Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 140 (2007) 503–528
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

Estimating dynamic panel data discrete choice models with fixed effects

Jesus M. Carro*

Department of Economics, Universidad Carlos III de Madrid, C/Madrid, 126, 28903-GETAFE (Madrid), Spain

Available online 18 September 2006

Abstract

This paper considers the estimation of dynamic binary choice panel data models with fixed effects. It is shown that the modified maximum likelihood estimator (MMLE) used in this paper reduces the order of the bias in the maximum likelihood estimator from $O(T^{-1})$ to $O(T^{-2})$, without increasing the asymptotic variance. No orthogonal reparametrization is needed. Monte Carlo simulations are used to evaluate its performance in finite samples where T is not large. In probit and logit models containing lags of the endogenous variable and exogenous variables, the estimator is found to have a small bias in a panel with eight periods. A distinctive advantage of the MMLE is its general applicability. Estimation and relevance of different policy parameters of interest in this kind of models are also addressed.

JEL classification: C23

Keywords: Panel data; Dynamic discrete choice; Fixed effects; Modified MLE

1. Introduction

This paper deals with the estimation of dynamic discrete choice models with fixed effects. These models, that take into account permanent unobserved heterogeneity and the dynamic processes, are of interest in many empirical applications in economics, because they allow us to distinguish between the sources of the time persistence on individual decisions observed in discrete panel data sets. Observed persistence may be due to

*Tel.: +34 91 624 9586; fax: +34 91 624 9875.
E-mail address: jcarro@eco.uc3m.es.

0304-4076/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2006.07.023

persistence in observable individual characteristics, true state dependence or permanent unobserved heterogeneity. As explained in Heckman (1981a), these sources of persistence in individual decisions have very different implications.[1]

It is well-known that permanent unobserved heterogeneity may bias estimates and lead to misleading conclusions about the effect of a variable if it is not taken into account. This is particularly true in dynamic models where we can have significant estimates of state dependence coefficients even when there is no state dependence and persistence is only due to permanent heterogeneity. In the econometric literature, there are two ways of treating unobserved heterogeneity: random effects and fixed effects. This paper follows the fixed effects approach since we do not want to impose any restriction on the conditional distribution of the unobserved heterogeneity.[2]

Monte Carlo experiments have shown that the traditional maximum likelihood estimator (MLE) of non-linear panel data models with fixed effects generally exhibits considerable bias in finite samples when the number of periods is not large.[3] It is well-understood how to estimate, and solve that problem, in a linear model since fixed-T consistent estimators are available. However, there are no general solutions for non-linear models, and in some cases, although a specific solution is available, it is not $\sqrt{N}$-consistent. For example, Honoré and Kyriazidou (2000) propose a fixed-T consistent estimator for dynamic discrete choice models with continuous exogenous regressors that needs some restrictive assumptions and it is not $\sqrt{N}$-consistent. As a matter of fact, calculations in Hahn (2001) indicate that $\sqrt{N}$-consistent estimation is infeasible in this case. Furthermore, results in Honoré and Tamer (2004) suggest that the parameters of dynamic discrete choice panel data models are not identified in a fixed-T context, even in simple cases.

The estimation of non-linear models with fixed effects by maximum likelihood suffers the so-called incidental parameters problem. Cox and Reid (1987) considered the general problem of doing inference for a parameter of interest in the absence of knowledge about nuisance parameters.[4] Their solution is based on getting a re-parametrization such that the nuisance parameters are information orthogonal to the other parameters, so as to limit the influence of the nuisance parameters. Then they develop a modification that reduces the order of the bias of the MLE, without increasing its asymptotic variance. Their general framework has been used for static binary choice panel data models with fixed effects in Arellano (2003), where the fixed effects are the nuisance parameters. Arellano (2003) expresses the modification in terms of the original parameters.

Cox and Reid modification is known to fail without an orthogonal reparametrization (see, for example Ferguson et al., 1991). Arellano (2003) shows that the modification he derives reduces the order of the bias of the MLE, under the assumption that an information orthogonal reparametrization exists. The problem is that an information orthogonal reparametrization may not exist, especially for the kind of models considered here. This paper shows that the modified MLE (MMLE), using the modification expressed in terms of the original parameters of the model as in Arellano (2003), reduces the order of

[1] An economic example that exhibits substantial persistence over time is female labor force participation and knowing whether or not it reflects true state dependence, is needed for understanding the behavioral relationships underlying participation decisions (e.g. Hyslop, 1999).

[2] In particular, it is difficult to deal with the initial conditions problem in the random effects approach. See Honoré (2002) for a discussion on this and other issues on the comparison of the two approaches.

[3] See Heckman (1981b) for an example.

[4] Incidental parameters are nuisance parameters whose number grows with the sample size.

the bias regardless of the existence of an information orthogonal reparametrization. This result holds not only for dynamic binary choice models, but also for general non-linear likelihood-based panel data models. Asymptotic properties for different N and T plans are studied, and its performance in finite samples evaluated through Monte Carlo simulations.

Although this MMLE is only consistent when T goes to infinity, it is shown to be useful in the estimation of models like a probit with lags of the endogenous variable and exogenous variables in panels with just eight time periods. This is because its reduction of the order of the asymptotic bias of the MLE, leads to a small finite sample bias in cases where T is not very large, as shown by Monte Carlo simulations. The method gives a more general framework for the estimation of non-linear models with fixed effects, compared to the restrictive assumptions needed for using other estimators. For instance, it can be used without being restricted to the logistic case.

Lancaster (2002) and Woutersen (2001) apply the information orthogonality idea to integrated likelihood, following a Bayesian approach to the problem. Lancaster applies it to linear panel data models with fixed effects. Woutersen derives the general properties of the integrated likelihood estimator. Furthermore, he shows that all the properties derived for the integrated likelihood estimator also hold for the modified profile likelihood proposed by Cox and Reid (1987).

The rest of the paper is organized as follows. Next section presents the kind of models whose estimation is studied in this paper, the alternative approach that we try to address and its asymptotic properties. Section 3 shows some simulations of this alternative approach to study its performance in finite samples and its usefulness for the estimation of the policy parameters of interest. In contrast to the linear case, each one of the parameters defined in the model does not capture on his own the marginal effect of the explanatory variables. The effects are different for each individual, they depend on the fixed effects and there is more than one measure that should be considered. In Section 4, a female labor force participation model is estimated as an empirical illustration. The last section concludes.

2. The model and the MMLE

2.1. The model

Let us consider the following panel data model:

$$y_{it} = 1\{\alpha y_{it-1} + x'_{it}\beta + \eta_i + v_{it} \geqslant 0\} \quad (t = 0, \ldots, T-1; i = 1, \ldots, N), \tag{1}$$

where $1\{c\}$ takes value one if condition c is satisfied and zero otherwise. $\{\eta_i\}_{i=1,\ldots,N}$ describe permanent unobserved heterogeneity among individuals and v_{it} reflects unobserved random variables and shocks that individuals receive every period. As previously said, we do not want to impose any restrictive assumptions on the distribution of η_i, so we take a fixed effect approach and therefore treat $\{\eta_i\}_{i=1,\ldots,N}$ as parameters to be estimated. For any variable or set of variables z, z_{it} denotes observation at period t for individual i, $z_i = \{z_{i0}, \ldots, z_{iT-1}\}$, i.e. the set of all observations for individual i, and z_i^t are the set of observations from the first period to period t for individual i, $z_i^t = \{z_{i0}, \ldots, z_{it}\}$.

Assuming that v_{it} follows a parametric distribution, a natural way of estimating this model is by maximum likelihood; to write down the probability of the sample and maximize it in all the parameters: β, $\eta_1, \ldots, \eta_N$. By doing so, it raises the incidental

parameters problem, first considered by Neyman and Scott (1948). The intuition of the incidental parameters problem is clear in this case. Only new observations for individual i give new information about η_i and more individuals, i.e. increments in N, do not help with the estimation of η_i and add more parameters to be estimated. Therefore, the MLE of η_i is only consistent when $T \to \infty$. In the MLE of model (1), the inconsistency of the estimations of η_i is transmitted to the estimator of the other parameters. Conditioning on the first observation, the log-likelihood is

$$l(\gamma, \eta_1, \ldots, \eta_N) = \sum_{i=1}^{N} l_i(\gamma, \eta_i) = \sum_{i=1}^{N} \sum_{t=1}^{T-1} \{y_{it} * \log F_{it} + (1 - y_{it}) * \log(1 - F_{it})\}, \tag{2}$$

where it is assumed that—v_{it} is independently distributed with cdf F and $\gamma = (\alpha, \beta')'$. Deriving with respect to γ, $\eta_1, \ldots, \eta_N$, we get the first order conditions $d_{\eta_i}(\gamma, \eta_i) \equiv \partial l_i(\gamma, \eta_i)/\partial \eta_i$ and $\mathbf{d}_{\gamma i}(\gamma, \eta_i) \equiv \partial l_i(\gamma, \eta_i)/\partial \gamma$. MLE of η_i for given γ, $\widehat{\eta}_i(\gamma)$, solves $d_{\eta_i}(\gamma, \eta_i) = 0$. The MLE of γ is given by the maximizer of the so-called concentrated log-likelihood, $\sum_{i=1}^{N} l_i(\gamma, \widehat{\eta}_i(\gamma))$, which solves the following first order condition:

$$\frac{1}{TN} \sum_{i=1}^{N} \left\{ \mathbf{d}_{\gamma i}(\gamma, \widehat{\eta}_i(\gamma)) + d_{\eta_i}(\gamma, \widehat{\eta}_i(\gamma)) \frac{\partial \widehat{\eta}_i(\gamma)}{\partial \gamma} \right\} = \frac{1}{TN} \sum_{i=1}^{N} \mathbf{d}_{\gamma i}(\gamma, \widehat{\eta}_i(\gamma)) = 0. \tag{3}$$

This first order condition or estimating equation of γ depends on $\widehat{\eta}_i$, and evaluated at the true value, γ_0, does not converge to zero in probability when $N \to \infty$ for fixed T, since $\widehat{\eta}_i$ does not converge to its true value, η_{i0}.

This problem can be overcome if the estimator of γ can be derived so that it does not depend on the incidental parameters. A way of doing this is by conditioning on sufficient statistics for η_i. However, it is not possible to find sufficient statistics for many of the non-linear models used in econometrics. In particular, the logistic assumption is needed and, even with that assumption, model (1) does not have a sufficient statistic. Manski (1987) maximum score estimator is not restricted to a specific distributional assumption, but it imposes strict exogeneity on all explanatory variables, excluding dynamic models.[5]

Honoré and Kyriazidou (2000) consider the estimation of fixed effects discrete choice models like (1) and propose a fixed T consistent estimator. This estimator requires the logistic assumption, ε_{it} to be serially independent over time, and additional restrictions from the conditional approach they follow to eliminate the dependence on the fixed effect. In the case $T = 4$ those additional restrictions are x_{i2} to equal x_{i3} or $(x_{i2} - x_{i3})$ to be continuously distributed with support in a neighborhood of 0, and $(x_{i1} - x_{i2})$ to have sufficient variation conditional on the event that $x_{i2} - x_{i3} = 0$. These restrictions rule out time-dummies, for instance. The rate of convergence is slower than $N^{-1/2}$. Furthermore, the rate of convergence is decreasing as the number of regressors increases. If the logistic assumption is relaxed, Manski's insight is used and in addition to the former limitations, the objective function is not differentiable, which makes the maximization more difficult. So, for example, α and β in a dynamic probit of the form of model (1) do not have a good estimator.

[5]See Arellano and Honoré (2001) and Arellano (2003) for surveys on fixed T solutions for discrete choice models.

2.2. Modifying the score of the concentrated likelihood

The traditional approach to the problem of estimating model (1) has been to look for a fixed T consistent estimator because most of the micropanels have much larger N than T, and the finite sample bias found when using some of the estimators that are consistent only when $T \to \infty$ is not negligible. Nevertheless, our goal is not necessarily to find a consistent estimator for fixed T, but an estimator with a good finite sample performance and a reasonable asymptotic approximation for the samples used in empirical studies. Moreover, as previously said, only partial solutions with restrictive assumptions have been found for fixed T, and, as suggested by the results in Honoré and Tamer (2004), identification problems arise in a fixed-T context when those assumptions are relaxed (see also Chamberlain, 1992 and Arellano, 2003). Also, as shown by Alvarez and Arellano (2003) for linear autoregressive models, the properties of some common estimators that are optimal when T is fixed, may be quite different when both T and N tend to infinity. In contrast to time series or single cross sections, panel data can exploit both dimensions for identification and inference. Besides, panels with $T = 2$ are not so common in practice and for values of T like 8 or 9 the finite sample bias of estimators that are only consistent when $T \to \infty$ might not be important. Given all this, we should not be restricted to fixed T consistent estimators and fixed T asymptotics.

Cox and Reid (1987) considered the general problem of doing inference for a parameter of interest in the absence of knowledge about nuisance parameters. Their formulation requires information orthogonality between the two types of parameters. That is, the expected information matrix to be block diagonal between the parameters of interest and the nuisance parameters. Therefore, they transform the nuisance parameters by reparametrization in order to get information orthogonality, and then modify the likelihood. Their general framework has been employed for static binary choice panel data models with fixed effects in Arellano (2003). The idea is to modify the concentrated log-likelihood to correct the first term on the asymptotic bias that comes from the estimation of the fixed effects. The first order condition or estimating equation of the modified likelihood is more nearly unbiased than the one using the concentrated likelihood. Arellano (2003) notes that the modified concentrated log-likelihood can be written in terms of the original parameters.[6]

For the model considered in this paper, the modified first order condition expressed in terms of the original parameters of the model is[7]

$$\mathbf{d}_{\gamma Mi}(\boldsymbol{\gamma}) = \mathbf{d}_{\gamma i}(\boldsymbol{\gamma}, \widehat{\eta}_i(\boldsymbol{\gamma})) - \frac{1}{2}\frac{1}{d_{\eta\eta i}(\boldsymbol{\gamma}, \widehat{\eta}_i(\boldsymbol{\gamma}))}\left(\mathbf{d}_{\gamma\eta\eta i}(\boldsymbol{\gamma}, \widehat{\eta}_i(\boldsymbol{\gamma})) + d_{\eta\eta\eta i}(\boldsymbol{\gamma}, \widehat{\eta}_i(\boldsymbol{\gamma}))\frac{\partial \widehat{\eta}_i(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right) + \frac{\partial}{\partial \eta_i}\left(\frac{1}{\mathrm{E}[d_{\eta\eta i}(\boldsymbol{\gamma}, \eta_i)]}\mathrm{E}[\mathbf{d}_{\gamma\eta i}(\boldsymbol{\gamma}, \eta_i)]\right)\Bigg|_{\eta_i = \widehat{\eta}_i(\boldsymbol{\gamma})} = 0, \tag{4}$$

where $\mathbf{d}_{\gamma i}(\boldsymbol{\gamma}, \widehat{\eta}_i(\boldsymbol{\gamma})) = \partial l_i(\boldsymbol{\gamma}, \eta_i)/\partial \gamma|_{\eta_i = \widehat{\eta}_i(\gamma)}$ is the standard first order condition from the concentrated likelihood, $\mathbf{d}_{\gamma\eta i}(\boldsymbol{\gamma}, \eta_i) = \partial^2 l_i/\partial\gamma\partial\eta_i$, $d_{\eta\eta i}(\boldsymbol{\gamma}, \eta_i) = \partial^2 l_i/\partial\eta_i^2$, $\mathbf{d}_{\gamma\eta\eta i}(\boldsymbol{\gamma}, \widehat{\eta}_i(\boldsymbol{\gamma})) =$

[6] It should be noted that the modified score expressed in terms of the original parameters used in Arellano (2003) and here, does not correspond exactly with the score from the Cox–Reid modified likelihood. The suppressed term is not invariant to reparametrizations and it is irrelevant for the purpose of bias reduction. See Arellano (2005).

[7] Here and all throughout the paper, even though it is not explicitly indicated, expectations are conditional on the same set of information as the likelihood.

$\partial^3 l_i/\partial\gamma\partial\eta_i^2|_{\eta_i=\widehat{\eta}_i(\gamma)}$, $d_{\eta\eta\eta i}(\boldsymbol{\gamma},\widehat{\eta}_i(\boldsymbol{\gamma})) = \partial^3 l_i/\partial\eta_i^3|_{\eta_i=\widehat{\eta}_i(\gamma)}$, and $\widehat{\eta}_i(\gamma)$ is obtained from the first order condition of η_i, as it is done in order to concentrate the log-likelihood. Therefore, the MMLE of γ, $\widehat{\gamma}_{\mathrm{MMLE}}$, is the value that makes $\mathbf{d}_{\gamma Mi}(\widehat{\gamma}_{\mathrm{MMLE}}) = 0$, i.e. the value of γ that solves the score equation (4).

2.3. Reduction of the order of the bias

Arellano (2003) shows that the modification of the standard ML score equation reduces the order of the bias under the assumption that an information orthogonal reparametrization exits. The problem is that such reparametrization may not exist. Moreover, it is know that in general an information orthogonal reparametrization is not feasible when γ is multidimensional, as in the kind of models considered here. In this paper all calculations are made using the original parametrization of the model, showing that the reduction of the order of bias made by the modification in (4) does not depend on an information orthogonal parametrization. The order of the bias is reduced in models like (1) where the parameters are not information orthogonal and it does not rely on the existence of an orthogonal reparametrization. The modification on the score of the concentrated log-likelihood (4) is a first order adjustment on the asymptotic bias, so the first order condition is more nearly unbiased.

Denote $\mathbf{d}_{\gamma i} = \partial l_i/\partial\boldsymbol{\gamma}$, $\mathbf{d}_{\gamma\eta i} = \partial l_i/\partial\boldsymbol{\gamma}\partial\eta_i$, and so on. Bold letters are used to denote vectors, e.g. $\mathbf{d}_{\gamma i} = (\partial l_i/\partial\alpha,\ \partial l_i/\partial\beta)'$, as oppose to scalars like $d_{\eta i} = \partial l_i/\partial\eta_i$. $d_{\eta i0}$, $\mathbf{d}_{\gamma i0}$, $\mathbf{d}_{\gamma\eta i0}$, etc., denote $d_{\eta i}(\boldsymbol{\gamma}_0,\eta_{i0})$, $\mathbf{d}_{\gamma i}(\boldsymbol{\gamma}_0,\eta_{i0})$, $\mathbf{d}_{\gamma\eta i}(\boldsymbol{\gamma}_0,\eta_{i0})$, etc., in other words the derivatives are evaluated at the true value of the parameters.

We are going to take the log-likelihood of the T observations of an individual i. That likelihood depends on $\boldsymbol{\gamma}$ and on η_i, but it does not depend on any other η_j where $j\neq i$. However, the result extends to the likelihood for all N individuals by simple additivity that exploits the independence of observations between individuals.

Expanding the score of the concentrated log-likelihood around η_{i0}, and evaluating it at $\boldsymbol{\gamma}_0$:

$$\mathbf{d}_{\gamma i}(\boldsymbol{\gamma}_0,\widehat{\eta}_i(\boldsymbol{\gamma}_0)) = \mathbf{d}_{\gamma i0} + \mathbf{d}_{\gamma\eta_i 0}(\widehat{\eta}_i(\boldsymbol{\gamma}_0)-\eta_{i0}) + \tfrac{1}{2}\mathbf{d}_{\gamma\eta\eta i0}(\widehat{\eta}_i(\boldsymbol{\gamma}_0)-\eta_{i0})^2 + \mathrm{O_p}(T^{-1/2}). \tag{5}$$

Under the usual regularity conditions the first three terms are $\mathrm{O_p}(T^{1/2})$, $\mathrm{O_p}(T^{1/2})$ and $\mathrm{O_p}(1)$, respectively, since, as usual, $\mathbf{d}_{\gamma\eta i0}$ is $\mathrm{O_p}(T)$, $(\widehat{\eta}_i(\boldsymbol{\gamma}_0)-\eta_{i0})$ is $\mathrm{O_p}(T^{-1/2})$, and $\mathbf{d}_{\gamma\eta\eta i0}$ is $\mathrm{O_p}(T)$. In the case of information orthogonal parameters made by Arellano (2003), $\mathbf{d}_{\gamma\eta i0}$ is $\mathrm{O_p}(\sqrt{T})$. That difference implies that we are going to need a higher expansion for $(\widehat{\eta}_i(\boldsymbol{\gamma}_0)-\eta_{i0})$ to get a reminder of the same order once it is multiplied by $\mathbf{d}_{\gamma\eta i0}$. Eq. (5) clearly shows that the score evaluated at $\boldsymbol{\gamma}_0$ differs from the value of the score that we want to get, $\mathbf{d}_{\gamma i0}$—i.e. the score evaluated at both γ_0 and η_{i0}—as much as $\widehat{\eta}_i(\boldsymbol{\gamma}_0)$ differs from η_{i0}.

From Chapter 7 of McCullagh (1987), we have the following asymptotic expansion for $(\widehat{\eta}_i(\boldsymbol{\gamma}_0)-\eta_{i0})$ (compare also with Akahira and Takeuchi, 1982 and Ferguson, 1992):

$$\begin{aligned}(\widehat{\eta}_i(\boldsymbol{\gamma}_0)-\eta_{i0}) = & -\frac{(1/T)d_{\eta i0}}{\mathrm{E}[(1/T)d_{\eta\eta i0}]} + \frac{1}{T}\frac{(1/\sqrt{T})(d_{\eta\eta i0}-\mathrm{E}[(1/T)d_{\eta\eta i0}])}{\mathrm{E}[(1/T)\,d_{\eta\eta i0}]^2}\frac{1}{\sqrt{T}}\,d_{\eta i0} \\ & + \frac{1}{2}\frac{1}{T}\frac{\mathrm{E}[(1/T)\,d_{\eta\eta\eta i0}](1/T)\,d_{\eta i0}d_{\eta i0}}{(-\mathrm{E}[(1/T)\,d_{\eta\eta i0}])^3} + \mathrm{O_p}(T^{-3/2}).\end{aligned} \tag{6}$$

We also need the expression for $(\widehat{\eta}_i(\gamma_0) - \eta_{i0})^2$:

$$(\widehat{\eta}_i(\gamma_0) - \eta_{i0})^2 = \frac{1}{T} \frac{(1/T)d_{\eta i0} d_{\eta i0}}{(\mathrm{E}[(1/T)d_{\eta\eta i0}])^2} + \mathrm{O_p}(T^{-3/2}). \tag{7}$$

Substituting some numerators in (6) and (7) by their expectations,[8] and making some simplifications

$$\begin{aligned}(\widehat{\eta}_i(\gamma_0) - \eta_{i0}) = & -\frac{(1/T)d_{\eta i0}}{\mathrm{E}[(1/T)d_{\eta\eta i0}]} + \frac{1}{T}\frac{\mathrm{E}[(1/T)d_{\eta\eta i0}d_{\eta i0}]}{\mathrm{E}[(1/T)d_{\eta\eta i0}]^2} \\ & + \frac{1}{2}\frac{1}{T}\frac{\mathrm{E}[(1/T)d_{\eta\eta\eta i0}]}{(\mathrm{E}[(1/T)d_{\eta\eta i0}])^2} + \mathrm{O_p}(T^{-3/2}),\end{aligned} \tag{8}$$

$$(\widehat{\eta}_i(\gamma_0) - \eta_{i0})^2 = \frac{1}{T}\frac{-1}{\mathrm{E}[(1/T)d_{\eta\eta i0}]} + \mathrm{O_p}(T^{-3/2}), \tag{9}$$

where we have made use of the information matrix identity,[9] and of the zero mean property of the score: $\mathrm{E}[(1/T)d_{\eta i0}] = 0$.

Replacing (8) and (9) in (5)

$$\begin{aligned}\mathbf{d}_{\gamma i}(\gamma_0, \widehat{\eta}_i(\gamma_0)) = & \mathbf{d}_{\gamma i0} - \mathbf{d}_{\gamma\eta i0}\frac{(1/T)d_{\eta i0}}{\mathrm{E}[(1/T)d_{\eta\eta i0}]} + \frac{1}{T}\mathbf{d}_{\gamma\eta i0}\frac{\mathrm{E}[(1/T)d_{\eta\eta i0}d_{\eta i0}]}{\mathrm{E}[(1/T)d_{\eta\eta i0}]^2} \\ & + \frac{1}{2}\frac{1}{T}\mathbf{d}_{\gamma\eta i0}\frac{\mathrm{E}[(1/T)d_{\eta\eta\eta i0}]}{(\mathrm{E}[(1/T)d_{\eta\eta i0}])^2} + \mathbf{d}_{\gamma\eta i0}\mathrm{O_p}(T^{-3/2}) \\ & - \frac{1}{2}\frac{1}{T}\mathbf{d}_{\gamma\eta\eta i0}\frac{1}{\mathrm{E}[(1/T)d_{\eta\eta i0}]} \\ & + \mathbf{d}_{\gamma\eta\eta i0}\mathrm{O_p}(T^{-3/2}) + \mathrm{O_p}(T^{-1/2}).\end{aligned} \tag{10}$$

Since both $\mathbf{d}_{\gamma\eta i0}$ and $\mathbf{d}_{\gamma\eta\eta i0}$ are $\mathrm{O_p}(T)$,

$$\begin{aligned}\mathbf{d}_{\gamma i}(\gamma_0, \widehat{\eta}_i(\gamma_0)) = & \mathbf{d}_{\gamma i0} - \frac{1}{\mathrm{E}[(1/T)d_{\eta\eta i0}]}\frac{1}{T}\mathbf{d}_{\gamma\eta i0}d_{\eta i0} + \frac{1}{T}\mathbf{d}_{\gamma\eta i0}\frac{\mathrm{E}[(1/T)d_{\eta\eta i0}d_{\eta i0}]}{\mathrm{E}[(1/T)d_{\eta\eta i0}]^2} \\ & + \frac{1}{2}\frac{1}{T}\mathbf{d}_{\gamma\eta i0}\frac{\mathrm{E}[(1/T)d_{\eta\eta\eta i0}]}{(\mathrm{E}[(1/T)d_{\eta\eta i0}])^2} - \frac{1}{2}\frac{1}{T}\mathbf{d}_{\gamma\eta\eta i0}\frac{1}{\mathrm{E}[(1/T)\,d_{\eta\eta i0}]} \\ & + \mathrm{O_p}(T^{-1/2}).\end{aligned} \tag{11}$$

Taking expectations,

$$\begin{aligned}\mathrm{E}[\mathbf{d}_{\gamma i}(\gamma_0, \widehat{\eta}_i(\gamma_0))] = & \frac{-1}{\mathrm{E}[(1/T)d_{\eta\eta i0}]}\left(\mathrm{E}\left[\frac{1}{T}\mathbf{d}_{\gamma\eta i0}d_{\eta i0}\right] + \frac{1}{2}\mathrm{E}\left[\frac{1}{T}\mathbf{d}_{\gamma\eta\eta i0}\right]\right) \\ & + \mathrm{E}\left[\frac{1}{T}\mathbf{d}_{\gamma\eta i0}\right]\frac{(\mathrm{E}[(1/T)d_{\eta\eta i0}d_{\eta i0}] + 1/2\,\mathrm{E}[(1/T)d_{\eta\eta\eta i0}])}{\mathrm{E}[(1/T)d_{\eta\eta i0}]^2} \\ & + \mathrm{O}(T^{-1}),\end{aligned} \tag{12}$$

[8] The reminders from this substitution ($\bar{x} = (1/T)\sum_t x_t = \mathrm{E}(x) + \mathrm{O_p}(T^{-1/2})$), once divided by the $1/T$ that multiplies each term, is of order $\mathrm{O_p}(T^{-3/2})$. This is why the first numerator in (6) cannot be substituted by its expectation; because that term is not multiplied by $1/T$.

[9] Information matrix identity: $-\mathrm{E}[(1/T)d_{\eta\eta i0}] = \mathrm{E}[(1/T)d_{\eta i0}d_{\eta i0}]$.

where the reminder is of order $O(T^{-1})$ because the $O_p(T^{-1/2})$ terms have zero mean (cf. Ferguson et al., 1991, p. 288 and appendix). Therefore, the score of the concentrated MLE has a bias of order O(1).

To make the expansion of the score of the modified log-likelihood we need the following results:

$$\frac{\partial}{\partial \eta_i}(\mathrm{E}[\mathbf{d}_{\gamma\eta i}(\gamma,\eta_i)]) = \mathrm{E}[\mathbf{d}_{\gamma\eta\eta i}(\gamma,\eta_i)] + \mathrm{E}[\mathbf{d}_{\gamma\eta i}(\gamma,\eta_i)d_{\eta i}(\gamma,\eta_i)], \tag{13}$$

$$\frac{\partial}{\partial \eta_i}(\mathrm{E}[d_{\eta\eta i}(\gamma,\eta_i)]) = \mathrm{E}[d_{\eta\eta\eta i}(\gamma,\eta_i)] + \mathrm{E}[d_{\eta\eta i}(\gamma,\eta_i)d_{\eta i}(\gamma,\eta_i)]. \tag{14}$$

Differencing $d_{\boldsymbol{\eta} i}(\gamma,\widehat{\eta}_i(\gamma)) = 0$ with respect to γ

$$\mathbf{d}_{\eta\gamma i}(\gamma,\widehat{\eta}_i(\gamma)) + d_{\eta\boldsymbol{\eta} i}(\gamma,\widehat{\eta}_i(\gamma))\frac{\partial \eta_i(\gamma)}{\partial \gamma} = 0, \tag{15}$$

$$\frac{\partial \eta_i(\gamma)}{\partial \gamma} = \frac{-1}{d_{\eta\boldsymbol{\eta} i}(\gamma,\widehat{\eta}_i(\gamma))}\,\mathbf{d}_{\eta\gamma i}(\gamma,\widehat{\eta}_i(\gamma)). \tag{16}$$

Therefore,

$$\frac{\partial \eta_i(\gamma_0)}{\partial \gamma} = \frac{-1}{\mathrm{E}[d_{\eta\boldsymbol{\eta} i0}]}\,\mathrm{E}[\mathbf{d}_{\eta\gamma i0}] + O_p(T^{-1/2}). \tag{17}$$

In Eq. (4), the modified score for i is

$$\begin{aligned}\mathbf{d}_{\gamma Mi}(\gamma) = \mathbf{d}_{\gamma i}(\gamma,\widehat{\eta}_i(\gamma)) &- \frac{1}{2}\frac{1}{d_{\eta\eta i}(\gamma,\widehat{\eta}_i(\gamma))}\left(\mathbf{d}_{\gamma\eta\eta i}(\gamma,\widehat{\eta}_i(\gamma)) + d_{\eta\eta\eta i}(\gamma,\widehat{\eta}_i(\gamma))\frac{\partial \widehat{\eta}_i(\gamma)}{\partial \gamma}\right)\\ &+ \frac{\partial}{\partial \eta_i}\left(\frac{1}{\mathrm{E}[d_{\eta\eta i}(\gamma,\eta_i)]}\,\mathrm{E}[\mathbf{d}_{\gamma\eta i}(\gamma,\eta_i)]\right)\Bigg|_{\eta_i=\widehat{\eta}_i(\gamma)}.\end{aligned} \tag{18}$$

Using the results in (13) and (14), and focusing on the difference $\mathbf{d}_{\gamma Mi}(\gamma) - \mathbf{d}_{\gamma i}(\gamma,\eta_i(\gamma))$, since we already have the bias of $\mathbf{d}_{\gamma i}(\gamma,\eta_i(\gamma))$ in Eq. (12):

$$\begin{aligned}\mathbf{d}_{\gamma Mi}(\gamma) - \mathbf{d}_{\gamma i}(\gamma,\eta_i(\gamma)) = &-\frac{1}{2}\frac{1}{d_{\eta\eta i}(\gamma,\widehat{\eta}_i(\gamma))}\left(\mathbf{d}_{\gamma\eta\eta i}(\gamma,\widehat{\eta}_i(\gamma)) + d_{\eta\eta\eta i}(\gamma,\widehat{\eta}_i(\gamma))\frac{\partial \widehat{\eta}_i(\gamma)}{\partial \gamma}\right)\\ &+ \frac{1}{\mathrm{E}[d_{\eta\eta i}(\gamma,\eta_i(\gamma))]}(\mathrm{E}[\mathbf{d}_{\gamma\eta\eta i}(\gamma,\eta_i(\gamma))] + \mathrm{E}[\mathbf{d}_{\gamma\eta i}(\gamma,\eta_i(\gamma))d_{\eta i}(\gamma,\eta_i(\gamma))])\\ &- \frac{1}{(\mathrm{E}[d_{\eta\eta i}(\gamma,\eta_i(\gamma))])^2}\,\mathrm{E}[\mathbf{d}_{\gamma\eta i}(\gamma,\eta_i(\gamma))](\mathrm{E}[d_{\eta\eta\eta i}(\gamma,\eta_i(\gamma))]\\ &+ \mathrm{E}[d_{\eta\eta i}(\gamma,\eta_i(\gamma))d_{\eta i}(\gamma,\eta_i(\gamma))]).\end{aligned} \tag{19}$$

Evaluating (19) at γ_0, substituting (17) on it, using the fact that $\eta_i(\gamma) = \eta_{i0} + O_p(T^{-1/2})$ and adding $1/T$ accordingly on the numerators and denominators while substituting the

different terms by their expectations,[10]

$$\begin{aligned}
&\mathbf{d}_{\gamma Mi}(\gamma_0) - \mathbf{d}_{\gamma i}(\boldsymbol{\gamma}_0, \eta_i(\boldsymbol{\gamma}_0)) \\
&\quad = -\frac{1}{2}\frac{1}{\mathrm{E}[(1/T)d_{\eta\eta i0}]}\left(\mathrm{E}\left[\frac{1}{T}\mathbf{d}_{\gamma\eta\eta i0}\right] - \frac{\mathrm{E}[(1/T)d_{\eta\eta\eta i0}]}{\mathrm{E}[(1/T)\,d_{\eta\boldsymbol{\eta} i0}]}\,\mathrm{E}\left[\frac{1}{T}\,\mathbf{d}_{\eta\gamma i0}\right]\right) \\
&\qquad + \frac{1}{\mathrm{E}[(1/T)d_{\eta\eta i0}]}\left(\mathrm{E}\left[\frac{1}{T}\,\mathbf{d}_{\gamma\eta\eta i0}\right] + \mathrm{E}\left[\frac{1}{T}\,\mathbf{d}_{\gamma\eta i0}d_{\eta i0}\right]\right) \\
&\qquad - \frac{(\mathrm{E}[(1/T)d_{\eta\eta\eta i0}] + \mathrm{E}[(1/T)d_{\eta\eta i0}d_{\eta i0}])}{(\mathrm{E}[(1/T)d_{\eta\eta i0}])^2}\,\mathrm{E}\left[\frac{1}{T}\,\mathbf{d}_{\gamma\eta i0}\right] + \mathrm{O_p}(T^{-1/2}).
\end{aligned} \tag{20}$$

Then,

$$\begin{aligned}
\mathrm{E}[\mathbf{d}_{\gamma Mi}(\gamma_0)] = \mathrm{E}[\mathbf{d}_{\gamma i}(\boldsymbol{\gamma}_0, \eta_i(\boldsymbol{\gamma}_0))] &+ \frac{1}{\mathrm{E}[(1/T)d_{\eta\eta i0}]}\left(\mathrm{E}\left[\frac{1}{T}\,\mathbf{d}_{\gamma\eta i0}d_{\eta i0}\right] + \frac{1}{2}\,\mathrm{E}\left[\frac{1}{T}\,\mathbf{d}_{\gamma\eta\eta i0}\right]\right) \\
&- \mathrm{E}\left[\frac{1}{T}\,\mathbf{d}_{\gamma\eta i0}\right]\frac{(\mathrm{E}[(1/T)d_{\eta\eta i0}d_{\eta i0}] + 1/2\mathrm{E}[(1/T)d_{\eta\eta\eta i0}])}{(\mathrm{E}[(1/T)d_{\eta\eta i0}])^2}.
\end{aligned} \tag{21}$$

Combining (21) with (12) shows that the modification of $\mathbf{d}_{\gamma i}(\boldsymbol{\gamma}, \widehat{\eta}_i(\boldsymbol{\gamma}))$ in (4) reduces the order of the bias of the score of the MLE from O(1) to $\mathrm{O}(T^{-1})$. For the estimator this implies that $\widehat{\gamma}_{\mathrm{MMLE}}$ is unbiased to order $\mathrm{O}(T^{-2})$, whereas $\widehat{\gamma}_{\mathrm{MLE}}$ has a bias of order $\mathrm{O}(T^{-1})$.

Therefore, under standard regularity conditions the MMLE of γ, $\widehat{\gamma}_{\mathrm{MMLE}}$, defined as the value that solves the score equation (4), has a bias of order $\mathrm{O}(T^{-2})$, as opposed to $\mathrm{O}(T^{-1})$ of the MLE.

A much simpler setting is considered to show some intuition behind the proposed modification and the presented result. Consider the model where $y_{it} \underset{iid}{\sim} \mathrm{N}(\eta_i, \sigma_0^2)$. The ML estimator of σ_0^2 is $\widehat{\gamma}_{\mathrm{MLE}} = (1/NT)\sum_i \sum_t (y_{it} - \widehat{\eta}_i)^2$. It is well-known that $\widehat{\gamma}_{\mathrm{MLE}}$ is not a consistent estimator of σ_0^2 when $N \to \infty$ with fixed T, since it converges to $((T-1)/T)\sigma_0^2$. In terms of the first order condition of the concentrated log-likelihood, the problem is that the expectation of the score $d_\gamma(\gamma)$, evaluated at the true value σ_0^2 is not equal to zero. $d_\gamma(\boldsymbol{\gamma})$ equals $-(NT/2)(1/\gamma) + (1/2\gamma^2)\sum_i \sum_t (y_{it} - \widehat{\eta}_i)^2$, and $\mathrm{E}[d_\gamma(\gamma)]$ evaluated at $\gamma = \sigma_0^2$ is equal to $-N/2\sigma_0^2$. So, $-N/2\sigma_0^2$ is the bias of order O(1) in the score equation. It can be easily shown that the modified score equation (4) in this example is $d_{\gamma M}(\gamma) = d_\gamma(\gamma) + N/2\gamma$. In this case the modification is only $N/2\gamma$ because σ_0^2 and η_i are information orthogonal and, therefore, many terms in (4) cancel out. From the modified score equation, $d_{\gamma M}(\gamma) = 0$, we obtain that $\widehat{\gamma}_{\mathrm{MMLE}} = (1/N(T-1))\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \widehat{\eta}_i)^2$, which is the fixed T consistent estimator of σ_0^2. The modification is removing the bias of the score equation. The expectation of the modified score is correctly centered: $\mathrm{E}[d_{\gamma M}(\gamma = \sigma_0^2)] = \mathrm{E}[d_\gamma(\gamma = \sigma_0^2)] + (N/2\sigma_0^2) = 0$. In general, as in the dynamic discrete choice models considered in this paper, the modified score will not be exactly centered at zero, but it will have a bias of smaller order of magnitude, as shown in this section.

2.4. Consistency and asymptotic normality

This subsection shows that the MMLE, as the MLE, follows a normal distribution asymptotically, under the classical asymptotic setting. Also, that the MMLE has no bias in

[10]Substituting each term by its expectation rises a remainder term of order $\mathrm{O_p}(T^{-1/2})$ with zero mean.

its asymptotic distribution even when N grows faster than T, provided T grows faster than $\sqrt[3]{N}$, in contrast to the MLE that has a bias in its asymptotic distribution unless T grows faster than N. All this is just an implication of the order of the bias of both estimators. It is important to note that, given the class of models considered here, for which the usual regularity conditions apply, we are just using the standard asymptotic results of the MLE. For the MMLE that standard asymptotic theory also applies, with the only differences coming from the reduction on the order of the bias.

Consistency: In the classical asymptotic setting the basic result for maximum likelihood estimation states that $\widehat{\gamma}_{\mathrm{MLE}}$ is consistent as $T \to \infty$. In the same setting the MMLE of γ is also consistent as $T \to \infty$, since the average score of both converge to the same object as T increases.[11]

From (11) and (12), by simply summing over independent i-observations, we can write

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \mathbf{d}_{\gamma i}(\gamma_0, \widehat{\eta}_i(\gamma_0)) = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \mathbf{d}_{\gamma i0} + \sqrt{\frac{N}{T}} b_N + \sqrt{\frac{N}{T^2}} a_N + \sqrt{\frac{N}{T^3}} c_N, \tag{22}$$

where $b_N = (1/N)\sum_{i=1}^{N} b_i$,

$$b_i = \frac{-1}{\mathrm{E}[(1/T) d_{\eta\eta i0}]} \left(\mathrm{E}\left[\frac{1}{T} \mathbf{d}_{\gamma\eta i0} d_{\eta i0} \right] + \frac{1}{2} \mathrm{E}\left[\frac{1}{T} \mathbf{d}_{\gamma\eta\eta i0} \right] \right) + \mathrm{E}\left[\frac{1}{T} \mathbf{d}_{\gamma\eta i0} \right] \frac{(\mathrm{E}[(1/T) d_{\eta\eta i0} d_{\eta i0}] + 1/2 \mathrm{E}[(1/T) d_{\eta\eta\eta i0})]}{\mathrm{E}[(1/T) d_{\eta\eta i0}]^2},$$

$a_N = (1/N)\sum_{i=1}^{N} a_i$, a_i is $\sqrt{T}$ times the $\mathrm{O_p}(T^{-1/2})$ terms of the remainder, $c_N = (1/N)\sum_{i=1}^{N} c_i$, and c_i is T times the $\mathrm{O_p}(T^{-1})$ terms on the remainder of (11). Therefore, b_i, a_i and c_i are $\mathrm{O_p}(1)$.

For the MMLE (22) is

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \mathbf{d}_{\gamma \mathbf{M} i}(\gamma_0) = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \mathbf{d}_{\gamma i0} + \sqrt{\frac{N}{T^2}} a_N + \sqrt{\frac{N}{T^3}} c_N. \tag{23}$$

In the context of the models considered in this paper, a standard central limit theorem applies to the true score $\mathbf{d}_{\gamma i}(\boldsymbol{\gamma}, \eta_i) = \partial l_i(\boldsymbol{\gamma}, \eta_i)/\partial \boldsymbol{\gamma}$ so that we have

$$V_{NT}^{-1/2} \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \mathbf{d}_{\gamma i}(\gamma_0, \widehat{\eta}_{i0}) \xrightarrow{d} \mathrm{N}(0, I), \tag{24}$$

where $V_{NT}^{-1/2}$ comes from the outer product of the score.

If $N/T \to c$, $0 < c < \infty$, from (22) and (24)

$$V_{NT}^{-1/2} \left\{ \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \mathbf{d}_{\gamma i}(\gamma_0, \widehat{\eta}_i(\gamma_0)) - \sqrt{\frac{N}{T}} b_N \right\} \xrightarrow{d} \mathrm{N}(0, I). \tag{25}$$

From a first order expansion of the concentrated score around the true value we obtain the usual expression for the estimator:

$$H_{NT} \sqrt{NT} (\widehat{\gamma}_{\mathrm{MLE}} - \gamma_0) = -\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \mathbf{d}_{\gamma i}(\gamma_0, \widehat{\eta}_i(\gamma_0)) + \mathrm{O_p}\left(\frac{1}{\sqrt{NT}} \right), \tag{26}$$

[11] It is important to make clear that for models like (1) not only the MLE but also MMLE are inconsistent for fixed T, since the modification corrects first order bias but not biases of smaller order.

where

$$H_{NT} = \frac{1}{NT}\sum_{i=1}^{N} \left.\frac{\partial \mathbf{d}_{\gamma i}(\gamma, \widehat{\eta}_i(\gamma))}{\partial \gamma}\right|_{\gamma=\gamma_0}.$$

Combining (25) and (26),

$$(H'_{NT} V_{NT}^{-1} H_{NT})^{1/2}\sqrt{NT}\left(\widehat{\gamma}_{\mathrm{MLE}} - \gamma_0 + \frac{1}{T} H_{NT}^{-1} b_N\right) \xrightarrow{d} \mathrm{N}(0, I). \tag{27}$$

Therefore, $\widehat{\gamma}_{\mathrm{MLE}}$ has a bias of order $\mathrm{O}(T^{-1})$ in its asymptotic distribution, whenever T does not grow faster than N. However, for the MMLE, if N and T grows at the same rate, from (23) and (24) we have

$$V_{NT}^{-1/2}\left\{\frac{1}{\sqrt{NT}}\sum_{i=1}^{N} \mathbf{d}_{\gamma \mathbf{M} i}(\gamma_0)\right\} \xrightarrow{d} \mathrm{N}(0, I). \tag{28}$$

Again, using the first order expansion of the modified score

$$H_{NT}^{\dagger}\sqrt{NT}(\widehat{\gamma}_{\mathrm{MMLE}} - \gamma_0) = -\frac{1}{\sqrt{NT}}\sum_{i=1}^{N} \mathbf{d}_{\gamma \mathbf{M} i}(\gamma_0) + \mathrm{O_p}\left(\frac{1}{\sqrt{NT}}\right), \tag{29}$$

where

$$H_{NT}^{\dagger} = \frac{1}{NT}\sum_{i=1}^{N} \left.\frac{\partial \mathbf{d}_{\gamma \mathbf{M} i}(\gamma)}{\partial \gamma}\right|_{\gamma=\gamma_0}.$$

Finally, combining it with (28) we have

$$(H_{NT}^{\dagger\prime} V_{NT}^{-1} H_{NT}^{\dagger})^{1/2}\sqrt{NT}(\widehat{\gamma}_{\mathrm{MMLE}} - \gamma_0) \xrightarrow{d} \mathrm{N}(0, I). \tag{30}$$

This result in (30) is not only when T grows at the same rate as N, but also even if $N/T \to \infty$, provided $\sqrt[3]{N}/T \to 0$. This is because the $\mathrm{O_p}(T^{-1/2})$ terms included in a_i in Eqs. (22) and (23) have zero mean.[12] As it has been already noted, this is an straightforward result from the order of the bias of $\widehat{\gamma}_{\mathrm{MMLE}}$.

Appendix A has the calculations needed in order to compute $d_{\alpha Mi}(\alpha, \beta)$ for a particular model of the type considered in this paper. In Appendix B, I address the problem of how to optimize a concentrated likelihood when $\widehat{\eta}_i(\alpha, \beta)$ cannot be analytically calculated.

Given the results presented in this section, the finite sample bias of the MMLE may be negligible for moderate T even though, in general, it is only consistent when $T \to \infty$. A main advantage of this way of estimating over other methods is its generality. Estimators like the ones mentioned in Section 2.1 are too specific and require very restrictive assumptions. However, MMLE can be applied to different models with different assumptions. For example, this method does not depend on the logistic assumption and it allows for time dummy variables. It could be used with more lags of the endogenous variables included on the set of explanatory variables. Also, MMLE could be generally applied to multinomial choice models and other non-linear model, not only to binary choice.

In addition to those properties, MMLE is a convenient estimator to compute the policy parameters of interest because the fixed effects are estimated as part of the estimation

[12] Woutersen (2001) proves that result for the integrated likelihood estimator. Li et al. (2003), consider a different adjustment to the MLE and derive its double asymptotic properties, too.

process whereas in the fixed T consistent estimation you get rid of them, and the marginal effects depend on the fixed effects. Furthermore, asymptotic properties in both N and T, have to be considered since the estimates of the marginal effects are only consistent when $T \to \infty$.

3. Monte Carlo evidence

In this section Monte Carlo simulations are used to evaluate the performance of the MMLE in different sample sizes to see if this new estimator has good properties in finite samples, and its asymptotic distribution is a good approximation.

3.1. Parameters of the model

The first model considered is a dynamic logit:

$$y_{it} = 1\{\alpha y_{it-1} + \beta x_{it} + \eta_i + v_{it} \geqslant 0\} \quad (t = 0, \ldots, T-1; i = 1, \ldots, N), \tag{31}$$

where x_{it} is an exogenous variable, η_i is an unobservable individual effect and $-v_{it}$ is independently distributed with cdf F conditional on η_i, so that

$$\Pr(y_{it} = 1 | \eta_i, y_i^{t-1}, x_i) = F(\alpha y_{it-1} + \beta x_{it} + \eta_i) = F_{it} \tag{32}$$

and F is the logistic cdf. I have considered this dynamic logit because under additional conditions Honoré and Kyriazidou (2000) have a consistent estimator for fixed T, so I have an estimator to compare it with. I will refer to the estimator proposed by them as HK. I design the experiment as they did, so that my results could be compared to the ones they report. One practical disadvantage of their estimator is that it requires to choose a bandwidth and the results may be negatively affected by that election. Another difference is that their estimator excludes observations for which $y_{i1} = y_{i2}$ and MMLE excludes observations for which $\sum_{t=1}^{T-1} y_{it} = 0$ or $\sum_{t=1}^{T-1} y_{it} = T-1$, like the MLE.[13] The proportion of observations used in the latter estimator is increasing with T whereas, in the former case, it remains constant.

As in Honoré and Kyriazidou (2000) I have made a 1000 replications, $\beta_0 = 1$, x_{it} is i.i.d. $N(0, \pi^2/3)$, v_{it} is *i.i.d.* logistically distributed, and $\eta_i = (x_{i0} + x_{i1} + x_{i2} + x_{i3})/4$, so that the fixed effects are correlated with x. For each simulated sample I have estimated by maximum likelihood and by modified maximum likelihood. HK estimations are taken from the tables reported in their paper. I report results with samples of different N and T size. I expect the MMLE to improve much more with T than with N, whereas HK estimator has a significant improvement with N since it is fixed T consistent. I show the median bias and the median absolute error (MAE) because they are robust to outliers, and to be able to compare with results presented in Honoré and Kyriazidou (2000).

Table 1 reports the estimates of the parameters for a value of α_0 equal to 0.5. For $T = 4$, though the bias is greatly reduced compared with MLE, the median bias and MAE of $\widehat{\alpha}_{\text{MMLE}}$ is far from the results obtained by Honoré and Kyriazidou. This is not surprising because, as I said, MMLE is not a fixed T consistent estimator whereas HK is, and I am

[13] Note that T is the total number of periods and $t = T-1$ is the last period we observe since the first one is $t = 0$.

Table 1
Logit design with different N and T values

		N	$T=4$			$T=8$		
			250	500	1000	250	500	1000
MLE	$\widehat{\beta}$	Bias	0.759	0.768	0.759	0.248	0.253	0.254
	$\widehat{\beta}$	MAE	0.759	0.768	0.759	0.248	0.253	0.254
	$\widehat{\alpha}$	Bias	−2.548	−2.513	−2.55	−0.757	−0.746	−0.741
	$\widehat{\alpha}$	MAE	2.548	2.513	2.55	0.757	0.746	0.741
MMLE	$\widehat{\beta}$	Bias	−0.054	−0.053	−0.057	0.012	0.015	0.015
	$\widehat{\beta}$	MAE	0.068	0.055	0.057	0.039	0.031	0.022
	$\widehat{\alpha}$	Bias	−0.554	−0.543	−0.563	−0.106	−0.104	−0.097
	$\widehat{\alpha}$	MAE	0.554	0.543	0.563	0.127	0.111	0.098
H and K	$\widehat{\beta}$	Bias	0.076	0.044	0.038	0.014	0.007	0.009
	$\widehat{\beta}$	MAE	0.154	0.113	0.086	0.05	0.037	0.027
	$\widehat{\alpha}$	Bias	−0.039	−0.052	−0.035	−0.053	−0.054	−0.041
	$\widehat{\alpha}$	MAE	0.403	0.256	0.178	0.131	0.098	0.075

Logit design: $y_{it}=1(\alpha y_{it-1}+\beta x_{it}+\eta_i+v_{it}\geqslant 0)$; $\beta_0=1$; $\alpha_0=0.5$; $\eta_i=(1/4)\sum_{t=1}^{4}x_{it}$; $x_{it}\sim N(0,\pi^2/3)$; $v_{it}\sim$logistic; 1000 Monte Carlo simulations. Median bias and median absolute error (MAE) are reported.

comparing both with the smallest T size we could have for estimating this kind of models.[14] However for a T as small as 8, the MMLE has a MAE comparable to HK. So, reducing the order of the bias allows us to use a consistent estimator when $T\to\infty$, with samples of moderate T size. As expected, the MMLE does not improve with N as HK does. Compared with MLE, MMLE performs better with $T=4$ than MLE with $T=8$.

I have simulated two different values of α because the larger the α, the greater the serial correlation of y_{it}, and I expect that the estimator performs worse, as it happens with the HK. Results are shown in Table 2. Again, estimates are greatly improved compared with MLE as they were for smaller values of α.

In order to asses the merits of the modification reducing the order of the bias with respect to T, Table 3 reports the results for 16 periods. The MLE of α_0 has still an important bias, however the MMLE is now the best one of the three estimators.

One of the advantages mentioned of MMLE with respect to the estimator proposed by Honoré and Kyriazidou estimator is its generality. The model simulated can be estimated by both methods, but if we want to estimate a probit instead of a logit, HK has to use Manski's insight and Honoré and Kyriazidou (2000) do not report estimates of α and β separately in this case. Nevertheless, MMLE works in the same way and keeps its theoretical properties regardless the distribution of v_{it}, to the extent that the MLE is a general method of estimation for different distributional assumptions. We now have the following model:

$$\Pr(y_{it}=1|\eta_i,y_i^{t-1},x_i)=\Phi(\alpha y_{it-1}+\beta x_{it}+\eta_i)=\Phi_{it}, \tag{33}$$

where Φ is the normal cdf.

[14]Estimates are conditional on the first observation.

Table 2
Logit design with $\alpha_0 = 2$, $T = 8$ and different values of N

		250			500			1000		
		MLE	MMLE	HK	MLE	MMLE	HK	MLE	MMLE	HK
$\widehat{\beta}$	Bias	0.270	0.019	0.016	0.265	0.015	0.014	0.265	0.016	0.016
$\widehat{\beta}$	MAE	0.270	0.045	0.064	0.265	0.032	0.044	0.265	0.023	0.034
$\widehat{\alpha}$	Bias	−0.654	−0.226	−0.195	−0.647	−0.218	−0.179	−0.647	−0.218	−0.16
$\widehat{\alpha}$	MAE	0.654	0.227	0.227	0.648	0.218	0.197	0.647	0.218	0.164

Table 3
Logit design with $T = 16$ and $N = 250$

	$\alpha_0 = 0.5$				$\alpha_0 = 2$			
	Results for $\widehat{\beta}$		Results for $\widehat{\alpha}$		Results for $\widehat{\beta}$		Results for $\widehat{\alpha}$	
	Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE
MLE	0.099	0.099	−0.312	0.312	0.108	0.108	−0.297	0.297
MMLE	0.005	0.023	−0.022	0.067	0.006	0.027	−0.044	0.084
HK	0.005	0.029	−0.053	0.074	−0.003	0.034	−0.200	0.201

Table 4 shows the simulation results for a probit with different values of N, T and α_0. The conclusions are the same as in the logit case and, in terms of MAE, they perform similarly. In Table 4 a situation with 10 periods is included. It is clear how it improves quickly with the number of periods and with 16 the median biases for the MMLE are less than 5% of the true values. Again, the MLE is severely biased even for the 16 periods case.

Estimation of the variance: The most common way of estimating the asymptotic variance–covariance matrix of the estimator in a maximum likelihood framework is using minus the inverse of the Hessian matrix—denoted by $(-H_{NT}^{\dagger})^{-1}$—evaluated at the estimated values. Looking at the asymptotic distribution in Eqs. (27) and (30) and given the asymptotic relations of the MMLE and MLE when both N and T go to infinity, there are four more consistent estimators of the variance–covariance matrix. In Table 5 I report the five measures and the comparison with the variances found in simulation and the percentage of times that the confidence intervals cover the true parameter values for each variance's estimator. The coverage of the intervals is less than 95% mainly because they are not centered. Centering them using the mean bias of the estimates, the coverage rate is very close to 95%. Looking at the results, all of them are quite similar and $(-H_{NT}^{\dagger})^{-1}$ is the easiest choice since it is calculated as part of the optimization process.[15]

3.2. Policy parameters of interest

In binary choice models like (1), β is of interest since some economic hypothesis impose testable restrictions on its sign or magnitude. Also, the β coefficients give the relative

[15]Although we are using a concentrated likelihood, the Hessian has to take into account that the η_i are also being estimated. In Appendix B, I explain how I have addressed this problem.

Table 4
Probit design with different N, T and α_0 values

α_0	T	N	MLE				MMLE			
			Results for $\widehat{\beta}$		Results for $\widehat{\alpha}$		Results for $\widehat{\beta}$		Results for $\widehat{\alpha}$	
			Bias	MAE	Bias	MAE	Bias	MAE	Bias	MAE
0.5	4	250	0.745	0.745	−2.665	2.665	−0.051	0.061	−0.450	0.450
		500	0.739	0.739	−2.634	2.634	−0.047	0.050	−0.434	0.434
		1000	0.715	0.715	−2.596	2.596	−0.053	0.053	−0.432	0.432
0.5	8	250	0.236	0.236	−0.781	0.781	−0.032	0.042	−0.078	0.119
		500	0.230	0.230	−0.777	0.777	−0.036	0.039	−0.077	0.090
		1000	0.232	0.232	−0.780	0.780	−0.035	0.035	−0.081	0.084
0.5	10	250	0.168	0.168	−0.591	0.591	−0.026	0.034	−0.057	0.094
		500	0.164	0.164	−0.578	0.578	−0.029	0.031	−0.044	0.072
0.5	16	250	0.086	0.086	−0.314	0.314	−0.016	0.027	−0.007	0.067
		500	0.089	0.089	−0.329	0.329	−0.013	0.019	−0.022	0.048
2	8	250	0.262	0.262	−0.691	0.691	−0.039	0.046	−0.248	0.248
		500	0.266	0.266	−0.700	0.700	−0.035	0.038	−0.256	0.256
		1000	0.260	0.266	−0.693	0.693	−0.038	0.038	−0.253	0.253
2	10	250	0.195	0.195	−0.536	0.536	−0.035	0.041	−0.174	0.179
		500	0.191	0.191	−0.534	0.534	−0.037	0.039	−0.173	0.174
2	16	250	0.104	0.104	−0.308	0.308	−0.018	0.028	−0.072	0.090
		500	0.108	0.108	−0.317	0.317	−0.014	0.021	−0.080	0.084

$y_{it} = 1(\alpha y_{it-1} + \beta x_{it} + \eta_i + v_{it} \geqslant 0)$; $\beta_0 = 1$; $\eta_i = (1/4)\sum_{t=1}^{4} x_{it}$; $x_{it} \sim \mathrm{N}(0, \pi^2/3)$; $v_{it} \sim \mathrm{N}(0, \pi^2/3)$; 1000 Monte Carlo simulations.

impact of the explanatory variables on the probabilities of $(y_{it} = 1)$. Nevertheless, even though the micropanel literature has emphasized the fixed T consistent estimation of β, in models like model (1) with two x variables, say x_1 and x_2, the effect of a change in x_1 over the expected y for an individual i (or the effect of a change in x_1 over the probability of $(y_{it} = 1)$, or the marginal effect of x_1) is

$$\frac{\partial}{\partial x_1} \mathrm{E}[y_{it}|x, \eta_i] = \frac{\partial}{\partial x_1} F(\beta_1 x_1 + \beta_2 x_2 + \eta_i) = \beta_1 f(\beta_1 x_1 + \beta_2 x_2 + \eta_i) \tag{34}$$

when x_1 is a continuous variable and

$$\mathrm{E}[y_{it}|x_{1b}, x_2, \eta_i] - \mathrm{E}[y_{it}|x_{1a}, x_2, \eta_i] = F(\beta_1 x_{1b} + \beta_2 x_2 + \eta_i) - F(\beta_1 x_{1a} + \beta_2 x_2 + \eta_i) \tag{35}$$

when we want to know the effect of changing x_1 from value x_{1a} to x_{1b}, as it happens if x_1 is a discrete variable. In Eq. (34) f denotes the pdf that corresponds with distribution F. These two measures depend on the levels of all the explanatory variables and on the permanent unobserved heterogeneity η_i. So, the effect differs among individuals due to their unobserved heterogeneity and the values of x and y that each one has.

Table 5
Estimates of the variance

	Mean	RMSE	CI, 95% (%)
	$\widehat{Var(\hat{\alpha})}$		
Value in 1000 simulations	0.011299		
$(H^{\dagger\prime} V^{\dagger-1} H^{\dagger\prime})^{-1}$	0.010755	9.726×10^{-4}	85.5
$(H^{\dagger\prime} V^{-1} H^{\dagger\prime})^{-1}$	0.011786	9.747×10^{-4}	87.7
$(H' V^{-1} H')^{-1}$	0.011734	9.407×10^{-4}	87.2
$-H^{-1}$	0.011212	4.123×10^{-4}	86.7
$-H^{\dagger-1}$	0.011259	4.219×10^{-4}	88.9
	$\widehat{Var(\hat{\beta})}$		
Value in 1000 simulations	0.001596		
$(H^{\dagger\prime} V^{\dagger-1} H^{\dagger\prime})^{-1}$	0.001682	1.908×10^{-4}	94.2
$(H^{\dagger\prime} V^{-1} H^{\dagger\prime})^{-1}$	0.001815	2.827×10^{-4}	94.8
$(H' V^{-1} H')^{-1}$	0.001781	2.464×10^{-4}	94.4
$-H^{-1}$	0.001696	1.497×10^{-4}	94.0
$-H^{\dagger-1}$	0.001715	1.712×10^{-4}	94.2

$y_{it} = 1(\alpha y_{it-1} + \beta x_{it} + \eta_i + v_{it} \geqslant 0)$; $N = 500$; $T = 8$; $\alpha_0 = 0.5$; $\beta_0 = 1$; $\eta_i = (1/4)\sum_{t=1}^{4} x_{it}$; $x_{it} \sim \mathrm{N}(0, \pi^2/3)$; $v_{it} \sim$logistic; 1000 Monte Carlo simulations.
Column CI 95%: percentage of 95% confidence intervals that contain the true value of the parameter across 1000 simulations. Intervals based on the normal asymptotic distribution. Intervals are not centered, so CI = $\hat{\alpha} \pm 1.96 \times \widehat{Var(\hat{\alpha})}$ and $\hat{\beta} \pm 1.96 \times \widehat{Var(\hat{\beta})}$. Mean bias $\hat{\alpha}_{\mathrm{MMLE}} = -0.090$; Mean bias $\hat{\beta}_{\mathrm{MMLE}} = 0.015$.

Usually, the mean effect for all individuals is what people want to calculate. But more than one mean can be considered, depending on the economic question. Some average measures found in literature are:

- The effect of an increment in x_1 over the probability of $y = 1$, for an individual with the average characteristics:

$$F(\beta_1(\mathrm{E}(x_{1it}) + 1) + \beta_2 \mathrm{E}(x_{2it}) + \mathrm{E}(\eta_i)) - F(\beta_1 \mathrm{E}(x_{1it}) + \beta_2 \mathrm{E}(x_{2it}) + \mathrm{E}(\eta_i)). \tag{36}$$

- The expected effect over the probability of $y = 1$ of going from x_{1a} to x_{1b} is

$$\mathrm{E}_{(\eta,x_2)|x_1}(\text{Eq. (35)}) = \int_{\eta_i, x_{2i}} [F(\beta_1 x_{1b} + \beta_2 x_2 + \eta_i) - F(\beta_1 x_{1a} + \beta_2 x_2 + \eta_i)] \, \mathrm{d}G_{(\eta,x_2)|x_1}(\eta_i, x_2 | x_{1a}), \tag{37}$$

  where G is the conditional distribution function. This is the parameter of interest estimated in Altonji and Matzkin (2001).
- Taking the average over the marginal distribution of η,

$$\int_{x_2} \int_{\eta} [F(\beta_1(x_{1b}) + \beta_2 x_2 + \eta_i) - F(\beta_1 x_{1a} + \beta_2 x_2 + \eta_i)] \, \mathrm{d}G_{\eta}(\eta_i) \, \mathrm{d}G_{x_2|x_1}(x_2 | x_{1a}). \tag{38}$$

Table 6
Sample counterparts of the population parameters of interest

Pop.	Sample counterpart
(36)	$[F(\beta_1(\overline{x}_1+1)+\beta_2\overline{x}_2+\overline{\eta})-F(\beta_1\overline{x}_1+\beta_2\overline{x}_2+\overline{\eta})]$
(37)	$(1/Na)\sum_{i=1}^{N}[F(\beta_1 x_{1b}+\beta_2 x_{2it}+\eta_i)-F(\beta_1 x_{1a}+\beta_2 x_{2it}+\eta_i)]1_{\{x_{1it}=x_{1a}\}}$[a]
(38)	$(1/Na)\sum_{i=1}^{N}\{(1/N)\sum_{j=1}^{N}[F(\beta_1 x_{1b}+\beta_2 x_{2it}+\eta_j)-F(\beta_1 x_{1a}+\beta_2 x_{2it}+\eta_j)]\}1_{\{x_{1it}=x_{1a}\}}$
(39)	$(1/N)\sum_{i=1}^{N}[F(\beta_1 x_{1b}+\beta_2 x_{2it}+\eta_i)-F(\beta_1 x_{1a}+\beta_2 x_{2it}+\eta_i)]$ where $\overline{x}=(1/N)\sum_{i=1}^{N}x_{it}$, $\overline{\eta}=(1/N)\sum_{i=1}^{N}\eta_i$ and $Na=\sum_{i=1}^{N}1\{x_{1it}=x_{1a}\}$.

[a]When x is a continuous variable, $1\{x_{it}=x_a\}$ will be substituted by a kernel density function.

This measure corresponds with the derivative of the average structural function (ASF) defined in Blundell and Powell (2003).[16]

- Chamberlain (1984) proposed as a parameter of interest the mean effect for a randomly drawn individual:

$$\int_{\eta_i,x_2}[F(\beta_1 x_{1b}+\beta_2 x_{2it}+\eta_i)-F(\beta_1 x_{1a}+\beta_2 x_{2it}+\eta_i)]\,\mathrm{d}G(\eta_i,x_{2it}). \tag{39}$$

All the above are population measures. If we have a random sample of (y_{it},x_{it},η_i) $i=1,\ldots,N;t=0,\ldots,T-1$, knowing β, the sample counterparts, for the effects on a specific period t, are in Table 6.

In Table 7 I show some of the parameters of interest described and their estimates by MMLE and MLE for a simulated sample. They can all be estimated, just replacing α, β and η_i by their estimates. The measure corresponding to Eq. (38) is different from the one corresponding to Eq. (37) because (38) is using the marginal distribution of the fixed effect; therefore, it is ignoring that there is positive correlation between the explanatory variables and the fixed effects. The MMLE is clearly improving the estimation comparing it with the MLE.

On the other hand, depending on the economic matter analyzed, we may need not only the mean, but also other descriptive statistics such as the variance, the percentiles or even the whole distribution of the effect on the population. In addition to that, the mean is very descriptive (in a statistical sense) in most of the linear models found in the literature but it may not capture relevant features of the distribution in binary choice models. In models like (1) individuals choose one option depending on whether they are above or below a threshold, and a change in x produces a change in the probability of being above that threshold. This means a greater effect on those who are close to the threshold and a small effect for those who are far away from the threshold. The mean effect may be between those two groups of individuals and it may not be relevant for most of the population.

[16]Blundell and Powell (2003) consider models with endogenous regressors. In model (1) the unobservable part has two components: an exogenous shock v_{it} and permanent unobserved heterogeneity η_i possibly correlated with the regressors. So the endogeneity in this case comes from η_i.

Table 7
Mean effects of $y_{it-1} = 1$ on the probability of $y_{it} = 1$ computed according to the different measures presented in Table 6

	Value	MMLE	MLE
$y_{it-1} = 0$			
Joint (37)	0.199	0.164	0.059
Marginal (38)	0.162	0.133	0.046
$y_{it-1} = 1$			
Joint (37)	0.198	0.157	0.056
Marginal (38)	0.160	0.132	0.046
Average (39)	0.198	0.160	0.057

$y_{it} = 1\{\alpha y_{it-1} + \beta x_{it} + \eta_i + v_{it} \geqslant 0\}$. Dynamic logit case: $\alpha_0 = 1, \beta_0 = -1$, $\eta_i = \mathrm{N}(0,1)$, $x_{it} = \eta_i + \mathrm{N}(0,1)$, v_{it} is i.i.d. logistically distributed. $N = 10\,000$, $T = 8$: Effects of $y_{it-1} = 1$ on the probability of $y_{it} = 1$: $\{F(\alpha + \beta x_{it} + \eta_i) - F(\beta x_{it} + \eta_i)\}$. The numbers in parentheses refer to the equation that define each measure.

Depending on the kind of economic study, we may only be interested on the effect over people with certain characteristics and situations, for example those who are near the threshold. In such case, the mean is not only a non-representative measure, but also could lead us to misleading conclusions.

I conduct a Monte Carlo experiment with 1000 replications of the logit model (31), with $\alpha_0 = 1$, $\beta_0 = 1.5$, $\eta_i = (x_{i0} + x_{i1} + x_{i2} + x_{i3})/2$, $N = 500$ and eight periods. Results on the estimation of the quantiles of the distribution of the effect of a change in x_{iT-1} over the probability of $(y_{iT-1} = 1)$ are in Table 8. The results are reported for all individuals in the full simulated sample with the true parameters' values, for the sample of movers, i.e. for the sample actually used on the estimation, with the true parameters' values, with the ML estimates of the parameters and with the modified maximum likelihood estimates of the parameters. The sample of movers excludes those i observations whose sum of y_{it} for the last $T-1$ periods is equal to zero or $T-1$. We call these individuals stayers, as opposed to movers, because they take the same decision over all the sample period. In this experiment the proportion of stayers is around 30%. The improvement using the MML estimates compared with ML estimates at all quantiles is clear in terms of both bias reduction and root mean square error, particularly at the highest quantiles. The MML estimates are quite close to the true value based on the sample of movers. However, if we compared them with the true value based on the full simulated sample, there are more differences for the lowest quantiles. Those that are less affected by a change on the explanatory variables, are those with higher probability of not changing their decision on the sample period. It is important to note that this problem, as the estimation of the model's parameters on finite samples, depends on the number of periods available. The estimation of the parameters of interest by MML presented in Tables 7 and 8, are consistent when T goes to infinity and they clearly improve the finite sample performance of the MLE.

Fig. 1 is the smoothed distribution of the effects described in Table 8, but from a simulation with $N = 10\,000$. The main feature is the bi-modality. Individuals around the first mode are those with a small effect due to that the levels of their observable variables or the levels of their individual effects are such that they have very high or very small probabilities of $y_{it} = 1$, and a change in x_{iT} scarcely affects them. Observations around the

Table 8
Quartiles of the distribution of the effect of a change in x_{iT-1} over the probability of $(y_{iT-1} = 1)$

	Full	Movers	MLE	MMLE
Minimum	0.0000	0.0001	0.0000	0.0001
Mean Bias			−0.0001	0.0000
RMSE			0.0002	0.0001
25%	0.0266	0.0395	0.0166	0.0385
Mean Bias			−0.0229	−0.0010
RMSE			0.0236	0.0059
Median	0.1195	0.1475	0.1189	0.1461
Mean Bias			−0.0286	−0.0013
RMSE			0.0327	0.0124
75%	0.2788	0.2968	0.3495	0.2909
Mean Bias			0.0526	−0.0059
RMSE			0.0572	0.0160
Maximum	0.3750	0.3750	0.5057	0.3667
Mean Bias			0.1307	−0.0083
RMSE			0.1337	0.0168

1000 Monte Carlo replications of the logit model (31), with $\alpha_0 = 1$, $\beta_0 = 1.5$, $\eta_i = (x_{i0} + x_{i1} + x_{i2} + x_{i3})/2$, $N = 500$ and $T = 8$. Mean bias and RMSE calculated with respect to the sample of movers, which is the sample actually used on the estimation of model's parameters.

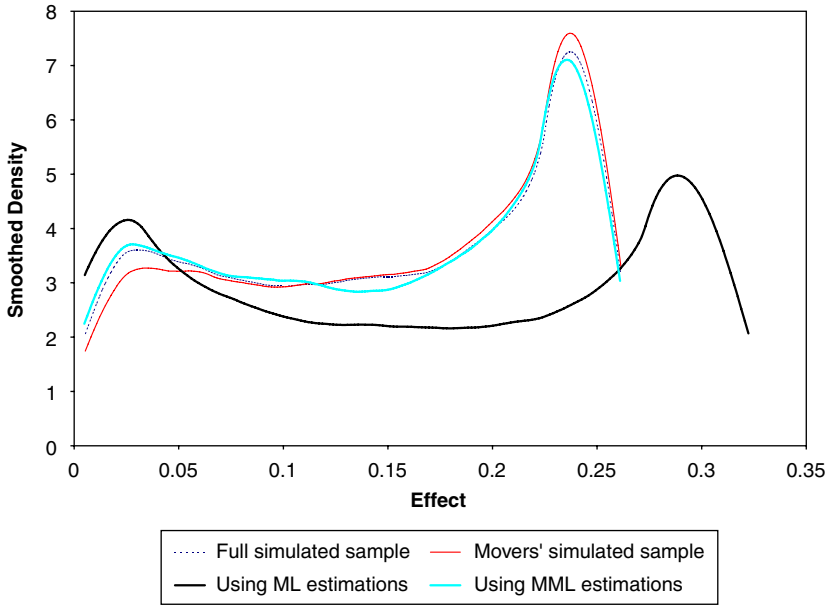

Fig. 1. Smoothed density of the effect of a change in x_{iT-1} over the probability of $(y_{iT-1} = 1)$ for a simulated sample from the logit design, with $T = 8$ and $N = 10000$.

second mode are those with higher effect and, in this simulation, most of the individuals are in that region. The mean effect (0.15) is between those two groups and it is only relevant for a very small part of the population. Although $\widehat{\beta}_{\text{MLE}}$ is severely biased, looking

at the mean effect based on ML estimations (0.16), it does not have much bias in this particular case because the MLE overestimates the frequency density of the smallest effects and the greatest effects, being balanced on average. Comparing the four densities, it can be noticed that the MLE misestimate the second mode and the densities around it significantly, whereas MMLE describes the distribution more accurately.

4. Empirical illustration

In this section, I illustrate the use of the modified maximum likelihood method by estimating an empirical model of female labor force participation. This empirical illustration is similar to some of the specifications estimated in Hyslop (1999), although there are some differences that makes a direct comparison difficult. Essentially, Hyslop uses a different sample period, random effects instead of fixed effects and AR(1) instead of white noise errors. In this empirical illustration, as in Hyslop (1999), children variables are assumed to be strictly exogenous with respect to ε_{it} in Eq. (40). However, children variables are allowed to be endogenous with respect to η_i. Moreover, in contrast to random effects approaches, no restrictions are placed on the form of the dependence between effects and children variables.[17]

I use data on 1461 married women corresponding to waves 12–22 of the Panel Study of Income Dynamics (PSID). Sample information is for 10 calendar years 1979–1988. Only women continuously married, aged between 18 and 60 in 1985 and whose husband is a labor force participant in each of the sample years, were included in the sample.[18]

The estimated equation is

$$y_{it} = 1\{\alpha y_{it-1} + x_{it}'\beta + \eta_i + \varepsilon_{it} \geqslant 0\} \quad (t = 0, \ldots, T-1;\ i = 1, \ldots, N), \tag{40}$$

y_{it} takes value one if women i participate in period t and zero otherwise. $x_{it} = (\#children0{-}2_{it-1}, \#children0{-}2_{it}, \#children3{-}5_{it}, \#children6{-}17_{it}, \log income_{it}$, time dummies, and a quadratic function of age), where $\#childrena{-}b$ is the number of children aged between a and b, $\log income$ is the log of husband's labor income deflated by Consumer Price Index and age is wife's age. ε_{it} is assumed to be an independent and identically distributed normal variable. So it is a dynamic probit model.

Most women in the sample participate at least one period. Just 8% never participate. Almost half of the them participate all the last nine periods. We are conditioning on the first observation to avoid the sample initial conditions problem.

Table 9 shows the results of the estimation of model (40) by MLE and MMLE. There are significant differences on the estimated parameters. The MLE is underestimating the true state dependence effect and overestimating the effect of the other variables. As a result, the impact of previous participation on the probability of participating is estimated to be, in absolute value approximately 1.4 times the impact of a child aged between 0 and 2 using MLE and 2.7 times using the MMLE. So, the estimate by MLE of the impact of previous participation relative to the impact of a child aged between 0 and 2 on the

[17] In Carro (2003) I study the same problem as in this empirical illustration but I consider more general specifications and assumptions. For example, I take into account specifically on the estimation of the model that the number of children variable could be affected by past participation decisions.

[18] As in Hyslop (1999), an individual is defined as a participant if they report both positive annual hours worked and annual labor earnings.

Table 9
Estimates of some of the parameters of model (40)

	α	#Children 0–2$_t$	#Children 3–5	#Children 6–17	Log(income)
MLE	0.753	−0.534	−0.283	−0.078	−0.253
	(0.043)	(0.064)	(0.055)	(0.043)	(0.055)
MMLE	1.081	−0.400	−0.183	−0.038	−0.209
	(0.042)	(0.058)	(0.050)	(0.039)	(0.051)

Standard errors are in parentheses.

probability of participating, which is approximately given by the ratio α/β, is a half of the value obtained when using MMLE.

5. Conclusion

I have applied the modified maximum likelihood estimator (MMLE) to dynamic panel data discrete choice models with fixed effects, using a modification expressed in terms of the original parameters of the model as in Arellano (2003). I have shown that the MMLE reduces the bias of the estimated parameters from $O(T^{-1})$ to $O(T^{-2})$ (without increasing the asymptotic variance), even if there is no information orthogonal reparametrization. Given that reduction on the order of the bias, the finite sample bias may be negligible for moderate T and the estimator has good asymptotic properties (in an N and T asymptotic) even in situations in which N grows faster than T. Monte Carlo experiments have shown that there is a small bias in probit and logit models with a lag of the endogenous variable and exogenous variables for eight time periods.

One of the main advantages of this approach over other methods for estimating panel data binary choice models is its generality. For example, it is not restricted to the logistic case. This method is generally applicable and it has the same asymptotic properties regardless of the distribution of the errors.

In addition to that, MMLE allows to get sensible estimates of the different policy parameters of interest considered in the literature: summary measures of the effect of a change in x over the probability of $y = 1$. In contrast to linear models, the expected effect is different for each individual and it depends on the fixed effects. The mean of that effect across all individuals may not be the parameter of interest, but we may need the distribution of the effect of a explanatory variable. Using MML estimates of model's parameters improves significantly the estimation of that distribution with respect to the ML case. Another advantage of the approach considered in the paper is that the fixed effect, needed for the calculation of the parameters of interest, is estimated as part of the estimation process whereas in the fixed T consistent estimation it is not. Also, asymptotics in both N and T has to be considered because the estimation of the marginal effect is consistent only when $T \to \infty$.

Acknowledgments

I am especially indebted to Manuel Arellano for his valuable advice and comments. I would like to thank an anonymous referee, an associate editor and the editor for

comments that greatly improved the paper, and Pedro Albarran, Cristina Barcelo and Pedro Mira for discussions and support. Thanks are also due to seminar participants at CEMFI, Universidad Carlos III, UCL, Universitat Pompeu Fabra, Universidad de Alicante, CAM-University of Copenhagen and "The Evaluation of Labour Market Policies" conference of the Network of Excellence in Amsterdam for useful comments. All remaining errors are mine.

Appendix A. Computation of the modified score

Let us consider the logit model used in the Monte Carlo experiments and implement the modification on it

$$y_{it} = 1\{\alpha y_{it-1} + \beta x_{it} + \eta_i + v_{it} \geqslant 0\} \quad (t = 0, \ldots, T-1; i = 1, \ldots, N), \tag{A.1}$$

where x_{it} is a vector of exogenous variables, η_i is an unobservable individual specific effect and $-v_{it}$ is independently distributed with cdf F conditional on η_i, $y_i^{t-1} = (y_{i0}, \ldots, y_{it-1})'$ and $x_i = (x_{i1}, \ldots, x_{it})'$, so that

$$\Pr(y_{it} = 1 | \eta_i, y_i^{t-1}, x_i) = F(\alpha y_{it-1} + \beta x_{it} + \eta_i) = F_{it}, \tag{A.2}$$

F is the logistic cdf.

From Eq. (4), an individual's modified score for α is

$$\begin{aligned} d_{\alpha Mi}(\alpha, \beta) = d_{\alpha \mathrm{Ci}}(\alpha, \beta) &- \frac{1}{2} \frac{d_{\alpha\eta\eta i}(\alpha, \beta, \widehat{\eta}_i(\alpha, \beta)) + d_{\eta\eta\eta i}(\alpha, \beta, \widehat{\eta}_i(\alpha, \beta)) \widehat{\partial \eta}_i(\alpha, \beta)/\partial \alpha}{d_{\eta\eta i}(\alpha, \beta, \widehat{\eta}_i(\alpha, \beta))} \\ &+ \frac{(\partial/\partial \eta_i)(\mathrm{E}[d_{\alpha\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i])}{\mathrm{E}[d_{\eta\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i]} \Big|_{\eta_i = \widehat{\eta}_i(\alpha, \beta)} \\ &- \frac{\mathrm{E}[d_{\alpha\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i]}{\mathrm{E}[d_{\eta\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i]} \Big|_{\eta_i = \widehat{\eta}_i(\alpha, \beta)} \\ &\times \frac{(\partial/\partial \eta_i)(\mathrm{E}[d_{\eta\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i])}{\mathrm{E}[d_{\eta\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i]} \Big|_{\eta_i = \widehat{\eta}_i(\alpha, \beta)}, \end{aligned} \tag{A.3}$$

where $d_{\alpha \mathrm{Ci}}(\alpha, \beta)$ is an individual's score from the concentrated likelihood,

$$d_{\alpha \mathrm{Ci}}(\alpha, \beta) = \frac{\partial l_i(\alpha, \beta, \eta_i(\gamma))}{\partial \alpha} = \sum_{t=1}^{T-1} \left(y_{it-1} + \frac{\widehat{\partial \eta}_i(\alpha, \beta)}{\partial \alpha} \right) [y_{it} - F(\alpha y_{it-1} + \beta x_{it} + \eta_i)], \tag{A.4}$$

$$d_{\alpha\eta i}(\alpha, \beta, \eta_i) = \frac{\partial^2 l_i}{\partial \alpha \partial \eta_i} = \sum_{t=1}^{T-1} y_{it-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i), \tag{A.5}$$

$$d_{\eta\eta i}(\alpha, \beta, \eta_i) = \frac{\partial^2 l_i}{\partial \eta_i^2} = -\sum_{t=1}^{T-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i), \tag{A.6}$$

f is the logistic pdf.

$$\mathrm{E}[d_{\alpha\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i] = -\sum_{t=1}^{T-1} \mathrm{E}[y_{it-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i) | y_{i0}, \eta_i, x_i], \tag{A.7}$$

$$\mathrm{E}[d_{\eta\eta i}(\alpha,\beta,\eta_i)|y_{i0},\eta_i,x_i] = -\sum_{t=1}^{T-1}\mathrm{E}[f(\alpha y_{it-1}+\beta x_i+\eta_i)|y_{i0},\eta_i,x_i], \tag{A.8}$$

$$\mathrm{E}[y_{it-1}f(\alpha y_{it-1}+\beta x_{it}+\eta_i)|y_{i0},\eta_i.x_i] = f(\alpha+\beta x_{it}+\eta_i)\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i), \tag{A.9}$$

$$\begin{aligned}&\mathrm{E}[f(\alpha y_{it-1}+\beta x_{it}+\eta_i)|y_{i0},\eta_i,x_i]\\ &\quad = f(\alpha+\beta x_{it}+\eta_i)\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)+f(\beta x_{it}+\eta_i)(1-\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i))\\ &\quad = \Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)(f(\alpha+\beta x_{it}+\eta_i)-f(\beta x_{it}+\eta_i))+f(\beta x_{it}+\eta_i).\end{aligned} \tag{A.10}$$

$\Pr(y_{it}=1|y_{i0},\eta_i,x_i)$ can be calculated recursively from:

$$\Pr(y_{i1}=1|y_{i0},\eta_i,x_i) = F(\alpha y_{i0}+\beta x_{i1}+\eta_i), \quad \text{starting point. For } t>1: \tag{A.11}$$

$$\begin{aligned}\Pr(y_{it}=1|y_{i0},\eta_i,x_i) = {}& \Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)(F(\alpha+\beta x_{it}+\eta_i)\\ &- F(\beta x_{it}+\eta_i))+F(\beta x_{it}+\eta_i).\end{aligned}$$

From (A.10),

$$\begin{aligned}&\frac{\partial}{\partial\eta_i}\mathrm{E}[f(\alpha y_{it-1}+\beta x_{it}+\eta_i)|y_{i0},\eta_i,x_i]\\ &\quad = \frac{\partial}{\partial\eta_i}\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)(f(\alpha+\beta x_{it}+\eta_i)-f(\beta x_{it}+\eta_i))\\ &\qquad + \Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)(f'(\alpha+\beta x_{it}+\eta_i)-f'(\beta x_{it}+\eta_i))+f'(\beta x_{it}+\eta_i).\end{aligned} \tag{A.12}$$

From (A.9),

$$\begin{aligned}&\frac{\partial}{\partial\eta_i}\mathrm{E}[y_{it-1}f(\alpha y_{it-1}+\beta x_{it}+\eta_i)|y_{i0},\eta_i,x_i]\\ &\quad = f'(\alpha+\beta x_{it}+\eta_i)\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)\\ &\qquad + f(\alpha+\beta x_{it}+\eta_i)\frac{\partial}{\partial\eta_i}\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i).\end{aligned} \tag{A.13}$$

$(\partial/\partial\eta_i)\Pr(y_{it}=1|y_{i0},\eta_i,x_i)$ are calculated recursively from:

$$\begin{aligned}\frac{\partial}{\partial\eta_i}\Pr(y_{it}=1|y_{i0},\eta_i,x_i) = {}& \frac{\partial}{\partial\eta_i}\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)(F(\alpha+\beta x_{it}+\eta_i)-F(\beta x_{it}+\eta_i))\\ &+\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)(f(\alpha+\beta x_{it}+\eta_i)\\ &-f(\beta x_{it}+\eta_i))+f(\beta x_{it}+\eta_i), \quad \text{for } t>1,\end{aligned} \tag{A.14}$$

$$\frac{\partial}{\partial\eta_i}\Pr(y_{i1}=1|y_{i0},\eta_i,x_i) = f(\alpha y_{i0}+\beta x_{i1}+\eta_i). \tag{A.15}$$

From the first order condition of η_i, $d_{\eta_i}(\alpha,\beta,\eta_i)=\sum_{t=1}^{T-1}(y_{it}-F_{it})$, $\widehat{\eta}_i(\alpha,\beta)$, solves

$$\sum_{t=1}^{T-1}y_{it} = \sum_{t=1}^{T-1}F(\alpha y_{it-1}+\beta x_{it}+\eta_i). \tag{A.16}$$

Deriving the previous equation with respect to α

$$0 = \sum_{t=1}^{T-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i)\left(\frac{\partial \widehat{\eta}_i(\alpha,\beta)}{\partial \alpha} + y_{it-1}\right).$$

Therefore,

$$\frac{\partial \widehat{\eta}_i(\alpha,\beta)}{\partial \alpha} = \frac{-\sum_{t=1}^{T-1} y_{it-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i)}{\sum_{t=1}^{T-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i)}. \tag{A.17}$$

The modified first order condition for β is calculated in the same way. In the logistic case $f_{it} = F_{it} * (1 - F_{it})$, which simplifies the first order condition of the likelihood, but these recursive procedures for computing the expectations needed for the modification work regardless of the density function f assumed.

Appendix B. Concentrating the likelihood and estimating with fixed effects

A problem that arises on the maximization of the log-likelihood function

$$\begin{aligned}\log L = \sum_{i=1}^{N}\sum_{t=1}^{T-1}\{&y_{it} * \log F(\alpha y_{it-1} + x_{it}\beta + \eta_i) \\ &+ (1 - y_{it}) * \log(1 - F(\alpha y_{it-1} + \beta x_{it} + \eta_i))\}\end{aligned} \tag{B.1}$$

is that we have to estimate N parameters corresponding to the fixed effects, implying a second derivative matrix with $N + 2$ rows and columns. A way of proceeding is using some results from matrix algebra suggested in Chamberlain (1980) and Greene (2004), in order to simplify the computation of the inverse of the Hessian. Alternatively Heckman and MaCurdy (1980) divided the optimization problem in two problems: one for α and β and another for $\{\eta_i\}_{i=1}^{N}$.

In this paper I compute both the MLE and the MMLE from the first order conditions of the concentrated likelihood, so I do not divide the procedure in two estimation problems. Since, due to non-linearity, we cannot get a explicit expression of the fixed effects estimators as functions of α and β, I make numerical substitution of them on the estimating of $\gamma = \alpha\beta'$ i.e. the estimator of γ solves

$$\sum_{i=1}^{N}\left\{d_{\gamma i}(\gamma, \widehat{\eta}_i(\gamma)) + d_{\eta_i}(\gamma, \widehat{\eta}_i(\gamma))\,\frac{\partial \widehat{\eta}_i(\gamma)}{\partial \gamma}\right\} = \sum_{i=1}^{N} d_{\gamma i}(\gamma, \widehat{\eta}_i(\gamma)) = 0, \tag{B.2}$$

where $\widehat{\eta}_i(\gamma)$ is the number that makes $d_{\eta_i}(\gamma, \eta_i) = 0$ for the value of γ in which we are evaluating the estimating equations; $d_{\eta_i}(\gamma, \eta_i) \equiv \partial l_i(\gamma, \eta_i)/\partial \eta_i$ and $d_{\gamma i}(\gamma, \eta_i) \equiv \partial l_i(\gamma, \eta_i)/\partial \gamma$. We use a Gauss–Newton-type algorithm to solve Eq. (B.2) with respect to γ, i.e. the value of γ that maximizes the function whose first order condition is (B.2). In each step of the algorithm $\widehat{\eta}_i$ is computed such that for the value of γ in that step (γ_s), $d_{\eta_i}(\gamma_s, \eta_i)$ equals zero. Thus, the equation for each of the η_i is nested in the algorithm that maximizes the concentrated likelihood. In each step, we have to solve N single non-linear equations, one for each of the fixed effects. $d_{\eta_i}(\gamma_s, \eta_i) = 0$ is easily solved by bracketing and bisection, and we use that N times. This method is faster than a Gauss–Newton-type procedure for this N problems. Here, we need to bracket the root of the equation. This can be done because we have some knowledge about the form of the equation since we know F and its derivatives.

The difference with respect to Heckman and MaCurdy's suggestion is that maximization with respect to α and β is not made for each given estimated value of the fixed effects. Instead of that the values of the fixed effects are changed accordingly in each step of the estimation process of α and β; just as if we were able to analytically find $\widehat{\eta}_i(\gamma)$.

To estimate the variance correctly, we take advantage of the fact that the equation $d_{\eta_i}(\gamma_s, \eta_i) = 0$ is nested on the algorithm. Thus, we calculate the second derivatives accounting for the fixed effects. That is, deriving (B.2) with respect to γ, the Hessian is equal to

$$\sum_{i=1}^{N}\left\{\frac{\partial^2 l_i(\gamma,\widehat{\eta}_i(\gamma))}{\partial\gamma\partial\gamma}+\left.\frac{\partial^2 l_i(\gamma,\eta_i)}{\partial\gamma\partial\eta_i}\right|_{\eta_i=\widehat{\eta}_i(\gamma)}\frac{\partial\widehat{\eta}_i(\gamma)}{\partial\gamma}\right.$$
$$\left.+\left(\left.\frac{\partial^2 l_i(\gamma,\eta_i)}{\partial\eta_i\partial\gamma}\right|_{\eta_i=\widehat{\eta}_i(\gamma)}+\left.\frac{\partial^2 l_i(\gamma,\eta_i)}{\partial\eta_i\partial\eta_i}\right|_{\eta_i=\widehat{\eta}_i(\gamma)}\frac{\partial\widehat{\eta}_i(\gamma)}{\partial\gamma}\right)\frac{\partial\widehat{\eta}_i(\gamma)}{\partial\gamma}+d_{\eta_i}(\gamma,\widehat{\eta}_i(\gamma))\frac{\partial^2\widehat{\eta}_i(\gamma)}{\partial\gamma\partial\gamma}\right\},$$

$d_{\eta_i}(\gamma,\widehat{\eta}_i(\gamma))(\partial^2\widehat{\eta}_i(\gamma)/\partial\gamma\partial\gamma) = 0$ because $d_{\eta_i}(\gamma,\eta_i) = 0$ at $\eta_i = \widehat{\eta}_i(\gamma)$.

Everything is the same for the MMLE, just replacing $d_{\gamma i}(\gamma,\widehat{\eta}_i(\gamma)) = \partial l_i(\gamma,\eta_i)/\partial\gamma|_{\eta_i=\widehat{\eta}_i(\gamma)}$ by the modified first order condition presented in the paper, $d_{\gamma Mi}(\gamma)$.

References

Akahira, M., Takeuchi, K., 1982. On asymptotic deficiency of estimators in pooled samples in the presence of nuisance parameters. Statistics and Decisions 1, 17–38.

Altonji, J.G., Matzkin, R.L., 2001. Panel data estimators for nonseparable models with endogenous regressors. Technical Working Paper 267, National Bureau of Economic Research.

Alvarez, J., Arellano, M., 2003. The time series and cross-section asymptotics of dynamic panel data estimators. Econometrica 71, 1121–1159.

Arellano, M., 2003. Discrete choice with panel data. Investigaciones Económicas XXVII (3), 423–458 ⟨ftp://ftp.funep.es/InvEcon/paperArchive/Sep2003/v27i3a1.pdf⟩.

Arellano, M., 2005. The Cox–Reid modified score: a comment. Manuscript, CEMFI, Madrid.

Arellano, M., Honoré, B., 2001. Panel data models: some recent developments. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 5. Elsevier Science, Amsterdam.

Blundell, R., Powell, J.L., 2003. Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont, M., Hansen, L.P., Turnovsky, S.J. (Eds.), Advances in Economics and Econometrics, Theory and Applications, Eighth World Congress, vol. II. Cambridge University Press, Cambridge.

Carro, J.M., 2003. Intertemporal female labor force participation with non-exogenous children. Mimeo, CEMFI, Madrid.

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

Chamberlain, G., 1984. Panel data. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. 2. Elsevier Science, Amsterdam.

Chamberlain, G., 1992. Binary response models for panel data: identification and information. Manuscript, Department of Economics, Harvard University.

Cox, D.R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. Journal of the Royal Statistical Society, Series B 49, 1–39.

Ferguson, H., 1992. Asymptotic properties of a conditional maximum-likelihood estimator. The Canadian Journal of Statistics 20 (1), 63–75.

Ferguson, H., Reid, N., Cox, D.R., 1991. Estimating equations from modified profile likelihood. In: Godambe, V.P. (Ed.), Estimating Functions. Oxford University Press, Oxford.

Greene, W., 2004. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. The Econometrics Journal 7, 98–119.

Hahn, J., 2001. The information bound of a dynamic panel logit model with fixed effects. Econometric Theory 17, 913–932.

Heckman, J.J., 1981a. Statistical models for discrete panel data. In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA.

Heckman, J.J., 1981b. The incidental parameters problem and the problem of initial conditions in estimating a discrete-time data stochastic process. In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, New York.

Heckman, J.J., MaCurdy, T.E., 1980. A life cycle model of female labour supply. Review of Economic Studies 47, 47–74.

Honoré, B., 2002. Nonlinear models with panel data. Portuguese Economic Journal 1, 163–173.

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

Honoré, B., Tamer, E., 2004. Bounds on parameters in dynamic discrete choice models. CAM Working paper 2004-23, University of Copenhagen.

Hyslop, D.R., 1999. State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. Econometrica 67, 1255–1294.

Lancaster, T., 2002. Orthogonal parameters and panel data. Review of Economic Studies 69, 647–666.

Li, H., Lindsay, B.G., Waterman, R.P., 2003. Efficiency of projected score methods in rectangular array asymptotics. Journal of the Royal Statistical Society, Series B 65, 191–208.

Manski, C., 1987. Semiparametric analysis of random effects linear models from binary panel data. Econometrica 55, 357–362.

McCullagh, P., 1987. Tensor Methods in Statistics. Chapman & Hall, London.

Neyman, J., Scott, E.L., 1948. Consistent estimates based on partially consistent observations. Econometrica 16, 1–32.

Woutersen, T., 2001. Robustness against incidental parameters and mixing distributions. Manuscript, Department of Economics, University of Western Ontario.