ and $T = 3$

Suppose we have observations y_{i1}, y_{i2} and y_{i3} on N individual units. Denote the set of all observations such that $y_{i1} + y_{i2} + y_{i3} = 1$ by $\mathcal{B}$ and define the sets

$$\begin{aligned} \mathcal{A}_1 &= \{y_{i1} = 1, y_{i2} = 0, y_{i3} = 0\}, \\ \mathcal{A}_2 &= \{y_{i1} = 0, y_{i2} = 1, y_{i3} = 0\}, \\ \mathcal{A}_3 &= \{y_{i1} = 0, y_{i2} = 0, y_{i3} = 1\}. \end{aligned}$$

It is now easily seen that (given the Markov property and (4))

$$\begin{aligned} \Pr(\mathcal{A}_1) &= \Pr(y_{i1} = 1) \Pr(y_{i2} = 0 \,|y_{i1} = 1) \Pr(y_{i3} = 0 \,|y_{i2} = 0) \\ &= \pi_i^* \left[1 - F(c_i + \rho)\right] \left[1 - F(c_i)\right] \\ &= \frac{F(c_i) \left[1 - F(c_i + \rho)\right] \left[1 - F(c_i)\right]}{1 - F(c_i + \rho) + F(c_i)}. \end{aligned}$$

Similarly

$$\begin{aligned} \Pr(\mathcal{A}_2) &= \frac{F(c_i) \left[1 - F(c_i + \rho)\right]^2}{1 - F(c_i + \rho) + F(c_i)}, \\ \Pr(\mathcal{A}_3) &= \frac{\left[1 - F(c_i + \rho)\right] \left[1 - F(c_i)\right] F(c_i)}{1 - F(c_i + \rho) + F(c_i)}, \end{aligned}$$

and

$$\Pr(\mathcal{B}) = \Pr(\mathcal{A}_1) + \Pr(\mathcal{A}_2) + \Pr(\mathcal{A}_3).$$

Also

$$\Pr(\mathcal{A}_i) = \Pr(\mathcal{A}_i \cap \mathcal{B}) = \Pr(\mathcal{B}) \Pr(\mathcal{A}_i \,|\mathcal{B}\,),$$

and

$$\Pr(\mathcal{A}_i \,|\mathcal{B}\,) = \frac{\Pr(\mathcal{A}_i)}{\Pr(\mathcal{B})} \text{ for } i = 1, 2, 3.$$

Hence

$$\begin{aligned} \Pr(\mathcal{A}_1 \,|\mathcal{B}\,) &= \frac{\left[1 - F(c_i)\right]}{\left[1 - F(c_i + \rho)\right] + 2\left[1 - F(c_i)\right]}, \\ \Pr(\mathcal{A}_2 \,|\mathcal{B}\,) &= \frac{\left[1 - F(c_i + \rho)\right]}{\left[1 - F(c_i + \rho)\right] + 2\left[1 - F(c_i)\right]}, \\ \Pr(\mathcal{A}_3 \,|\mathcal{B}\,) &= 1 - \Pr(\mathcal{A}_1 \,|\mathcal{B}\,) - \Pr(\mathcal{A}_2 \,|\mathcal{B}\,). \end{aligned}$$

In the exponential case, $1 - F(c_i) = \exp(-c_i)$ and $1 - F(c_i + \rho) = \exp(-c_i - \rho)$, and

$$\begin{aligned} \Pr(\mathcal{A}_1 \,|\mathcal{B}\,) &= \frac{1}{\exp(-\rho) + 2}, \ \Pr(\mathcal{A}_2 \,|\mathcal{B}\,) = \frac{\exp(-\rho)}{\exp(-\rho) + 2}, \\ \Pr(\mathcal{A}_3 \,|\mathcal{B}\,) &= \frac{1}{\exp(-\rho) + 2}, \end{aligned}$$

which do not depend on the incidental parameters. It is clear that conditioning on $y_{i1}+y_{i2}+y_{i3} = 0$ and $y_{i1}+y_{i2}+y_{i3} = 3$ will not help. It only remains to consider the case where the conditioning set is $y_{i1} + y_{i2} + y_{i3} = 2$. Denoting

$$\begin{aligned} \mathcal{C}_1 &= \{y_{i1} = 1, y_{i2} = 1, y_{i3} = 0\}, \ \mathcal{C}_2 = \{y_{i1} = 0, y_{i2} = 1, y_{i3} = 1\}, \\ \mathcal{C}_3 &= \{y_{i1} = 1, y_{i2} = 0, y_{i3} = 1\}, \ \mathcal{D} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 = \{y_{i1} + y_{i2} + y_{i3} = 2\}. \end{aligned}$$

It is easily seen that

$$\begin{aligned}\Pr(\mathcal{C}_1\,|\mathcal{D}\,) &= \frac{F(\rho+c_i)}{2F(\rho+c_i)+F(c_i)},\ \Pr(\mathcal{C}_2\,|\mathcal{B}\,)=\frac{F(\rho+c_i)}{2F(\rho+c_i)+F(c_i)},\\ \Pr(\mathcal{C}_3\,|\mathcal{B}\,) &= \frac{F(c_i)}{2F(\rho+c_i)+F(c_i)}.\end{aligned}$$

These conditional probabilities depend on c_i even if $F(\cdot)$ has an exponential form. Consequently, the only appropriate conditioning is $y_{i1}+y_{i2}+y_{i3}=1$.

The conditional likelihood function for the exponential model is given by

$$\begin{aligned}L_c(\rho) &= \prod_{i\in\mathcal{B}}\left(\frac{1}{\exp(-\rho)+2}\right)^{y_{i1}+y_{i3}}\prod_{i\in\mathcal{B}}\left(\frac{\exp(-\rho)}{\exp(-\rho)+2}\right)^{y_{i2}}\\ &= \prod_{i\in\mathcal{B}}\left(\frac{1}{\exp(-\rho)+2}\right)^{y_{i1}+y_{i2}+y_{i3}}\prod_{i\in\mathcal{B}}\left(\exp(-\rho)\right)^{y_{i2}},\end{aligned}$$

and

$$\begin{aligned}\ln L_c(\rho) &= -\sum_{i\in\mathcal{B}}\ln\left[\exp(-\rho)+2\right]-\rho\sum_{i\in\mathcal{B}}y_{i2} \qquad (25)\\ &= -\ln\left[\exp(-\rho)+2\right]\sum_{i=1}^{N}I(y_{i1}+y_{i2}+y_{i3}=1)-\rho\sum_{i=1}^{N}y_{i2}I(y_{i1}+y_{i2}+y_{i3}=1),\end{aligned}$$

where $I(A)=1$ is A is true and $I(A)=0$ if A is not true. The conditional log-likelihood function can be written more compactly as

$$\ln L_c(\rho)=n_{\mathcal{B}}\left\{-\ln\left[\exp(-\rho)+2\right]-\rho\;\hat{p}\right\},$$

where $n_{\mathcal{B}}=\sum_{i=1}^{N}I(y_{i1}+y_{i2}+y_{i3}=1)$, and

$$\hat{p}=\frac{\sum_{i=1}^{N}y_{i2}I(y_{i1}+y_{i2}+y_{i3}=1)}{\sum_{i=1}^{N}I(y_{i1}+y_{i2}+y_{i3}=1)}=\frac{\sum_{i=1}^{N}I(y_{i1}=0,y_{i2}=1,y_{i3}=0)}{\sum_{i=1}^{N}I(y_{i1}+y_{i2}+y_{i3}=1)}.$$

Also since

$$\frac{\partial\ln L_c(\rho)}{\partial\rho}=n_{\mathcal{B}}\left\{\frac{\exp(-\rho)}{2+\exp(-\rho)}-\;\hat{p}\right\},$$

then the conditional maximum likelihood estimator of ρ is given by

$$\hat{\rho}=-\ln\left(\frac{2\hat{p}}{1-\hat{p}}\right). \qquad (26)$$

The standard error for $\hat{\rho}$ can be obtained using the second derivative of the conditional log-likelihood function. We have

$$Var(\hat{\rho})=\frac{1}{n_{\mathcal{B}}}\frac{\left[2+\exp(-\rho)\right]^2}{2\exp(-\rho)}.$$

7.4 Proof of Theorem 1

Given assumption (A3), and using (12) we have

$$\pi_i^*=\Pr(y_{i1}=1|c_i)=\frac{1-e^{-c_i}}{1-e^{-c_i}(1-e^{-\rho_0})},$$

and it is evident that this choice of initial distribution makes y_{it} stationary conditional on c_i. Thus $\pi_i^* = \Pr(y_{it} = 1|c_i) = \Pr(y_{i1} = 1|c_i)$ for $t \geq 1$. To simplify notation we first note that e_{it} defined by (15) can also be written as:[15]

$$e_{it} = e^{\rho \Delta y_{i,t-1}}(y_{it} - 1) + 1 - y_{i,t-1}.$$

Let $f_i(\rho) = e_{it} y_{i,t-2}$, and note that

$$\begin{aligned} E\left[e^{\rho\Delta y_{i,t-1}}(y_{it}-1)y_{i,t-2}\right] &= E\left[E(y_{it}-1|c_i, y_{i,t-1}, y_{i,t-2}, \ldots)e^{\rho\Delta y_{i,t-1}}y_{i,t-2}\right] \\ &= -E(e^{-c_i-\rho_0 y_{i,t-1}}e^{\rho\Delta y_{i,t-1}}y_{i,t-2}) \\ &= -E(e^{-c_i-(\rho_0-\rho)y_{i,t-1}-\rho y_{i,t-2}}y_{i,t-2}) \\ &= -E\left[E(e^{-(\rho_0-\rho)y_{i,t-1}}|c_i, y_{i,t-2}, y_{i,t-3}, \ldots)e^{-c_i-\rho y_{i,t-2}}y_{i,t-2}\right] \\ &= -E\left[(e^{-(\rho_0-\rho)}(1-e^{-c_i-\rho_0 y_{i,t-2}}) + e^{-c_i-\rho_0 y_{i,t-2}})e^{-c_i-\rho y_{i,t-2}}y_{i,t-2}\right] \\ &= -E(e^{-c_i-(\rho_0-\rho)-\rho y_{i,t-2}}y_{i,t-2} - e^{-2c_i-(\rho_0-\rho)-(\rho+\rho_0)y_{i,t-2}}y_{i,t-2} \\ &\qquad + e^{-2c_i-(\rho+\rho_0)y_{i,t-2}}y_{i,t-2}) \\ &= -e^{-\rho_0}E(e^{-c_i}\pi_i^*) + e^{-2\rho_0}E(e^{-2c_i}\pi_i^*) - e^{-\rho_0-\rho}E(e^{-2c_i}\pi_i^*). \end{aligned}$$

Also

$$\begin{aligned} E\left[(1-y_{i,t-1})y_{i,t-2}\right] &= E\left[E(1-y_{i,t-1}|c_i, y_{i,t-2}, y_{i,t-3}, \ldots)y_{i,t-2}\right] \\ &= E(e^{-c_i-\rho_0 y_{i,t-2}}y_{i,t-2}) = e^{-\rho_0}E(e^{-c_i}\pi_i^*). \end{aligned}$$

Summing up we obtain

$$E\left[f_i(\rho)\right] = (e^{-\rho_0} - e^{-\rho})e^{-\rho_0}E(e^{-2c_i}\pi_i^*).$$

Clearly $E(e^{-2c_i}\pi_i^*) \leq 1$. On the other hand, using assumption (A1),

$$E(e^{-2c_i}\pi_i^*) = E\left(\frac{e^{-2c_i}(1-e^{-c_i})}{1-e^{-ci}(1-e^{-\rho_0})}\right) \geq E\left(\frac{1}{2}e^{-2c_i}(1-e^{-c_i})\right) \geq \frac{K}{2}\Pr(e^{-2c_i}(1-e^{-c_i}) \geq K).$$

Assumption (A1) implies that $0 < e^{-c_i} < 1$ almost surely, thus it is possible to choose K so that the right hand side is positive. Thus $E\left[f_i(\rho)\right]$ is continuous in ρ and equals zero if and only if $\rho = \rho_0$. This satisfies Assumption 1.1 of Harris and Mátyás (1999).

Consider now $f_i'(\rho) = e^{\rho\Delta y_{i,t-1}}\Delta y_{i,t-1}(y_{it}-1)y_{i,t-2}$, which is clearly continuous and bounded by $e^{\max(R)}$ for all $\rho \in R$. It follows that,

$$|f_i(\rho) - f_i(\rho')| \leq e^{\max(R)}|\rho - \rho'|,$$

for all $\rho, \rho' \in R$ and so f is Lipschitz. This, together with assumptions (A4) and (A5) implies that $N^{-1}\sum_{i=1}^{N} f_i(\rho)$ converges uniformly to $E\left[f_i(\rho)\right]$ by Corollary 3.1 of Newey (1991).[16] This satisfies Assumption 1.2 of Harris and Mátyás (1999) and it follows from the their Theorem 1.1 that $\widehat{\rho}$ is consistent.

[15] Since $\gamma = 1 - \exp(-\rho)$, and because $y_{i,t-1}$ and $y_{i,t-2}$ take 0 and 1 values only, then it is easily verified that $(1 - \gamma y_{i,t-2}) / (1 - \gamma y_{i,t-1})$ and $e^{\rho\Delta y_{i,t-1}}$ give the same values for all admissible choices of $y_{i,t-1}$ and $y_{i,t-2}$.

[16] See the discussion in Harris and Mátyás (1999) pp. 14-17.

The continuity of $f_i'(\rho)$ satisfies Assumption 1.7 of Harris and Mátyás (1999). $f_i''(\rho) = e^{\rho \Delta y_{i,t-1}}(\Delta y_{i,t-1})^2(y_{it} - 1)y_{i,t-2}$ is bounded again by $e^{\max(R)}$. Thus $f_i'(\rho)$ itself is Lifschitz and employing assumption (A6) it follows again from Newey (1991) that $N^{-1}\sum_{i=1}^{N} f_i'(\rho)$ converges uniformly to $E\left[f_i'(\rho)\right]$. By Theorem 4.1.5 of Amemiya (1985), $N^{-1}\sum_{i=1}^{N} f_i'(\widehat{\rho})$ converges to $Ef_i'(\rho_0)$. This satisfies Assumption 1.8 of Harris and Mátyás (1999).

Now let $i \neq j$. By Assumption (A2), $f_i(\rho)$ and $f_j(\rho)$ are independent conditional on c_i and c_j. Therefore, $E\left[f_i(\rho)f_j(\rho)\right] = E\left[E(f_i(\rho)|c_i, c_j)E(f_j(\rho)|c_i, c_j)\right]$. Assumption (A2) again implies that conditional on c_i, $f_i(\rho)$ is independent of c_j. Thus $E(f_i(\rho)|c_i, c_j) = E(f_i(\rho)|c_i)$. It follows that $E\left[f_i(\rho)f_j(\rho)\right] = E\left[E(f_i(\rho)|c_i)E(f_j(\rho)|c_j)\right]$. Since $E(f_i(\rho_0)|c_i) = 0$, we have that $E\left[f_i(\rho_0)f_j(\rho_0)\right] = 0$ for $i \neq j$ and so var $\left[N^{-1/2}\sum_{i=1}^{N} f_i(\rho_0)\right] = N^{-1}\sum_{i=1}^{N} E\left[f_i^2(\rho_0)\right]$. Thus Assumption (A7) implies the last necessary assumption of Harris and Mátyás (1999), namely their Assumption 1.9.

REFERENCES

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

Amemiya, T. (1985): *Advanced Econometrics.* Cambridge: Harvard University Press.

Andersen, E. B. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society*, Series B, 32, 283-301.

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

Arellano, M. (2003): "Discrete Choice with Panel Data." *Investigaciones Económicas*, XXVII (3), 423-458.

Arellano, M., and S. Bonhomme (2011): "Nonlinear Panel Data Analysis," *Annual Review of Economics*, 3, 395-424.

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 (2003): "Binary Choice Panel Data Models with Predetermined Variables," *Journal of Econometrics*, 115, 125-157.

Arellano, M., and B. Honoré (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, Vol. 5, ed. by J. Heckman and E. Leamer, Amsterdam: North-Holland.

Bartolucci, F., and V. Nigro (2010): "A Dynamic Model for Binary Panel Data with Unobserved Heterogeneity Admitting a $\sqrt{n}$-Consistent Conditional Estimator," *Econometrica*, 78, 719-733.

Bartolucci, F., and V. Nigro (2012): "Pseudo Conditional Maximum Likelihood Estimation of the Dynamic Logit Model for Binary Panel Data," *Journal of Econometrics*, 170, 102-116.

Bartolucci, F., R. Bellio, A. Salvan, and N. Sartori (2016): "Modified Profile Likelihood for Fixed-Effects Panel Data Models," *Econometric Reviews*, 35, 1271-1289.

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

Chamberlain, G. (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman and B. Singer, Cambridge: Cambridge University Press.

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

Chamberlain, G. (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78**,** 159-168.

Carrasco, M. (2012): "A Regularization Approach to the Many Instruments Problem," *Journal of Econometrics*, 170, 383-398.

Carro, J. M. (2007): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects," *Journal of Econometrics*, 140, 503-528.

Cox, D. R. (1958): "Some Problems Connected with Statistical Inference," *The Annals of Mathematical Statistics*, 29(2), 357-372.

Harris, D., and L. Mátyás (1999): "Introduction to the Generalised Method of Moments Estimation," in *Generalized Method of Moments Estimation*, ed. by L. Mátyás, Cambridge, U.K.: Cambridge University Press.

Hahn, J. (1999): "How Informative is the Initial Condition in the Dynamic Panel Data Model with Fixed Effects?" *Journal of Econometrics*, 93, 309-326.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica*, 50, 1029-1054.

Hayakawa, K. and M. H. Pesaran (2015), Robust standard errors in transformed likelihood estimation of dynamic panel data models with cross-sectional heteroskedasticity, *Journal of Econometrics*, 188, 111-134.

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

Heckman, J. (1981b): "Heterogeneity and State Dependence," in *Studies in Labor Markets*, ed. by S. Rosen, Chicago: University of Chicago Press.

Honoré, B. (2002): "Nonlinear Models with Panel Data," *Portuguese Economic Journal*, 1, 163-179.

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

Hsiao, C. (2014): *Analysis of Panel Data*, Third Edition, Cambridge, U.K.: Cambridge University Press.

Hsiao, C., M. H. Pesaran, and K. A. Tahmiscioglu (2002): "Maximum Likelihood Estimation of Fixed Effects Dynamic Panel Data Models Covering Short Time Periods," *Journal of Econometrics*, 109, 107–150.

Lee, Y. and P.C.B. Phillips (2015): "Model selection in the presence of incidental parameters," *Journal of Econometrics*, 188(2), 474-489.

Magnac, T. (2001): "Subsidised Training and Youth Employment: Distinguishing Unobserved Heterogeneity from State Dependence in Labour Market Histories," *Economic Journal*, 110**,** 805-837.

Magnac, T. (2004): "Panel Binary Variables and Sufficiency: Generalizing Conditional Logit," *Econometrica*, 72**,** 1859-1876.

Mehrhoff, J. (2009): "A Solution to the Problem of Too Many Instruments in Dynamic Panel Data GMM," Discussion Paper, Series 1, Economic Studies, Deutsche Bundesbank, Frankfurt.

Newey, W. K. (1991). "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59, 1161-1167.

Pesaran, M. H. (2015). *Time Series and Panel Data Econometrics.* Oxford, U.K.: Oxford University Press.

Pesaran, M. H., and A. Timermann (2009): "Testing Dependence Among Serially Correlated Multicategory Variables," *Journal of the American Statistical Association*, 104, 325-337.

Roodman, D. (2009): "A Note on the Theme of Too Many Instruments," *Oxford Bulletin of Economics and Statistics*, 71, 135-158.

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

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

Wooldridge, J. M. (2005): "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel-Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics,* 20, 39–54.