# ESTIMATING DYNAMIC PANEL DATA DISCRETE CHOICE MODELS WITH FIXED EFFECTS

Jesús M. Carro

CEMFI Working Paper No. 0304

January 2003

CEMFI
Casado del Alisal 5; 28014 Madrid
Tel. (34) 914 290 551. Fax (34) 914 291 056
Internet: www.cemfi.es

I am especially indebted to Manuel Arellano for his valuable advice and comments. I would like to thank Pedro Albarran, Cristina Barcelo and Pedro Mira for helpful discussions and support. Thanks are also due to seminar participants at CEMFI, Universidad Carlos III, UCL and "The Evaluation of Labour Market Policies" conference of the Network of Excellence in Amsterdam for useful comments. All errors are mine.

CEMFI Working Paper No. 0304
January 2003

# ESTIMATING DYNAMIC PANEL DATA DISCRETE CHOICE MODELS WITH FIXED EFFECTS

Abstract

In this paper, I consider the estimation of dynamic binary choice panel data models with fixed effects. I use a Modified Maximum Likelihood Estimator (MMLE) that reduces the order of the bias in the Maximum Likelihood Estimator from $O(T^{-1})$ to $O(T^{-2})$, without increasing the asymptotic variance. I evaluate its performance in finite samples where $T$ is not large, using Monte Carlo simulations. 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. Identification issues about policy parameters of interest that arise in this kind of models are also addressed. In contrast with linear models, parameters of interest typically depend on the distribution of the individual effects. I discuss the relevance of mean effects across individuals and show an instance in which the entire distribution is needed. Compared with simple MLE, simulation results show that MMLE improves significantly the estimation of the distribution of the effect of interest.

*JEL Codes*: C23, J22.
*Keywords*: Panel data, dynamic discrete choice, fixed effects, modified MMLE.

Jesús M. Carro
CEMFI
carro@cemfi.es

# 1 Introduction

This paper deals with the estimation of dynamic discrete choice models with fixed effects. These models, that take into account unobserved permanent heterogeneity and the underlying 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. That observed persistence may be due to persistence on observable individual characteristics, true state dependence or permanent unobserved heterogeneity. In the last two cases we observe that for given observable characteristics individuals choose an option more frequently when they have chosen it in the past. However, these two sources of persistence in individual decisions have very different implications and we want to separably identify each of them and estimate their relative importance. There is true state dependence if previous choices affects current utility. In contrast, if the source of persistence is permanent unobserved heterogeneity individuals have higher propensity to take that decision, but there is no effect of previous choices on current utility and past experience has no a behavioral effect (see Heckman, 1981a). This is important not only to study persistence, but also to know the effect of a variable on decisions or in program evaluations. 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.[1]

Furthermore, it is well-known that permanent unobserved heterogeneity may bias estimates and lead to misleading conclusions about the effect of a variable if we do not control for it. This is particularly true in dynamic models: the state dependence coefficients are seriously biased getting significative coefficients even when there is no state dependence and persistence is only due to permanent heterogeneity. A great bias is also found in the coefficients that describe the effect of observed characteristics when they are correlated with the unobserved ones. In econometric literature, there are two ways

[1] See Hyslop (1999) and Chay & Hyslop (2000) for examples of studies of dynamic structure and persistence on female labor force and welfare participation.

of treating unobserved heterogeneity: random effects and fixed effects. Random effects are used when some knowledge about the distribution of the unobserved heterogeneity given the observables is assumed. Fixed effects are used when that distribution is left completely unrestricted. In the latter case, the effect is treated as one different parameter for each individual.

The problem with random effects comes from the difficulty in establishing a distribution of the heterogeneity, particularly when it may be related to other observed variables. An additional problem is the so-called "initial conditions problem": the process is not observed from the beginning; thus, we have to assume something about the initial conditions and its relations with the other variables. Even if you condition in the first sample observation, you need to specify the distribution of the random effects given the observables and the initial conditions. This is because even if the random effects are independent of the observables, their distribution conditional on the first sample observation depends on the observables. A misspecification on sample initial conditions or random effects distribution will lead to inconsistent estimates. So we would obtain more robust estimates (robust to misspecifications) if we leave unrestricted the unobserved heterogeneity as in the fixed effects approach.

There is an extensive research on how to estimate linear panel data models with fixed effects, but there are no general solutions for non-linear cases. This is an open area of research. For instance, probit models with explanatory variables including both exogenous variables and the lagged dependent variable do not have a $\sqrt{N}$ consistent estimator. Monte Carlo experiments have shown that the traditional maximum likelihood estimator exhibits considerable bias in finite samples when $T$ is not large.[2]

The estimation of non-linear models with fixed effects by maximum likelihood suffers the so-called incidental parameters problem. Here, the fixed effects are the incidental parameters. Cox and Reid (1987) considered the general problem of doing inference for a parameter of interest in the absence of knowledge about nuisance parameters.[3] Their solution is based on getting a parametrization such that the nuisance parameters are

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

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

information orthogonal to the other parameters, as to limit the influence of the nuisance parameters. Then, they develop a modification that reduce the order of the bias of the maximum likelihood estimator (MLE), without increasing its asymptotic variance. Their general framework has been used for static binary choice panel data models with fixed effects in Arellano (2001). I apply Cox and Reid idea to dynamic panel data discrete choice models, studying the asymptotic properties for different $N$ and $T$ plans, and I evaluate its performance in finite samples. Although this modified MLE 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, because it reduces the order of the bias of the MLE. The method gives a general framework for the estimation of non-linear models with fixed effects, compared with the restrictive assumptions needed for using other estimators. For instance, it can be used regardless of the distribution of the errors assumed.

Lancaster (1997) 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) that is the base of this paper.

The rest of paper is organized as follows. Next section presents the kind of models whose estimation is studied in this paper, the problem that I try to solve and a brief comment on some approaches taken in the literature for specific cases. Section 2 also discusses the nature of parameters of interest (policy parameters), i.e. useful measures of the effects with the kind of models considered here. That is not a trivial question, since, in contrast with the linear case, each one of the parameters defined in the model does not capture on his own the effect of the explanatory variables, the effects are different for each individual, they depend on the fixed effects and there are more than one measure that should be considered. Section 3 presents the alternative approach that we try to address and its asymptotic properties. Section 4 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 Section 5 I estimate a female labor force participation model as an empirical illustration. The last section concludes.

# 2 The Model and Parameters of Interest

## 2.1 The Model

Let us consider the following panel data model:

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

where $1\{c\}$ takes value one if condition $c$ is satisfied and zero otherwise. $\{\eta_i\}_{i=1,...,N}$ describe permanent unobserved heterogeneity among individuals and $v_{it}$ reflects unobserved random variables and shocks that individuals receive every period. As previously explained, I do not want to make any assumption nor restriction on the distribution of $\eta_i$, so I take a fixed effect approach and, therefore, treat $\{\eta_i\}_{i=1,...,N}$ as parameters to be estimated. In all the paper, for any variable (or set of variables) $z$, $z_{it}$ denotes observation at period $t$ for individual $i$, $z_i = \{z_{i0}, ..., z_{iT-1}\}$, 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 $i$ ($z_i^t = \{z_{i0}, ..., z_{it}\}$).

Examples of the use of this kind of models can be found in Chay and Hyslop (2000) and Hyslop (1999). The former estimates different specifications of the model with alternative assumptions about the unobserved individual heterogeneity and initial conditions, using female welfare and labor force participation data. Hyslop (1999) studies the dynamic structure of labor force participation of U.S. married women using both linear probability and probit specifications. However, he uses a random effects approach for the estimation of the dynamic probit model.

Assuming that $v_{it}$ follows certain 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: $\beta$, $\eta_1$, ..., $\eta_N$. By doing so, it gives rise to the incidental parameters problem, first considered by Neyman and Scott (1948). It implies that for any panel

of finite length, estimators of individual fixed effects are necessarily inconsistent. The intuition of incidental parameters problem is clear in this case. Only new observations for individual $i$ give new information about $\eta_i$ and more individuals, i.e. increments in $N$, do not help with the estimation of $\eta_i$ and add more parameters to be estimated. Therefore, the maximum likelihood estimator (MLE) of $\eta_i$ is only consistent when $T \to \infty$. In the maximum likelihood estimator of model (1), the inconsistency of the estimations of $\eta_i$ is transmitted to the estimator of the other parameters. The log-likelihood conditioning on the first observation is

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

where it is assumed that $-v_{it}$ are independently distributed with cdf $F$ and $\gamma = (\alpha, \beta')'$. Deriving with respect to $\gamma$, $\eta_1$, ..., $\eta_N$, we get the first order conditions $d_{\eta_i}(\gamma, \eta_i) \equiv \frac{\partial l_i(\gamma, \eta_i)}{\partial \eta_i}$ and $d_{\gamma i}(\gamma, \eta_i) \equiv \frac{\partial l_i(\gamma, \eta_i)}{\partial \gamma}$. Note that $l(\gamma, \eta_1, ..., \eta_N)$ is defined for $i$ observations such that $\sum_{t=1}^{T-1} y_{it}$ is not zero or $T - 1$. MLE of $\eta_i$ for given $\gamma$, $\widehat{\eta}_i(\gamma)$, solves $d_{\eta_i}(\gamma, \eta_i) = 0$. The MLE of $\gamma$ 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:

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

This first order condition or estimating equation of $\gamma$ depends on $\widehat{\eta}_i$, and evaluated at the true value, $\gamma_0$, 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, $\eta_{i0}$, in such situation.

This problem can be overcome if the estimator of $\gamma$ can be derived so that it does not depend on the incidental parameters. This is what is done in the linear case, where estimators of $\gamma$ based on regression in first differences or deviations from group means (whiting groups estimator) are consistent for large $N$ and fixed $T$ because they do not depend on $\eta_i$. However, first differentiating or deviations from means does not work in the case of non-linear models: $\Delta y_{it} = 1\{\alpha y_{it-1} + x'_{it}\beta + \eta_i + v_{it} \geq 0\} - 1\{\alpha y_{it-2} + x'_{it-1}\beta + \eta_i + v_{it-1} \geq 0\}$ still depends on $\eta_i$.

A way of getting rid of the incidental parameters and therefore a consistent estimator for fixed $T$ in binary choice models is by conditioning on a sufficient statistic of $\eta_i$, using

conditional MLE. However it is not possible to find a sufficient statistic for many of the non-linear models used in econometrics. In particular, logistic assumption is needed and, even with that assumption, model (1) does not have a sufficient statistic. Manski's maximum score estimator is not restricted to a specific distributional assumptions, but it imposes strict exogeneity on all explanatory variables, excluding dynamic models.[4]

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 logistic assumption, $\varepsilon_{it}$ be serially independent over time, $x_{i2}$ equal $x_{i3}$ or $(x_{i2} - x_{i3})$ be continuously distributed with support in a neighborhood of 0 and $(x_{i1} - x_{i2})$ 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 logistic assumption is relaxed, Manski's insight is used and now, in addition to the former limitations, the parameters are identified only up to scale and the objective function is not differentiable, which makes the maximization more difficult. Honoré and Kyriazidou do not derive the asymptotic distribution in this latter case, but they expect the limiting distribution to be non-normal like in the maximum score estimator, and the rate of convergence to be slower than $N^{-1/3}$. So $\alpha$ and $\beta$ in a dynamic probit of the form of model (1) do not have a good estimator.

## 2.2 Policy Parameters of Interest

What we want to measure with the econometric models is the effect of $x$ on $y$. In the linear model that effect is $\beta$ and the expected effect of a change in $x$ over $y$ is the same for all individuals, so the average effect on the population is equal to the expected effect for an individual, $\beta$. In binary choice models $\beta$ is of interest since some economic hypothesis impose testable restrictions on its sign or magnitude. Also, the $\beta$ coefficients give the relative impact of the explanatory variables on the probabilities of $(y_{it} = 1)$. Nevertheless, even though the micropanel literature has emphasized the fixed

[4] See Arellano & Honore (2001) and Arellano (2001) for surveys on fixed $T$ solutions for discrete choice models.

$T$ consistent estimation of $\beta$, 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 and individual $i$ (or the effect of a change in $x_1$ over the probability of $(y_{it} = 1)$) is:

$$\frac{\partial}{\partial x_1} E\left[y_{it}|x, \eta_i\right] = \frac{\partial}{\partial x_1} F(\beta_1 x_1 + \beta_2 x_2 + \eta_i) = \beta_1 f(\beta_1 x_1 + \beta_2 x_2 + \eta_i) \tag{2}$$

when $x$ is a continuous variable and

$$E\left[y_{it}|x_{1b}, x_2, \eta_i\right] - E\left[y_{it}|x_{1a}, x_2, \eta_i\right] = F(\beta_1 x_{1b} + \beta_2 x_2 + \eta_i) - F(\beta_1 x_{1a} + \beta_2 x_2 + \eta_i) \tag{3}$$

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 equation (2) $f$ 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 $\eta_i$. So, the effect differs among individuals due to their unobserved heterogeneity and the values of the $x$ and $y$ that each one has. Also, the effect for an individual is different in each period if any of the explanatory variables has changed during that time.

Usually, the mean effect for all individuals is what people want to calculate. But more than one mean can be considered. A measure found in literature is the effect of an increment in $x_1$ over the probability of $y = 1$, for an individual with the average characteristics:

$$F(\beta_1(E(x_{1it}) + 1) + \beta_2 E(x_{2it}) + E(\eta_i)) - F(\beta_1 E(x_{1it}) + \beta_2 E(x_{2it}) + E(\eta_i)) \tag{4}$$

As noted by Chamberlain (1984), this effect may not be relevant for most of the population since an individual with the average characteristics may not represent any individual in the population.

The expected effect over the probability of $y = 1$ of going from $x_{1a}$ to $x_{1b}$ is $E_{(\eta,x_2)|x_1}$(equation (3) ) =

$$\int_{\eta_i, x_{2i}} \left[F(\beta_1 x_{1b} + \beta_2 x_2 + \eta_i) - F(\beta_1 x_{1a} + \beta_2 x_2 + \eta_i)\right] dG_{(\eta,x_2)|x_1}(\eta_i, x_2|x_{1a}) \tag{5}$$

where $G$ is the distribution function. This is the parameter of interest estimated in Altonji and Matzkin (2001). They present it for the continuous case as the expected

value of the partial derivative of the probability of $y = 1$ with respect to $x$, holding the distribution of the unobservables constant, i.e. $E_{(\eta,x_2)|x_1}$(equation (2)). This is done so, because we want to isolate the effect of the explanatory variable $x_1$ from its correlation with the unobserved heterogeneity.

An alternative mean, proposed in the literature is taking the average over the marginal distribution of $\eta$,

$$\int_{x_2}\int_{\eta}\left[F(\beta_1(x_{1b})+\beta_2 x_2+\eta_i)-F(\beta_1 x_{1a}+\beta_2 x_2+\eta_i)\right]dG_\eta(\eta_i)dG_{x_2|x_1}(x_2|x_{1a}). \quad (6)$$

This measure corresponds with the derivative of the Average Structural Function (ASF) defined in Blundell and Powell (2000). They present it as easier to identify since the conditional distribution of $\eta$ is not needed.[5] This measure abstracts from the correlation between the $x$ variables and $\eta$.

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

$$\int_{\eta_i,x_2}\left[F(\beta_1 x_{1b}+\beta_2 x_{2it}+\eta_i)-F(\beta_1 x_{1a}+\beta_2 x_{2it}+\eta_i)\right]dG(\eta_i,x_{2it}) \quad (7)$$

where the joint distribution of $\eta$ and $x_2$ is used. This measure takes into account the correlation between $x_2$ and $\eta$, but if we want to measure the effect at certain level of $x_1 = x_{1a}$ the distribution of the unobserved types is not the same as for other level. In this latter case the effect is given by equation (5).

Equations (7) and (5) are the answer to different questions. For example, if we want to measure the effect of having a third child ($x_{1a} = 2$, $x_{1b} = 3$) over the probability of female labor force participation ($y = 1$) for those that already have two children, we should take into account that there is a correlation between the number of children and the unobserved preferences ($\eta_i$), and that some unobserved types are more likely to have two children than others. So the relevant distribution of $\eta$ is the conditional one and we should compute (5). Measure (7) will be the answer to a question like, what would be the expected effect of having a child ($x_{1a} = 0$, $x_{1b} = 1$) for all individuals in

[5]Blundell and Powell (2000) consider models with endogenous regressors. In model (1) the unobservable part has two components: an exogenous shock $v_{it}$ and permanent unobserved heterogeneity $\eta_i$ possibly correlated with the regressors. So the endogeneity in this case comes from $\eta_i$.

the population? If $x_1$ were a treatment indicator such that $x_{1b} = 1$ and $x_{1a} = 0$, (7) would be the Average Treatment Effect and (5) would be the average Treatment on the Untreated.

The previous parameters of interest give an expected effect at each level of $x_1$. If we want to know the average effect for the population of an increment in their level of $x_1$ on a specific period $t$: $E$(equation (3)) =

$$\int_{\eta_i, x_{1it}, x_{2it}} \left[F(\beta_1(x_{1it}+1) + \beta_2 x_{2it} + \eta_i) - F(\beta_1 x_{1it} + \beta_2 x_{2it} + \eta_i)\right] dG_{(\eta,x)}(\eta_i, x_{1it}, x_{2it}) \tag{8}$$

This is a summary measure for the population to evaluate the effect of an increment given the actual level of the variable from which individuals come from. This is the parameter of interest if, for example, we want to know the average effect of an increment in the external income over the probability of working, considering the level of income that each one has. In this example, $x$ is continues, so the effect is given by $E(2) = \int_{\eta, x_{it}} \beta_1 f(x'_{it}\beta + \eta_i) dG_{(\eta,x)}(\eta_i, x_{it})$. Notice that the parameter of interest expressed in equation (8) is equal to $E_{x_1}$(equation (5)) -replacing $x_{1a} = x_{1it}$ and $x_{1b} = x_{1it} + 1$- but it is not equal to $E_{x_1}$(equation (7)).

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

In model (1), there is also a long run effect, since there are dynamic effects through $\alpha y_{it-1}$. This long run effect of a change in $x_{it}$ over the probability of $(y = 1)$ for a specific individual $i$ on period $t$ is

$$\frac{\partial}{\partial x}\left\{\frac{F(\beta x_{it} + \eta_i)}{1 - F(\alpha + \beta x_{it} + \eta_i) + F(\beta x_{it} + \eta_i)}\right\}$$

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.[6] In addition to that, the

[6] This is a decision based on economic motivations. For example, in the context of treatment effects and program evaluation studies, Heckman and Smith (1997) discuss situations in which the distribution of the effect on the population and not the mean of it, is what we need to estimate.

Table 1: Sample counterparts of the population paremeters of interest

| Pop. | Sample Counterpart |
|---|---|
| (4) | $[F(\beta_1(\overline{x}_1+1)+\beta_2\overline{x}_2+\overline{\eta})-F(\beta_1\overline{x}_1+\beta_2\overline{x}_2+\overline{\eta})]$ |
| (5) | $\frac{1}{Na}\sum_{i=1}^{N}[F(\beta_1 x_{1b}+\beta_2 x_{2it}+\eta_i)-F(\beta_1 x_{1a}+\beta_2 x_{2it}+\eta_i)]\,1_{\{x_{1it}=x_{1a}\}}$* |
| (6) | $\frac{1}{Na}\sum_{i=1}^{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\left[F(\beta_1 x_{1b}+\beta_2 x_{2it}+\eta_j)-F(\beta_1 x_{1a}+\beta_2 x_{2it}+\eta_j)\right]\right\}1_{\{x_{1it}=x_{1a}\}}$ |
| (7) | $\frac{1}{N}\sum_{i=1}^{N}[F(\beta_1 x_{1b}+\beta_2 x_{2it}+\eta_i)-F(\beta_1 x_{1a}+\beta_2 x_{2it}+\eta_i)]$ |
| (8) | $\frac{1}{N}\sum_{i=1}^{N}[F(\beta_1(x_{1it}+1)+\beta_2 x_{2it}+\eta_i)-F(\beta_1 x_{1it}+\beta_2 x_{2it}+\eta_i)]$ |

where $\overline{x}=\frac{1}{N}\sum_{i=1}^{N}x_{it}$, $\overline{\eta}=\frac{1}{N}\sum_{i=1}^{N}\eta_i$ and $Na=\sum_{i=1}^{N}1\{x_{1it}=x_{1a}\}$.
* When $x$ is a continuos variable, $1\{x_{it}=x_a\}$ will be sustituted by a kernel density function.

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 $x$ produce 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, as it is captured by the form of function $f$. The group with small effect of a change in $x$ on the probability of choosing ($y_{it}=1$) contains individuals whose characteristics makes them choose an option with high probability. The mean effect may be between those two groups of individuals and it may not be relevant for most of the population. More appropriate measures are those that help us to evaluate the effect over both groups separately and to know the relative importance of them. 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.

For instance, suppose that in model (1), $y_{it}$ indicates whether individual $i$ owns a

car and $x_{it}$ is their level of income at period $t$ and we want to study the effect of $x$ over $y$. People with very high level of income is going to have a car and a small change in income is scarcely going to affect their decision - i.e. (2) is very small-. People with very low level of income is not going to have it and a change in income is not probably going to change that, since they are very far below from the threshold level. In this last case, (2) is going to be very small too. If we want to know the effect of changes in the level of income over $y_{it}$ we may prefer to focus on those that, due to their level of income, are near to the threshold (not very large and not very small), and therefore are significantly affected by a small change in $x$. In such situation we need to look at the distribution of the effects and not only to the mean.

# 3 Modifying 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 commented in the previous section, only partial solutions with restrictive assumptions have been found for fixed $T$, and identification problems arise in that context when those assumptions are relaxed (see Chamberlain, 1992 and Arellano, 2001). Also, as shown by Alvarez and Arellano (1998) for linear autoregresive 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 do not need to restrict ourselves 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 be block diagonal between the parameters of interest and the nuisance parameters. 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 a fixed effects formulation in Arellano (2001).

Reparametrization is made from the original parameters $(\gamma, \eta_i)$ to $(\gamma, \lambda_i)$ so that[7]

$$E\left(\frac{\partial^2 l_i(\gamma_0, \lambda_{0i})}{\partial\gamma\partial\lambda_i}\right) = 0 \tag{9}$$

If we could get orthogonality, that is $\frac{\partial^2 l_i(\gamma,\lambda_i)}{\partial\gamma\partial\lambda_i}$ to equal zero for all $i$, there would not be incidental parameters problem, since the estimation of $\gamma = (\alpha, \beta')'$ would not depend on the estimation of $\lambda_i$. But this strong form of orthogonality cannot be generally achieved and, in particular, does not hold for models like (1). Information orthogonality ensures the estimations of $\lambda_i(\gamma)$ changes slowly with $\gamma$, but the MLE of $\gamma$ is the same, since it is invariant to reparametrization. So, $\widehat{\gamma}_{MLE}$ has the same bias of order $O(T^{-1})$ with both parametrizations. This is why they modified the concentrated likelihood. In our case, the modified concentrated log likelihood of Cox and Reid (1987) is:

$$L_M(\alpha, \beta) = \sum_{i=1}^{N} l_{Mi}(\alpha, \beta) = \sum_{i=1}^{N} l_i^*\left(\gamma, \widehat{\lambda}_i(\gamma)\right) - \frac{1}{2}\log\left[-d_{\lambda\lambda}^*\left(\gamma, \widehat{\lambda}_i(\gamma)\right)\right] \tag{10}$$

where $l_i^*$ is the concentrated log-likelihood for all the observations of individual $i$, reparametrized from $(\alpha, \beta, \eta_i)$ to $(\alpha, \beta, \lambda_i)$, i.e. $l_i^*(\gamma, \lambda_i) = l_i(\gamma, \eta_i(\gamma, \lambda_i))$. The modification term is $d_{\lambda\lambda}^*\left(\alpha, \beta, \widehat{\lambda}_i(\alpha, \beta)\right) = \partial^2 l_i^*/\partial\lambda_i^2$. They make a first order adjustment that tries to correct the asymptotic bias that comes from the estimation of the fixed effect.[8] The

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

[8] Concentrated likelihood and modified concentrated likelihood in this paper correspond to the profile likelihood and modified profile likelihood in Cox & Reid (1987). In that paper they justify the modification as a way to approximate the conditional likelihood (conditional on incidental parameters or on a sufficient statistic of them), using $\widehat{\lambda}_i(\gamma)$ as the conditioning statistic. They interpret the modifications term as penalizing values of $\alpha$ and $\beta$ for which the information about the fixed effects is relatively large.

first order condition or estimating equation of the modified likelihood is more nearly unbiased than that using the concentrated likelihood. Ferguson, Reid and Cox (1991) proved that result in general, and Arellano (2001) made the calculation for the static panel data case showing that the bias in the expected modified score is of order $O(T^{-1})$ as opposed to $O(1)$ in the expected concentrated score without modification.[9] For estimators this means that the order of the bias is reduced from $O(T^{-1})$ of the MLE to $O(T^{-2})$ of the modified maximum likelihood estimator (MMLE). For models like (1) both MLE and MMLE are inconsistent for fixed $T$, since the modification corrects first order bias but not biases of smaller order.[10] Arellano (2001) derived the asymptotic properties of MMLE when $T/N$ tends to a constant -i.e. $T$ grows at the same rate as $N$- and compared it with MLE. His results can be summarized as follows:

Consistency: Both, the ML and the MML estimators of $\gamma = (\alpha, \beta')'$ are consistent as $T \to \infty$ regardless of $N$.

Asymptotic normality: When $\frac{N}{T} \to c$, $0 < c < \infty$ ($N$ and $T$ grows at the same rate):

$$\left(H_{NT}^{'} V_{NT}^{-1} H_{NT}\right)^{1/2} \sqrt{NT} \left(\widehat{\gamma}_{MLE} - \gamma_0 + \frac{1}{T} H_{NT}^{-1} b_N\right) \xrightarrow{d} N(0, I) \tag{11}$$

$$\left(H_{NT}^{*'} V_{NT}^{-1} H_{NT}^{*}\right)^{1/2} \sqrt{NT} \left(\widehat{\gamma}_{MMLE} - \gamma_0\right) \xrightarrow{d} N(0, I) \tag{12}$$

where

$$V_{NT} = \frac{1}{NT} \sum_{i=1}^{N} \frac{\partial l_i^*(\gamma_0, \lambda_{i0})}{\partial \gamma} \frac{\partial l_i^*(\gamma_0, \lambda_{i0})}{\partial \gamma}', \tag{13}$$

$$H_{NT} = \frac{1}{NT} \sum_{i=1}^{N} \frac{\partial^2 l_i^*(\gamma_0, \widehat{\lambda}_i(\gamma_0))}{\partial \gamma \partial \gamma'} \tag{14}$$

and

$$H_{NT}^{*} = \frac{1}{NT} \sum_{i=1}^{N} \frac{\partial^2 l_{Mi}(\gamma_0)}{\partial \gamma \partial \gamma'} \tag{15}$$

The MLE has a bias of order $O(T^{-1})$ in its asymptotic distribution. The bias term

[9] In Appendix A, I've included that calculations for our case.

[10] For models whose bias of smaller order is zero, the MMLE is a consistent estimator when $N \to \infty$ and $T$ is fixed, because the modification is correcting all the bias. An example is the model: $y_{it} = \alpha y_{it-1} + \eta_i + v_{it}$, studied in Lancaster (1997).

disappears in the case of the MMLE because $b_N = \frac{1}{N}\sum_{i=1}^{N} \frac{E\left[d^*_{\gamma\lambda\lambda i}(\gamma,\lambda_{i0})\right]}{2E\left[d^*_{\lambda\lambda i}(\gamma,\lambda_{i0})\right]}$ is the term that the modification corrected.

As a matter of fact, the MMLE, has no bias in its asymptotic distribution not only when $N$ and $T$ grows at the same rate as proved in Arellano (2001), but also when $N$ grows faster than $T$ $\left(\frac{T}{N} \to 0\right)$, provided $T$ grows faster than $N^{1/3}$, i.e. $\frac{T}{\sqrt[3]{N}} \to \infty$. This result is straight forward from the order of the bias of $\widehat{\gamma}_{MMLE}$, which, as stated before, is $O(T^{-2})$.[11]

Arellano (2001) writes modified log-likelihood in equation (10) in terms of the original parameters as:

$$L_M(\gamma) = \sum_{i=1}^{N} l_{Mi}(\gamma) = \sum_{i=1}^{N} l_i\left(\gamma, \widehat{\eta}_i(\gamma)\right) - \frac{1}{2}\log\left[-d_{\eta\eta i}\left(\gamma, \widehat{\eta}_i(\gamma)\right)\right] + \log\left(\left.\frac{\partial\lambda_i}{\partial\eta_i}\right|_{\eta_i=\widehat{\eta}_i(\gamma)}\right) \tag{16}$$

where $l_i\left(\gamma, \widehat{\eta}_i(\gamma)\right)$ is the concentrated log-likelihood of individual $i$'s observations and $d_{\eta\eta i}\left(\gamma, \widehat{\eta}_i(\gamma)\right) = \left.\frac{\partial^2 l_i}{\partial\eta_i^2}\right|_{\eta_i=\widehat{\eta}_i(\gamma)}$. The two terms that modified $l_i\left(\gamma, \widehat{\eta}_i(\gamma)\right)$ comes from the modification in equation (10): $d^*_{\lambda\lambda} = \frac{\partial^2 l_i^*}{\partial\lambda_i^2} = \frac{\partial^2 l_i}{\partial\eta_i^2}\left(\frac{\partial\eta_i}{\partial\lambda_i}\right)^2 + \frac{\partial l_i}{\partial\eta_i}\frac{\partial^2\eta_i}{\partial\lambda_i^2} = \frac{\partial^2 l_i}{\partial\eta_i^2}\left(\frac{\partial\eta_i}{\partial\lambda_i}\right)^2$. In equation (16) is clearly seen that explicit reparametrization from $(\gamma, \eta_i)$ to $(\gamma, \lambda_i)$ is not needed. As in the integrated likelihood estimator studied by Woutersen (2001), only the Jacobian term $\frac{\partial\lambda_i}{\partial\eta_i}$ is required. But, since the estimating equation is the score equation, the Jacobian is not even needed for estimation. We need the derivative of the Jacobian with respect to $\gamma$, which is simpler to obtain because it is given by the partial differential equations implied by the orthogonalization condition (9). The orthogonalization $\eta(\gamma, \lambda_i)$ must satisfy the partial differential equations

$$\frac{\partial\eta_i}{\partial\gamma} = -\frac{1}{E\left[d_{\eta\eta i}(\gamma,\eta_i)\right]} E\left[d_{\gamma\eta i}(\gamma,\eta_i)\right]$$

Then, for $\alpha$, the orthogonalization implies that:

$$\frac{\partial}{\partial\alpha}\log\left(\frac{\partial\lambda_i}{\partial\eta_i}\right) = \frac{\partial}{\partial\eta_i}\left(\frac{E\left[d_{\alpha\eta i}(\gamma,\eta_i)\right]}{E\left[d_{\eta\eta i}(\gamma,\eta_i)\right]}\right) \tag{17}$$

Given (16) and (17), the modified maximum likelihood estimator of $\alpha$, $\widehat{\alpha}_{MMLE}$, is

[11] Woutersen (2001) proves that result for the integrated likelihood estimator.

the value of $\alpha$ that solves the following score equation:

$$\sum_{i=1}^{N} d_{\alpha M i}(\gamma) = \sum_{i=1}^{N} d_{\alpha\,\mathrm{Ci}}(\gamma) - \sum_{i=1}^{N} \frac{1}{2} \frac{\frac{\partial}{\partial \alpha} d_{\eta\eta i}(\gamma, \widehat{\eta}_i(\gamma))}{d_{\eta\eta i}(\gamma, \widehat{\eta}_i(\gamma))} + \sum_{i=1}^{N} \frac{\frac{\partial}{\partial \eta_i} \left( E\left[ d_{\gamma\eta i}(\gamma, \eta_i) \right] \right) \Big|_{\eta_i = \widehat{\eta}_i(\gamma)}}{E\left[ d_{\eta\eta i}(\gamma, \widehat{\eta}_i(\gamma)) \right]}$$
$$- \sum_{i=1}^{N} \frac{E\left[ d_{\gamma\eta i}(\gamma, \widehat{\eta}_i(\gamma)) \right]}{E\left[ d_{\eta\eta i}(\gamma, \widehat{\eta}_i(\gamma)) \right]} \frac{\frac{\partial}{\partial \eta_i} \left( E\left[ d_{\eta\eta i}(\gamma, \eta_i) \right] \right) \Big|_{\eta_i = \widehat{\eta}_i(\gamma)}}{E\left[ d_{\eta\eta i}(\gamma, \widehat{\eta}_i(\gamma)) \right]} = 0 \tag{18}$$

where $d_{\alpha\,\mathrm{Ci}}(\gamma) = \frac{\partial l_i(\gamma, \eta_i)}{\partial \gamma} \Big|_{\eta_i = \widehat{\eta}_i(\gamma)}$ is the standard score from the concentrated likelihood, $d_{\alpha\eta i}(\gamma, \eta_i) = \frac{\partial^2 l_i}{\partial \alpha \partial \eta_i}$, $d_{\eta\eta i}(\gamma, \eta_i) = \frac{\partial^2 l_i}{\partial \eta_i^2}$ and $\widehat{\eta}_i(\gamma)$ is gotten from the first order condition of $\eta_i$, as it is in the concentrated maximum likelihood. The score equation for $\beta$ is the same as (18), just replacing $\alpha$ by $\beta$.

In appendix B, I show all the calculations needed in order to compute $d_{\alpha M i}(\alpha, \beta)$ for a particular model. In Appendix C, I address the problem of how to optimize a concentrated likelihood given that, as it happens in the kind of models I interested in, $\widehat{\eta}_i(\alpha, \beta)$ cannot be analytically calculated.

The modified maximum likelihood could achieve the goal stated at the beginning of this section and the finite sample bias may be negligible for moderate $T$ even that, in general, it is only consistent when $T \to \infty$, because it reduces the order of the bias. Also, a main advantage of this way of estimating over other methods for panel data binary choice models is its generality. Estimators like the ones mentioned in Section 2 are too specific and require very restrictive assumptions. However, MMLE can be applied to different models with different assumptions. For example this method allows for time dummy variables whereas Honoré and Kyriazidou's does not. MMLE could also be applied to multinomial choice models and other non-linear model, not only to binary choice.

In addition to these properties, MMLE is a convenient estimator to compute the policy parameters of interest because the fixed effects are estimated as part of the estimation process whereas in the fixed $T$ consistent estimation you get read of them, and the effect of interest depends on the fixed effects, as discussed in Section 2. Furthermore, asymptotic properties in both $N$ and $T$, have to be considered here since the estimates of the parameters of interest are only consistent when $T \to \infty$.

# 4 Monte Carlo Evidence

In this section Monte Carlo simulations are used to evaluate the performance of MMLE in different sample sizes to see if this new estimator and its asymptotic distribution when both $N$ and $T$ goes to infinity and $T \propto N^{\alpha}$ with $\alpha > \frac{1}{3}$, is a good approximation and have a good properties in finite samples.

The first model I consider is a dynamic logit:

$$y_{it} = 1\{\alpha y_{it-1} + \beta x_{it} + \eta_i + v_{it} \geq 0\} \ (t = 0, ..., T-1; i = 1, ..., N) \tag{19}$$

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

$$\Pr(y_{it} = 1 | \eta_i, y_i^{t-1}, x_i) = F(\alpha y_{it-1} + \beta x_{it} + \eta_i) = F_{it} \tag{20}$$

and $F$ is the logistic cdf. I consider 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 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 with the ones they report. One practical disadvantage of their estimator that deserves to be pointed out 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.[12] 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 make a thousand replications, $\beta_0 = 1$, $x_{it}$ is i.i.d. $N(0, \pi^2/3)$, $\varepsilon_{it}$ 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 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

[12] 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 2: Logit design with different N and T values

| | | | T=4 | | | T=8 | | |
|---|---|---|---|---|---|---|---|---|
| | | N | 250 | 500 | 1000 | 250 | 500 | 1000 |
| | $\widehat{\beta}$ | Bias | 0.759 | 0.768 | 0.759 | 0.248 | 0.253 | 0.254 |
| MLE | $\widehat{\beta}$ | 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 |
| | $\widehat{\beta}$ | Bias | -0.054 | -0.053 | -0.057 | 0.012 | 0.015 | 0.015 |
| MMLE | $\widehat{\beta}$ | MAE | 0.068 | 0.051 | 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 |
| | $\widehat{\beta}$ | Bias | 0.076 | 0.044 | 0.038 | 0.014 | 0.007 | 0.009 |
| H and K | $\widehat{\beta}$ | 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: $y_{it} = 1\left(\alpha y_{it-1} + \beta x_{it} + \eta_i + \varepsilon_{it} \geq 0\right)$; $\beta_0 = 1$; $\alpha_0 = 0.5$; $\eta_i = \frac{1}{4}\sum_{t=1}^{4} x_{it}$; $x_{it} \sim N\left(0, \frac{\pi^2}{3}\right)$; $\varepsilon_{it}$ ∼logistic; 1000 Monte Carlo simulations. Median Bias and Median Absolute Error (MAE) are reported.

$T$ 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 present the median bias an median absolute error (MAE) because are robust to outliers and to be able to compare with results presented in Honoré and Kyriazidou (2000).

Table 2 presents parameters estimation for a value of $\alpha_0$ equal to 0.5. For $T = 4$, though the bias is greatly reduced compared with MLE, the median bias and MAE of $\widehat{\alpha}_{mmle}$ is far from the results got 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 comparing both with the smallest $T$ size we could have for estimating this kind

Table 3: 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 4: 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 |

of models.[13] However for a $T$ as small as 8, the MMLE has a median absolute error 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 for two different values of $\alpha$ because the larger the $\alpha$ the greater the serial correlation of $y_{it}$ and I expect that the estimator performs worse, as it happens with the HK. Results are presented in Table 3. Again, estimates are greatly improved compared with MLE as they were for smaller values of $\alpha$.

In order to asses the merits of the modification reducing the order of the bias with respect to $T$, I present in Table 4 the results for 16 periods. The MLE of $\alpha_0$ has still an important bias, however the MMLE is now clearly the best one of the three estimators.

[13] Take into account that we condition on the first observation to avoid the initial conditions problem.

Table 5: Probit design with different N, T and $\alpha_0$ values

| | | | MLE | | | | MMLE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Results for $\widehat{\beta}$ | | Results for $\widehat{\alpha}$ | | Results for $\widehat{\beta}$ | | Results for $\widehat{\alpha}$ | |
| $\alpha_0$ | $T$ | $N$ | 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 |

$y_{it} = 1\left(\alpha y_{it-1} + \beta x_{it} + \eta_i + \varepsilon_{it} \geq 0\right)$; $\beta_0 = 1$; $\eta_i = \frac{1}{4}\sum_{t=1}^{4} x_{it}$; $x_{it} \sim N\left(0, \frac{\pi^2}{3}\right)$; $\varepsilon_{it} \sim N\left(0, \frac{\pi^2}{3}\right)$; 1000 Monte Carlo simulations.

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 the same model with a time dummy variable for instance, can not be estimated using HK, whereas it can by modified maximum likelihood. Also, if we want to estimate a probit instead of a logit, HK does not identified separably $\alpha$

and $\beta$. Nevertheless, MMLE works in the same way and keeps its theoretical properties regardless the distribution of $v_{it}$, to the extend that the maximum likelihood estimator is a general method of estimation for different distributional assumptions. As explained in appendix B, the estimator is expressed in terms of a general distribution and density functions, such that we only have to substitute for the appropriate functions. Let us see how it works in finite sample for the same model we have already estimated, model (19), but using a probit assumption. That is

$$\Pr(y_{it} = 1|\eta_i, y_i^{t-1}, x_i) = \Phi(\alpha y_{it-1} + \beta x_{it} + \eta_i) = \Phi_{it} \tag{21}$$

where $\Phi$ is the normal cdf.

Table 5 presents simulations results for a probit with different values of $N$, $T$ and $\alpha_0$. The conclusions are the same as in the logit case and, in terms of median absolute error, they perform similarly. In this Table 5 I include the situation with ten periods. It can be seen 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 sixteen 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 $(-H_{NT}^*)^{-1}$- evaluated at the estimated values. Looking at the asymptotic distribution in equations (11) and (12) 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 6 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 due to 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 $(-H_{NT}^*)^{-1}$ is the easiest choice since it is calculated as part of the

Table 6: Estimates of the variance

| | Mean | RMSE | CI, 95% |
|---|---|---|---|
| | | $\widehat{Var(\widehat{\alpha})}$ | |
| Value in 1000 simulations | 0.011299 | | |
| $(H^{*\prime}V^{*-1}H^{*\prime})^{-1}$ | 0.0107554 | 9.726*$10^{-4}$ | 85.5% |
| $(H^{*\prime}V^{-1}H^{*\prime})^{-1}$ | 0.0117856 | 9.747*$10^{-4}$ | 87.7% |
| $(H^{\prime}V^{-1}H^{\prime})^{-1}$ | 0.0117337 | 9.407*$10^{-4}$ | 87.2% |
| $-H^{-1}$ | 0.0112123 | 4.123*$10^{-4}$ | 86.7% |
| $-H^{*-1}$ | 0.0112586 | 4.219*$10^{-4}$ | 88.9% |
| | | $\widehat{Var(\widehat{\beta})}$ | |
| Value in 1000 simulations | 0.0015955 | | |
| $(H^{*\prime}V^{*-1}H^{*\prime})^{-1}$ | 0.0016823 | 1.908*$10^{-4}$ | 94.2% |
| $(H^{*\prime}V^{-1}H^{*\prime})^{-1}$ | 0.0018146 | 2.827*$10^{-4}$ | 94.8% |
| $(H^{\prime}V^{-1}H^{\prime})^{-1}$ | 0.0017809 | 2.464*$10^{-4}$ | 94.4% |
| $-H^{-1}$ | 0.001696 | 1.497*$10^{-4}$ | 94.0% |
| $-H^{*-1}$ | 0.0017146 | 1.712*$10^{-4}$ | 94.2% |

$y_{it} = 1\left(\alpha y_{it-1} + \beta x_{it} + \eta_i + \varepsilon_{it} \geq 0\right)$; $N = 500$; $T = 8$; $\alpha_0 = 0.5$; $\beta_0 = 1$; $\eta_i = \frac{1}{4}\sum_{t=1}^{4} x_{it}$; $x_{it} \sim N\left(0, \frac{\pi^2}{3}\right)$; $\varepsilon_{it}$ ~logistic; 1000 Monte Carlo simulations.
Colum CI 95%: percentage of 95% confidence intervales that contain the true value of the parameter across 1000 simulations. Intervals based on the normal asintotic distribution. Intervals are not centered, so CI = $\widehat{\alpha} \pm 1.96 * \widehat{Var(\widehat{\alpha})}$ and $\widehat{\beta} \pm 1.96 * \widehat{Var(\widehat{\beta})}$. Mean bias $\widehat{\alpha}_{MMLE} = -0.090$; Mean bias $\widehat{\beta}_{MMLE} = 0.015$

optimization process.[14]

**Policy Parameters of Interest.** I present some of the parameters of interest described in Section 2 for a simulated sample, and their estimates by MMLE and MLE, in table 7. They can all be estimated, just replacing $\alpha$, $\beta$ and $\eta_i$ by their estimates. The measure corresponding to equation (6) is different from the one corresponding to

[14] Although we are using a concentrated likelihood, the Hessian has to take into account that the $\eta_i$ are also being estimated. In Appendix C, I explain how I have addressed this problem.

Table 7: Mean Effects of $y_{it-1} = 1$ on the probability of $y_{it} = 1$ computed according to the diffent measures presented in table 1.

| | Value | MMLE | MLE |
|---|---|---|---|
| $y_{it-1} = 0$ | | | |
| joint (5) | 0.199 | 0.164 | 0.059 |
| marginal (6) | 0.162 | 0.133 | 0.046 |
| $y_{it-1} = 1$ | | | |
| joint (5) | 0.198 | 0.157 | 0.056 |
| marginal (6) | 0.160 | 0.132 | 0.046 |
| Average (7) | 0.198 | 0.160 | 0.057 |

$y_{it} = 1\{\alpha y_{it-1} + \beta x_{it} + \eta_i + v_{it} \geq 0\}$. Dynamic logit case. $\alpha_0 = 1, \beta_0 = -1$, $\eta_i = N(0,1)$, $x_{it} = \eta_i + N(0,1)$ , $\varepsilon_{it}$ is i.i.d. logistically distributed. $N = 10000$, $T = 8$. 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.

equation (5) because (6) 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.

According to the discussion at the end of Section 2, I compute the expected effect of a change in $x$ over the probability of $(y_{it} = 1)$ for each individual, given in equation (2) and look at the distribution of that effect. I use the MML estimates of the parameters of model (19), based on a simulated sample of the described logit design with $\alpha_0 = 0.5$, $\beta_0 = 1$, $N = 500$ and eight periods.

In Figure 1 I present smoothed densities of the effects of a change in $x_{iT-1}$ over the probability of $(y_{iT-1} = 1)$ 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, as explained before, those $i$ observations whose sum of $y_{it}$ for the

Table 8: Descriptive statistics of the individual effects of a change in $x_{iT-1}$ over the probability of $(y_{iT-1} = 1)$

| | Full | Movers | MMLE | MLE |
|---|---|---|---|---|
| Mean | 0.1476 | 0.1519 | 0.1454 | 0.1613 |
| Std. Deviation | 0.0797 | 0.0783 | 0.0805 | 0.1052 |
| Skewness | -0.2707 | -0.3358 | -0.2488 | -0.0530 |
| Percentiles | | | | |
| 1% | 0.0050 | 0.0064 | 0.0040 | 0.0017 |
| 10% | 0.0303 | 0.0348 | 0.0280 | 0.0177 |
| 25% | 0.0747 | 0.0818 | 0.0716 | 0.0583 |
| Median | 0.1581 | 0.1647 | 0.1560 | 0.1630 |
| 75% | 0.2246 | 0.2270 | 0.2236 | 0.2670 |
| 90% | 0.2458 | 0.2462 | 0.2454 | 0.2998 |
| 99% | 0.25 | 0.25 | 0.2493 | 0.3069 |

Logit design with $\alpha_0 = 0.5$, $N = 10000$ and $T = 8$. The first column is for the full sample, the second column is for the sample of movers, i.e. the sample actually used on the estimation, with the true parameters' values, the third column uses the modified maximum likelihood estimates of the parameters and the last column uses the ML estimates of the parameters.

last $T - 1$ periods is equal to zero or $T - 1$. We call these individuals stayers, as oppose to movers, because they take the same decision all the sample period. In this experiment the proportion of stayers is around 10%. 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 second mode are those with higher effect and, in this simulation, most of the individuals are in that region. The mean effect is between those two groups and it is only relevant for a very small part of the population.

Figure 2 is the same but with $N = 10000$. Comparing the four densities in both graphs, it can be noticed that the MLE misestimate the second mode and the densities

around that mode significantly, whereas MMLE describes the distribution more accurately. The main difference between the sample of movers and the full sample is that the former underestimates the density of the small effects, because most of the excluded observations are those whose probability of taking value one is always very high or very low -this is why they take value one or zero in all observed periods- and, therefore a variation in $x$ has little impact on their decision.

The same conclusions can be noticed comparing the percentiles and other descriptive statistics of the distributions in Table 8. No great differences are found if we look at the mean or the median, but there are important differences between MLE and the rest of the estimations looking at any other statistic, particularly at the highest percentiles. Although $\widehat{\beta}_{MLE}$ is severely biased, the mean effect based on ML estimations does not have much bias in this particular case because it overestimates the frequency density of the smallest effects and underestimates the frequency of greatest effects, being finally balanced on average.

Figure 2 and Table 8 are based on a simulation of a large sample. In order to evaluate the small samples performance of the estimation of the distribution of the effects, I conduct a Monte Carlo experiment with 1000 replications of the logit model (19), 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 are in Table 9. The improvement using the MML estimates compared with ML estimates at all quantiles and at the estimation of the maximum effect 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. The reason for this situation is the higher proportion of stayers in this experiment compared with the experiment in table 8. In this case that proportion is around 30%, whereas on the graphs presented in this section it was around 10%. The differences are found only on the lowest quantiles because those that are less affected by a change on the explanatory variables, are those with higher probability of

Table 9: Quantiles of the distribution of the individual effects 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 (19), 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.

not changing their decision on the sample period and, therefore, not being used on the estimation of $\alpha$ and $\beta$. 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. In any case, the estimation of the parameters of interest by MML presented in tables 7, 8 and 9, are consistent when $T$ goes to infinity and they clearly improve the finite sample performance of the MLE.

# 5 Empirical Illustration

In this section, I illustrate 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 assumend to be strictly exogenous with respect to $\varepsilon_{it}$ in equation (22). However, children variables are endogenous with respect to $\eta_i$. Moreover, in contrast with random effects approaches, no restrictions are placed on the form of the dependence between effects and children variables.[15]

I use data on 1461 married women corresponding to waves 12-22 of the Panel Study of Income Dynamics (PSID). Sample information is for the ten calendar years 1979-88. 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.[16]

The equation estimated is:

$$y_{it} = 1\left\{\alpha y_{it-1} + x_{it}'\beta + \eta_i + \varepsilon_{it} \geq 0\right\} \ \ (t = 0, ..., T-1; \ i = 1, ..., N) \qquad (22)$$

$y_{it}$ takes value one if wife $i$ participate in period $t$ and zero otherwise. $x_{it} = (\#kids0 - 2_{it-1}, \ \#kids0 - 2_{it}, \ \#kids3 - 5_{it}, \ \#kids6 - 17_{it}, \ \log income_{it}, \ (age_{it}/10), \ (age_{it}/10)^2,$ time dummies), where $\#kidsa - b$ 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. $\varepsilon_{it}$ is assumed to be independent and identically distributed normal variable. So it is a dynamic probit model.

Table 10 shows some descriptive statistics for the explanatory variables. Table 11 contains the distribution of the number of periods that wives in the sample participate, looking at the last nine sample periods. Most of them participate at least one period. Just eight percent never participate. Almost half of the them participate all the last nine periods. We can not use in our estimation those who never participate or participate all the periods. So the sample we use for estimation is restricted to 664 women. We

[15] In Carro (2002) I study the same problem as in this empirical illustration but I consider more general specifications and assumptions, like take into account specifically on the estimation of the model that the number of children variable could be affected by past participation decisions. In this empirical illustration, as in Hyslop (1999), number of children variables are assumed to be strictly exogenous with respect to $\varepsilon_{it}$ in model (22). However, here they are freely correlated with the other unobservable part $\eta_i$ of equation (22).

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

Table 10: Descriptive statistics

| Variable | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|
| $\#kids0-2$ | 0.235 | 0.472 | 0 | 4 |
| $\#kids3-5$ | 0.288 | 0.515 | 0 | 3 |
| $\#kids6-17$ | 1.036 | 1.105 | 0 | 7 |
| *income* | 42093 | 0.447 | 153 | 1340221 |
| *age* in 1980 | 33.3 | 8.84 | 16 | 56 |

Table 11: Distribution of the number of years that wifes participate, looking at the last nine sample periods

| | Number of years worked | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total |
| Freq. | 121 | 62 | 62 | 57 | 71 | 69 | 94 | 110 | 139 | 676 | 1461 |
| Percent | 8.28 | 4.24 | 4.24 | 3.90 | 4.86 | 4.72 | 6.43 | 7.53 | 9.51 | 46.27 | 100 |

look at the last nine year instead to the ten years we have in our sample, because we are conditioning on the first observation to avoid the sample initial conditions problem.

Table 12 presents the results of the estimation of model (22) by MLE and Modified MLE. There are significant differences on the estimated parameters. As expected, the MLE is underestimating the true state dependence effect and overestimating the effect of the other variables. As a result of that, the impact of previous participation on the probability of participating is, in absolute value, 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 probability of participating, which is approximately given by the ratio $\alpha/\beta$, is a half of the value gotten when using Modified MLE.

In the spirit of the discussion about the policy parameters of interest, one may be interested in calculating the effect on the participation decisions of having one more child

Table 12: Estimates of model (22)

| Parameter | ML Estimates | MML Estimates |
|---|---|---|
| $\alpha$ | 0.755 | 1.082 |
| | (0.043) | (0.042) |
| # Children 0-2$_{t-1}$ | -0.039 | 0.004 |
| | (0.054) | (0.049) |
| # Children 0-2 | -0.534 | -0.400 |
| | (0.064) | (0.058) |
| # Children 3-5 | -0.281 | -0.182 |
| | (0.055) | (0.050) |
| # Children 6-17 | -0.075 | -0.036 |
| | (0.043) | (0.039) |
| Log(income) | -0.252 | -0.208 |
| | (0.055) | (0.051) |
| Age/10 | 2.329 | 1.780 |
| | (0.627) | (0.573) |
| Age$^2$/100 | -0.244 | -0.183 |
| | (0.042) | (0.047) |

Standard errors are in parentheses. Time dummies are also being estimated.

on a particular period, which is of the form of equation (3). In Figure 3 I present the distribution of that effect on the sample of movers, using both the ML and the Modified ML estimates of the parameters. As it happened on the simulated experiment, the MLE is misestimating the distribution of the effect. Also, the MLE overestimates the mean effect, since the mean is -0.156 for the MLE and -0.116 for the MMLE. Figures 4 and 5 present the dynamic effect over twenty periods of having one more child in period one. The first graph is the effect for an individual whose characteristic at period zero are the average characteristic of the sample. This is the kind of effect calculated in

Hyslop (1999). The dynamic effect based on the maximum likelihood estimates is not only overestimating the effect, as clearly shown in the graph, but also misestimating the dynamics, since there is less persistence of the effect over time, compared with the MMLE. It might be more relevant to look at the average effect for the individuals on the sample, instead of the effect for an individual with the average characteristics. Figure 5 presents that average effect. There are differences on magnitude on those graphs. At period three, the effect for an individual with the average characteristics is slightly above 0.20, whereas the mean effect for all individuals is slightly above 0.15, according to the Modified MLE. The conclusion on the comparison of the two estimators in this second graph are the same as in the previous.

# 6 Conclusion

I have applied the modified maximum likelihood estimator (MMLE) to dynamic panel data discrete choice models with fixed effects. This reduces the bias of the estimated parameters from $O(T^{-1})$ to $O(T^{-2})$ (without increasing the asymptotic variance), so that 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, this method allows for time dummy variables whereas Honoré and Kyriazidou's does not. For the probit model, MMLE identifies separably the coefficients $(\alpha, \beta)$ of the explanatory variables and not just $\beta/\alpha$. The method is generally applicable and it has the same asymptotic properties regardless of the distribution of the errors.

In addition, 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 with linear models, that expected effect

is different for each individual and it depends on the fixed effects and on the level of the variables. I have shown that the mean of that effect across all individuals may not be the parameter of interest because for some economic studies we may need to estimate the whole distribution on the population 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 effect for each individual, is estimated as part of the estimation process whereas in the fixed $T$ consistent estimation you get read of them. Also, the asymptotics in both $N$ and $T$ has to be considered because the estimation of that effect is consistent only when $T \to \infty$.

# A Appendix

This appendix shows how the modification on the concentrated log-likelihood (10) is a first order adjustment on the asymptotic bias of the score of the concentrated log-likelihood, so the first order condition is more nearly unbiased and the order of the bias is reduced. The notation follows that in Arellano(2001).

Denote $d^*_{\gamma i} = \frac{\partial l^*_i}{\partial \lambda_i}$, $d^*_{\gamma\lambda i} = \frac{\partial l^*_i}{\partial \gamma \partial \lambda_i}$, and so on. Making an expansion around $\lambda_{i0}$(true value of the fixed effect of individual $i$) of the score of the concentrated log-likelihood evaluated at $\gamma_0$ :

$$d^*_{\gamma i}(\gamma_0, \widehat{\lambda}_i(\gamma_0)) = d^*_{\gamma i}(\gamma_0, \lambda_{i0}) + d^*_{\gamma\lambda_i}(\gamma_0, \lambda_{i0})(\widehat{\lambda}_i(\gamma_0) - \lambda_{i0}) + \frac{1}{2} d^*_{\gamma\lambda\lambda i}(\gamma_0, \lambda_{i0})(\widehat{\lambda}_i(\gamma_0) - \lambda_{i0})^2 + r \quad \text{(A1)}$$

where $r$ is the remainder term. In this equation is clear that the score evaluated at the true value of $\gamma$ ($\gamma_0$), differs from the value of the score that we want to get, i.e. the score evaluated at both $\gamma_0$ and $\lambda_{i0}$ ($d^*_{\gamma i}(\gamma_0, \lambda_{i0})$), as much as $\widehat{\lambda}_i(\gamma_0)$ differs from $\lambda_{i0}$.

The estimator $\widehat{\lambda}_i(\gamma_0)$ solves $d^*_{\lambda_i}(\gamma_0, \widehat{\lambda}_i(\gamma_0)) = 0$. Expanding $d^*_{\lambda_i}(\gamma_0, \widehat{\lambda}_i(\gamma_0))$:

$$0 = d^*_{\gamma i}(\gamma_0, \widehat{\lambda}_i(\gamma_0)) = d^*_{\lambda_i}(\gamma_0, \lambda_{i0}) + d^*_{\lambda\lambda i}(\gamma_0, \lambda_{i0})(\widehat{\lambda}_i(\gamma_0) - \lambda_{i0}) + r$$

From that:

$$\left(\widehat{\lambda}_i(\gamma_0) - \lambda_{i0}\right) = -\frac{d^*_{\lambda_i}(\gamma_0, \lambda_{i0})}{d^*_{\lambda\lambda i}(\gamma_0, \lambda_{i0})} + r' \quad \text{(A2)}$$

Substituting in (A1) the expressions for $\left(\widehat{\lambda}_i(\gamma_0) - \lambda_{i0}\right)$ and $\left(\widehat{\lambda}_i(\gamma_0) - \lambda_{i0}\right)^2$, after some calculations and taking expectations:

$$E\left[d^*_{\gamma i}(\gamma_0, \widehat{\lambda}_i(\gamma_0))|.\right] = E\left[d^*_{\gamma i}(\gamma_0, \lambda_{i0})|.\right] + \frac{E\left[d^*_{\gamma\lambda\lambda i}(\gamma_0, \lambda_{i0})|.\right]}{2E\left[d^*_{\lambda\lambda i}(\gamma_0, \lambda_{i0})|.\right]} + O\left(T^{-1}\right) \quad \text{(A3)}$$

where the conditioning set is $(\lambda_{i0}, x_i, y_{i0})$. $d^*_{\gamma i}(\gamma_0, \widehat{\lambda}_i(\gamma_0))$ is biased and the leading term of the bias, $\frac{E\left[d^*_{\gamma\lambda\lambda i}(\gamma_0, \lambda_{i0})|.\right]}{2E\left[d^*_{\lambda\lambda i}(\gamma_0, \lambda_{i0})|.\right]}$, is $O(1)$.

If we do the same with the score of the modified log-likelihood (10):

$$E\left[d^*_{M\gamma i}(\gamma_0)|.\right] = E\left[d^*_{\gamma i}(\gamma_0, \widehat{\lambda}_i(\gamma_0))|.\right] - \frac{E\left[d^*_{\gamma\lambda\lambda i}(\gamma_0, \lambda_{i0})|.\right]}{2E\left[d^*_{\lambda\lambda i}(\gamma_0, \lambda_{i0})|.\right]} \quad \text{(A4)}$$

where the last term in this expression comes from the derivative of the modification in (10), i.e. the derivative with respect to $\gamma$ of $-\frac{1}{2}\log\left[-d^*_{\lambda\lambda}(\gamma, \lambda_i)\right]$. So, (A4) is equal to (A3) minus the leading term of the bias of the score of the standard concentrated log-likelihood. Therefore, the bias in the modified score is of order $O(T^{-1})$ as opposed to $O(1)$ in the expected score without modification. For the estimators this imply that the MMLE is reducing the order of the bias in the MLE form $O(T^{-1})$ to $O(T^{-2})$.

# B Appendix: Computation of the Modified Score

Let's consider the logit model used in the Monte Carlo experiments and implement the modification on it.

$$y_{it} = 1\{\alpha y_{it-1} + \beta x_{it} + \eta_i + v_{it} \geq 0\} \ (t = 0, ..., T-1; i = 1, ..., N) \tag{B1}$$

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

$$\Pr(y_{it} = 1|\eta_i, y_i^{t-1}, x_i) = F(\alpha y_{it-1} + \beta x_{it} + \eta_i) = F_{it} \tag{B2}$$

$F$ is the logistic cdf.

As explained in the paper, an individual's modified score, in terms of the original parameterization, is of the form:

$$\begin{aligned} d_{\alpha Mi}(\alpha, \beta) &= d_{\alpha\,\mathrm{Ci}}(\alpha, \beta) - \frac{1}{2} \frac{\frac{\partial}{\partial \alpha} d_{\eta\eta i}(\alpha, \beta, \widehat{\eta}_i(\alpha, \beta))}{d_{\eta\eta i}(\alpha, \beta, \widehat{\eta}_i(\alpha, \beta))} + \frac{\frac{\partial}{\partial \eta_i} \left( E\left[ d_{\alpha\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i \right] \right) \Big|_{\eta_i = \widehat{\eta}_i(\alpha, \beta)}}{E\left[ d_{\eta\eta i}(\alpha, \beta, \widehat{\eta}_i(\alpha, \beta)) | y_{i0}, \eta_i, x_i \right]} \\ &\quad - \frac{E\left[ d_{\alpha\eta i}(\alpha, \beta, \widehat{\eta}_i(\alpha, \beta)) | y_{i0}, \eta_i, x_i \right]}{E\left[ d_{\eta\eta i}(\alpha, \beta, \widehat{\eta}_i(\alpha, \beta)) | y_{i0}, \eta_i, x_i \right]} \frac{\frac{\partial}{\partial \eta_i} \left( E\left[ d_{\eta\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i \right] \right) \Big|_{\eta_i = \widehat{\eta}_i(\alpha, \beta)}}{E\left[ d_{\eta\eta i}(\alpha, \beta, \widehat{\eta}_i(\alpha, \beta)) | y_{i0}, \eta_i, x_i \right]} \end{aligned} \tag{B3}$$

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

$$d_{\alpha\,\mathrm{Ci}}(\alpha, \beta) = \frac{\partial l_i(\alpha, \beta, \eta_i(\gamma))}{\partial \alpha} = \sum_{t=1}^{T-1} \left( y_{it-1} + \frac{\partial \widehat{\eta}_i(\alpha, \beta)}{\partial \alpha} \right) \left[ y_{it} - F(\alpha y_{it-1} + \beta x_{it} + \eta_i) \right], \tag{B4}$$

$$d_{\alpha\eta i}(\alpha, \beta, \eta_i) = \frac{\partial^2 l_i}{\partial \alpha \partial \eta_i} = -\sum_{t=1}^{T-1} y_{it-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i), \tag{B5}$$

$$d_{\eta\eta i}(\alpha, \beta, \eta_i) = \frac{\partial^2 l_i}{\partial \eta_i^2} = -\sum_{t=1}^{T-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i), \tag{B6}$$

$f$ is the logistic pdf.

$$E\left[ d_{\alpha\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i \right] = -\sum_{t=1}^{T-1} E\left[ y_{it-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i) | y_{i0}, \eta_i, x_i \right] \tag{B7}$$

$$E\left[ d_{\eta\eta i}(\alpha, \beta, \eta_i) | y_{i0}, \eta_i, x_i \right] = -\sum_{t=1}^{T-1} E\left[ f(\alpha y_{it-1} + \beta x_i + \eta_i) | y_{i0}, \eta_i, x_i \right] \tag{B8}$$

$$E\left[ y_{it-1} f(\alpha y_{it-1} + \beta x_{it} + \eta_i) | y_{i0}, \eta_i . x_i \right] = f(\alpha + \beta x_{it} + \eta_i) \Pr\left( y_{it-1} = 1 | y_{i0}, \eta_i, x_i \right) \tag{B9}$$

$$\begin{aligned} E\left[ f(\alpha y_{it-1} + \beta x_{it} + \eta_i) | y_{i0}, \eta_i, x_i \right] &= f(\alpha + \beta x_{it} + \eta_i) \Pr\left( y_{it-1} = 1 | y_{i0}, \eta_i, x_i \right) + \\ &\quad + f(\beta x_{it} + \eta_i) \left( 1 - \Pr\left( y_{it-1} = 1 | y_{i0}, \eta_i, x_i \right) \right) \\ &= \Pr\left( y_{it-1} = 1 | y_{i0}, \eta_i, x_i \right) \left( f(\alpha + \beta x_{it} + \eta_i) - f(\beta x_{it} + \eta_i) \right) + \\ &\quad + f(\beta x_{it} + \eta_i) \end{aligned} \tag{B10}$$

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

$$\begin{aligned}\Pr(y_{i1}=1|y_{i0},\eta_i,x_i) &= F(\alpha y_{i0}+\beta x_{i1}+\eta_i), \text{ starting point. For } t>1: \quad \text{(B11)}\\ \Pr(y_{it}=1|y_{i0},\eta_i,x_i) &= \Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)\left(F(\alpha+\beta x_{it}+\eta_i)-F(\beta x_{it}+\eta_i)\right)+F(\beta x_{it}+\eta_i)\end{aligned}$$

From (B10), $\frac{\partial}{\partial\eta_i}E\left[f(\alpha y_{it-1}+\beta x_{it}+\eta_i)|y_{i0},\eta_i,x_i\right]=$

$$\begin{aligned}&\frac{\partial}{\partial\eta_i}\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)\left(f(\alpha+\beta x_{it}+\eta_i)-f(\beta x_{it}+\eta_i)\right)\\ &+\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)\left(f'(\alpha+\beta x_{it}+\eta_i)-f'(\beta x_{it}+\eta_i)\right)+f'(\beta x_{it}+\eta_i) \quad \text{(B12)}\end{aligned}$$

From (B9), $\frac{\partial}{\partial\eta_i}E\left[y_{it-1}f(\alpha y_{it-1}+\beta x_{it}+\eta_i)|y_{i0},\eta_i,x_i\right]=$

$$f'(\alpha+\beta x_{it}+\eta_i)\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)+f(\alpha+\beta x_{it}+\eta_i)\frac{\partial}{\partial\eta_i}\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i) \quad \text{(B13)}$$

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

$$\begin{aligned}\frac{\partial}{\partial\eta_i}\Pr(y_{it}=1|y_{i0},\eta_i,x_i) &= \frac{\partial}{\partial\eta_i}\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)\left(F(\alpha+\beta x_{it}+\eta_i)-F(\beta x_{it}+\eta_i)\right)\\ &\quad+\Pr(y_{it-1}=1|y_{i0},\eta_i,x_i)\left(f(\alpha+\beta x_{it}+\eta_i)-f(\beta x_{it}+\eta_i)\right)+\\ &\quad+f(\beta x_{it}+\eta_i), \text{ for } t>1 \quad \text{(B14)}\\ \frac{\partial}{\partial\eta_i}\Pr(y_{i1}=1|y_{i0},\eta_i,x_i) &= f(\alpha y_{i0}+\beta x_{i1}+\eta_i) \quad \text{(B15)}\end{aligned}$$

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

$$\sum_{t=1}^{T-1}y_{it}=\sum_{t=1}^{T-1}F(\alpha y_{it-1}+\beta x_{it}+\eta_i) \quad \text{(B16)}$$

Deriving the previous equation with respect to $\alpha$ :

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

Therefore:

$$\frac{\partial\widehat{\eta}_i(\alpha,\beta)}{\partial\alpha}=\frac{-\sum_{t=1}^{T-1}y_{it-1}f(\alpha y_{it-1}+\beta x_{it}+\eta_i)}{\sum_{t=1}^{T-1}f(\alpha y_{it-1}+\beta x_{it}+\eta_i)} \quad \text{(B17)}$$

The modified first order condition for $\beta$ 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 of computing the expectations needed for the modification works regardless of the density function $f$ assumed.

# C Appendix: Concentrating the likelihood and estimating with fixed effects

A problem that arise on the maximization of the log likelihood function

$$\log L = \sum_{i=1}^{N}\sum_{t=1}^{T-1}\{y_{it} * \log F(\alpha y_{it-1} + x_{it}\beta + \eta_i) + (1 - y_{it}) * \log(1 - F(\alpha y_{it-1} + \beta x_{it} + \eta_i))\} \quad \text{(C1)}$$

is that we have to estimate $N$ parameters corresponding to the fixed effects, implying a second derivatives matrix with $N + 2$ rows and columns. A way of proceeding is using some results from matrix algebra suggested in Chamberlain (1980) in order to simplify the computation of the inverse of the Hessian. Heckman and MaCurdy (1980) divided the optimization problem in two problems:
1.- The maximization with respect to $\alpha$ and $\beta$ of the log likelihood given a value of the fixed effects $\{\eta_i\}_{i=1}^{N}$.
2.- The maximization of the log likelihood function for each $\eta_i$ given the estimation of $\alpha$ and $\beta$: $\log L_i|(\alpha, \beta) = \sum_{t=1}^{T-1}\{y_{it} * \log F_{it} + (1 - y_{it}) * \log(1 - F_{it})\}$. This gives $N$ isolated maximization problems.

They suggested iterating back and forth between the two problems until convergence is achieved. Apart from the issue of whether or not the back and forth iteration will converge to the true maximum of the log-likelihood, a shortcoming of this way of proceeding is that the estimated variance for the estimator of $\alpha$ and $\beta$ will be too small. The Hessian of the log-likelihood function is not block diagonal, so the estimator of the variance based on the first program does not obtain the correct submatrix of the inverse information matrix.

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 nonlinearity, we can not get a explicit expression of the fixed effects estimators as functions of $\alpha$ and $\beta$, I make numerical substitution of them on the estimating equations of $\gamma = (\alpha, \beta)'$, i.e. the estimator of $\gamma$ solves

$$\sum_{i=1}^{N}\left\{d_{\gamma i}(\gamma, \widehat{\eta}_i(\gamma)) + d_{\eta_i}(\gamma, \widehat{\eta}_i(\gamma))\frac{\partial \widehat{\eta}_i(\gamma)}{\partial \gamma}\right\} = \sum_{i=1}^{N} d_{\gamma i}(\gamma, \widehat{\eta}_i(\gamma)) = 0 \quad \text{(C2)}$$

where $\widehat{\eta}_i(\gamma)$ is the number that makes $d_{\eta_i}(\gamma, \eta_i) = 0$ for the value of $\gamma$ in which we are evaluating the estimating equations; $d_{\eta_i}(\gamma, \eta_i) \equiv \frac{\partial l_i(\gamma, \eta_i)}{\partial \eta_i}$ and $d_{\gamma i}(\gamma, \eta_i) \equiv \frac{\partial l_i(\gamma, \eta_i)}{\partial \gamma}$. So, we use a Gauss-Newton type algorithm to solve equation (C2) with respect to $\gamma$, and in each step $\widehat{\eta}_i$ is computed such that for the value of $\gamma$ in that step ($\gamma_s$), $d_{\eta_i}(\gamma_s, \eta_i)$ equals zero. Thus, the equation for each of the $\eta_i$ is nested in the algorithm that maximizes the concentrated likelihood. In each step, we have to solve $N$ single nonlinear equations, one for each of the fixed effects. $d_{\eta_i}(\gamma_s, \eta_i) = 0$ is easily solve 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 $\alpha$ and $\beta$ is not made for each given estimated value of the fixed effects. Instead of that the

values of the fixed effects are change accordingly in each step of the estimation process of $\alpha$ and $\beta$; just as if we were able to analytically find $\widehat{\eta}_i(\gamma)$.

To overcome the already mentioned problem of estimating the variance, 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 (C2) with respect to $\gamma$, the Hessian is equal to:

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

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

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

# References


[1] Altonji, J. G. and R. L. Matzkin (2001): " Panel Data Estimators for Nonseparable Models with Endogenous Regressors", *Technical Working Paper 267,* National Bureau of Economic Research.

[2] Alvarez, J. and M. Arellano (1998): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators", *Working Paper 9808,* CEMFI, Madrid; *forthcoming in Econometrica.*

[3] Arellano, M.(2001): "Discrete Choice with Panel Data" *Working Paper 0101,* CEMFI, Madrid.

[4] Arellano, M. and B. Honoré (2001): "Panel Data Models: Some Recent Developments", in Heckman, J. J. and E. Leamer (eds.) *Handbook of Econometrics*, vol. 5, Elsevier Science, Amsterdam.

[5] Blundell, R. and J. L. Powell (2000): "Endogeneity in Nonparametric and Semiparametric Regression Models" *mimeo.* Prepared for the Econometric Society World Meetings Seattle, August 2000.

[6] Chamberlain, G. (1984): "Panel Data", in Griliches, Z. and M.D. Intriligator (eds.) *Handbook of Econometrics*, vol. 2, Elsevier Science, Amsterdam.

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

[8] Chamberlain, G. (1992): "Binary Response Models for Panel Data: Identification and Information" *Manuscript, Department of Economics*, Harvard University.

[9] Chay, K. Y. and D. R. Hyslop (2000) "Identification and Estimation of Dynamic Binary Response Panel Data Models: Empirical Evidence using Alternative Approaches" *Manuscript,* UCLA.

[10] Carro, J. M. (2002) "Intertemporal Female Labor Force Participation With Non-Exogenous Children", *in progress*, CEMFI, Madrid.

[11] Cox, D. R. and N. Reid (1987): "Parameter Orthogonality and Approximate Conditional Inference." *Journal of the Royal Statistical Society, Series B*, 49, 1-39.

[12] Ferguson, H., N. Reid and D. R. Cox (1991): "Estimating Equations from Modified Profile Likelihood", in Godambe, V. P. (ed.), *Estimating Functions*, Oxford University Press.

[13] Heckman, J.J. (1981a): "Statistical Models for Discrete Panel Data" in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. McFadden (eds.), MIT Press.

[14] Heckman, J.J. (1981b): "The incidental parameters Problem and the Problem of Initial Conditions in Estimating a Discrete-Time Data Stochastic Process" in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. McFadden (eds.), MIT Press.

[15] Heckman, J. J. and J. Smith (1997), with the assistance of N. Clements :"Making the most out of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts", *Review of Economic Studies*, 64, 487-535.

[16] Heckman, J. J. and T. E. MaCurdy (1980): "A life Cycle Model of Female Labour Supply", *Review of Economic Studies*, 47, 47-74.

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

[18] Hyslop, D. R. (1999): "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women", *Econometrica*, 67, 1255-1294.

[19] Lancaster, T. (1997): "Orthogonal Parameters and Panel Data", *Working Paper No. 97-32*, Department of Economics, Brown University.

[20] Manski, C. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data", *Econometrica*, 55, 357-362.

[21] Neyman J., and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations", *Econometrica*, 16, 1-32.

[22] Woutersen, T. (2001): "Robustness Against Incidental Parameters and Mixing Distributions", *Manuscript, Department of Economics*, University of Western Ontario.

Figure 1: Smoothed density of the effect of a change in $x_{iT-}$ over the probability of $(y_{iT-1} = 1)$ for a simulated sample from the Logit design, with $T = 8$ and $N = 500$.

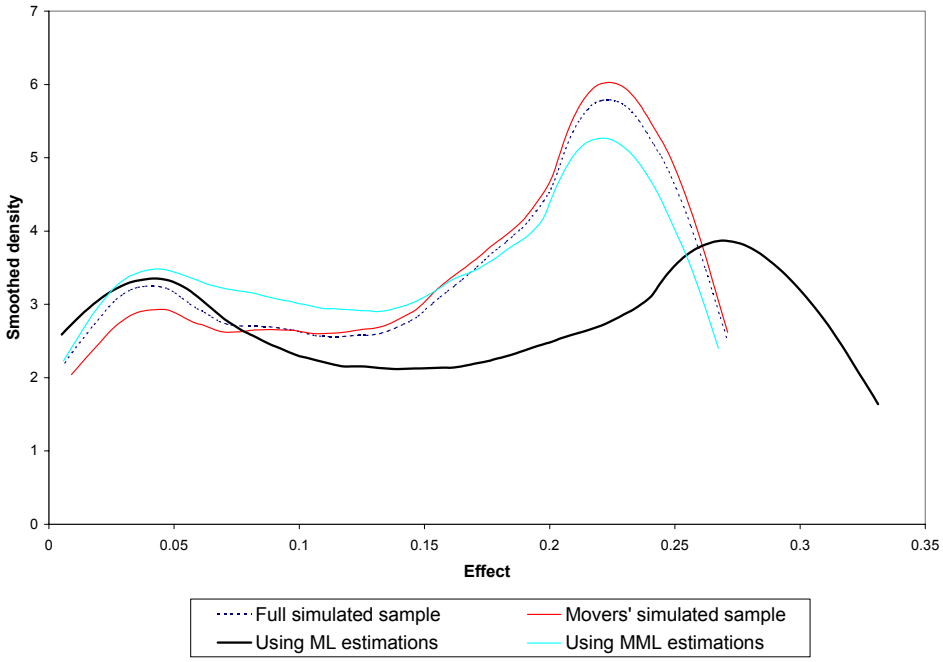

Figure 2: Smoothed density of the effect of a change in $x_{iT-1}$ over the probability of $(y_{iT-1} = 1)$ for a simulated sample from the Logit design, with $T = 8$ and $N = 10000$.

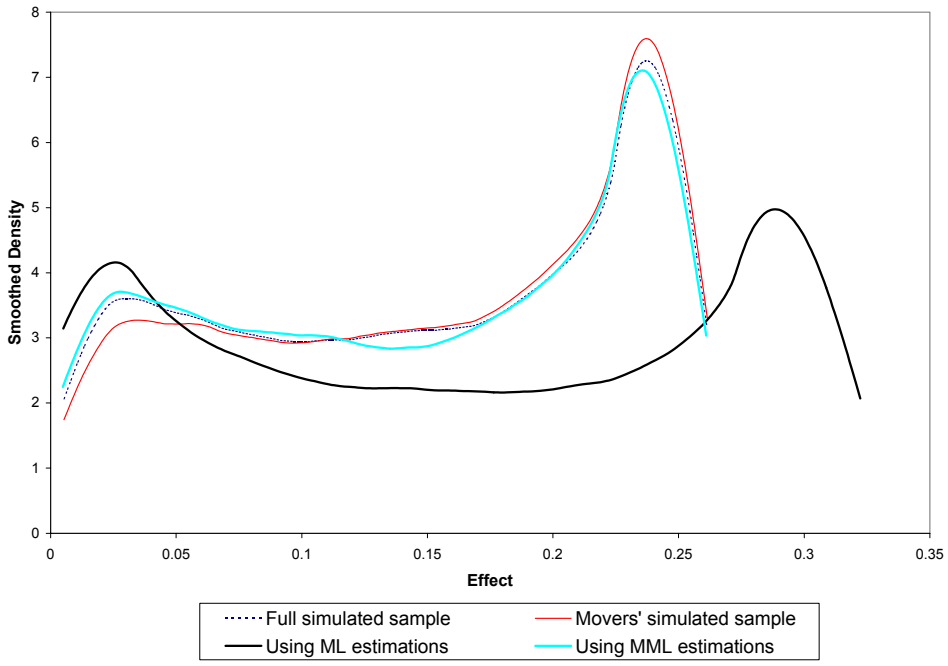


Figure 3: Smoothed density of the effect of having one more child beetween 0 and 2 years old at period t=1

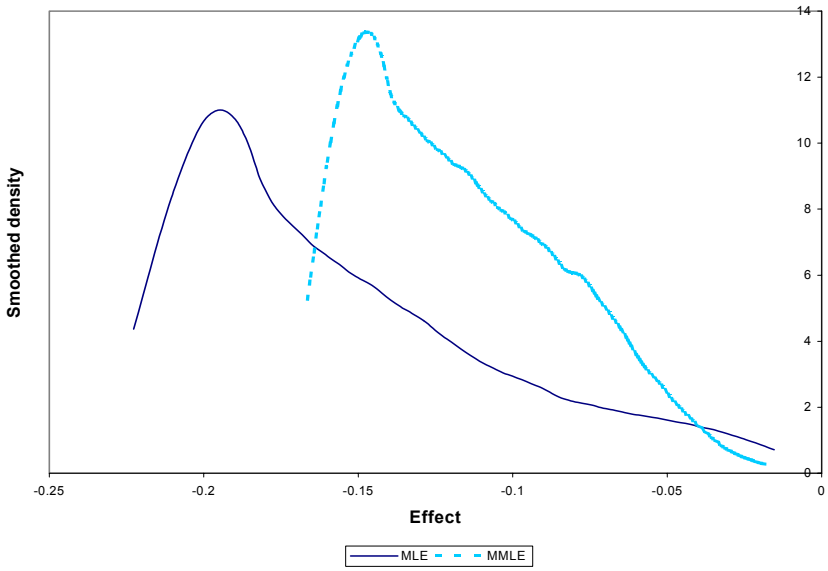

Figure 4: Effect of a birth in period one for an individual with the average characteristics.

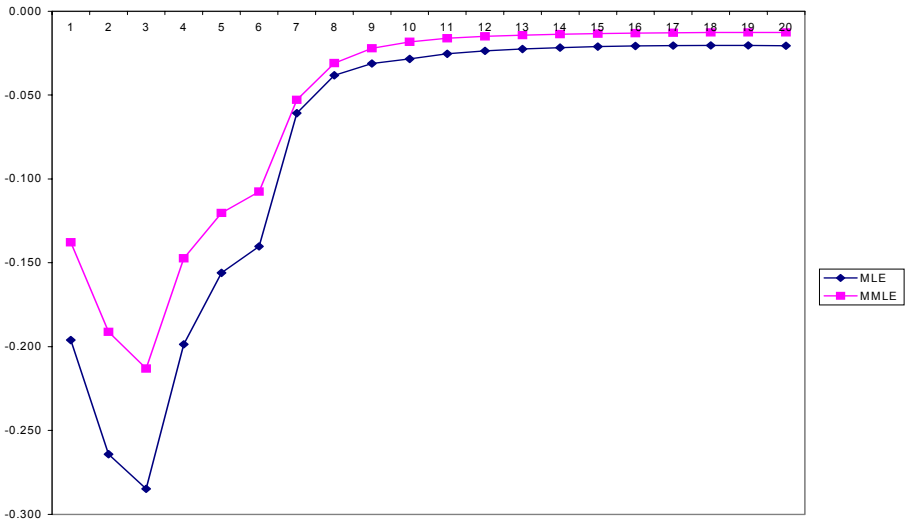


Figure 5: Dynamic Mean Effect of a birth in period one