Identification and estimation of marginal effects in nonlinear panel models

Victor Chernozhukov
Iván Fernández-Val
Jinyong Hahn
Whitney Newey

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP25/08

Identification and Estimation of Marginal Effects in Nonlinear Panel Models [1]

Victor Chernozhukov, MIT
Iván Fernández-Val, BU
Jinyong Hahn, UCLA
Whitney Newey, MIT

September 9, 2008

[1]First version of May 2007. We thank J. Angrist, B. Graham, and seminar participants of CEMFI, CEMMAP Microeconometrics: Measurement Matters Conference, CEMMAP Inference in Partially Identified Models with Applications Conference, Harvard/MIT, MIT, UC Berkeley, USC, and 2007 WISE Panel Data Conference for helpful comments. Chernozhukov, Fernández-Val, and Newey gratefully acknowledge research support from the NSF.

Abstract

This paper gives identification and estimation results for marginal effects in nonlinear panel models. We find that linear fixed effects estimators are not consistent, due in part to marginal effects not being identified. We derive bounds for marginal effects and show that they can tighten rapidly as the number of time series observations grows. We also show in numerical calculations that the bounds may be very tight for small numbers of observations, suggesting they may be useful in practice. We give an empirical illustration.

1 Introduction & Motivation

Marginal effects are commonly used in practice to quantify the effect of variables on an outcome of interest. They are known as average treatment effects, average partial effects, and average structural functions in different contexts (e.g., see Wooldridge, 2002, Blundell and Powell, 2003). In panel data marginal effects average over unobserved individual heterogeneity. Chamberlain (1984) gave important results on identification of marginal effects in nonlinear panel data using control functions. Our paper gives identification and estimation results for marginal effects in panel data under strict exogeneity, time stationarity, and discrete regressors.

It is sometimes thought that marginal effects can be estimated using linear models, as shown by Hahn (2001) in an example and Wooldridge (2005) under strong independence conditions. We find that the situation is more complicated. The marginal effect may not be identified. Furthermore, with a binary regressor the linear model uses the wrong weighting in estimation when the number of time periods T exceeds three. We show that correct weighting can be obtained by averaging individual regression coefficients. We also derive bounds for the marginal effect when it is not identified.

We find that these bounds can be wide when no restrictions are placed on the outcome, but tighten substantially for some semiparametric models. In binary choice models with additive heterogeneity we find in numerical results that the bounds can be very tight even when T is small. We also give theorems showing that the bounds tighten quickly as T grows.

These results suggest how the bounds can be used in practice. Although they can be difficult to compute for large T, their tightness for small T makes it feasible to compute them for different small time intervals and combine results to improve efficiency. To illustrate their usefulness we provide an empirical illustration based on Chamberlain's (1984) labor force participation example.

This paper is closely related to Honoré and Tamer (2006) and Chernozhukov, Hahn, and Newey (2004). These papers derived bounds for slope coefficients in autoregressive and static models respectively. Here we focus on marginal effects and give results on the rate of convergence of bounds as T grows. Also, we find that the linear programming algorithm proposed by Honoré and Tamer (2006) needs to be replaced in practice by some other method, and here propose using quadratic minimum distance. We give empirical results.

Browning and Carro (2007) give results on marginal effects in autoregressive panel models. They find that more than additive heterogeneity is needed to describe some interesting application. They also find that marginal effects are not generally identified in dynamic models. Graham and Powell (2008) consider identification with continuous regressors.

Hahn and Newey (2004) gave theoretical and simulation results showing that fixed effects estimators of marginal effects in nonlinear models may have little bias, as suggested by Wooldridge (2002). Fernández-Val (2008) found that averaging fixed effects estimates of individual marginal effects has bias that shrinks faster as T grows than does the bias of slope coefficients. We show that, with small T, fixed effects consistently estimates an identified component of the marginal effects. We also give numerical results showing that the bias of fixed effects estimators of the marginal effect is very small in a range of examples.

The bounds approach we take is different from the bias correction methods of Hahn and Kuersteiner (2002), Alvarez and Arellano (2003), Woutersen (2002), Hahn and Newey (2004), Hahn and Kuersteiner (2007), and Fernández-Val (2008). The bias corrections are based on large T approximations. The bounds approach takes explicit account of possible nonidentification for fixed T. Inference accuracy of bias corrections will depend on T being the right size relative to the number of cross-section observations n, while inference for bounds does not.

In Section 2 we give a general nonparametric conditional mean model with correlated unobserved individual effects and analyze the properties of linear estimators. Section 3 gives bounds for marginal effects in these models and results on the rate of convergence of these bounds as T grows. Section 4 gives similar results, with tighter bounds, in a binary choice model with a location shift individual effect. Section 5 gives results and numerical examples on calculation of population bounds. Section 6 discusses estimation and Section 7 inference. Section 8 gives an empirical example.

2 A Conditional Mean Model and Linear Estimators

The data consist of n observations of time series $Y_i = (Y_{i1}, ..., Y_{iT})'$ and $X_i = [X_{i1}, ..., X_{iT}]'$, for a dependent variable Y_{it} and a vector of regressors X_{it}. We will assume throughout that (Y_i, X_i), $(i = 1, ..., n)$, are independent and identically distributed observations. A case we consider in some depth is binary choice panel data where $Y_{it} \in \{0, 1\}$. For simplicity we also give some results for binary X_{it}, where $X_{it} \in \{0, 1\}$.

A general model we consider is a nonseparable conditional mean model as in Wooldridge (2005). Here there is an unobserved individual effect α_i and a function $m(x, \alpha)$ such that

$$E[Y_{it}|X_i, \alpha_i] = m(X_{it}, \alpha_i), (t = 1, ..., T). \tag{1}$$

The individual effect α_i may be a vector of any dimension. For example, α_i could include individual slope coefficients in a binary choice model, where $Y_{it} \in \{0, 1\}$, $F(\cdot)$ is a CDF, and

$$\Pr(Y_{it} = 1|X_i, \alpha_i) = E[Y_{it}|X_i, \alpha_i] = F(X_{it}'\alpha_{i2} + \alpha_{i1}).$$

Such models have been considered by Browning and Carro (2007) in a dynamic setting. More familiar models with scalar α_i are also included. For example, the binary choice model with an individual location effect has

$$\Pr(Y_{it} = 1|X_i, \alpha_i) = E[Y_{it}|X_i, \alpha_i] = F(X_{it}'\beta^* + \alpha_{i1}).$$

This model has been studied by Chamberlain (1980, 1984, 1992), Hahn and Newey (2004), and others. The familiar linear model $E[Y_{it}|X_i, \alpha_i] = X_{it}'\beta^* + \alpha_i$ is also included as a special case of the general conditional mean model.

The two critical assumptions made in equation (1) are that X_i is strictly exogenous conditional on α and that $m(x, \alpha)$ does not vary with time. These conditions lead to identification from differences across time. Without time stationarity, identification becomes more difficult.

Our primary object of interest is the marginal effect given by

$$\mu_0 = \frac{\int [m(\tilde{x}, \alpha) - m(\bar{x}, \alpha)] Q^*(d\alpha)}{D},$$

where $\tilde{x}$ and $\bar{x}$ are two possible values for the X_{it} vector, Q^* denotes the marginal distribution of α, and D is the distance, or number of units, corresponding to $\tilde{x} - \bar{x}$. This object gives the average, over the marginal distribution, of the per unit effect of changing x from $\bar{x}$ to $\tilde{x}$. It is the average treatment effect in the treatment effects literature. For example, suppose $\bar{x} = (\bar{x}_1, x_2')'$ where $\bar{x}_1$ is a scalar, and $\tilde{x} = (\tilde{x}_1, x_2')'$. Then $D = \tilde{x}_1 - \bar{x}_1$ would be an appropriate distance measure and

$$\mu_0 = \frac{\int [m(\tilde{x}_1, x_2, \alpha) - m(\bar{x}_1, x_2, \alpha)] Q^*(d\alpha)}{\tilde{x}_1 - \bar{x}_1},$$

would be the per unit effect of changing the first component of X_{it}. Here one could also consider averages of the marginal effects over different values of x_2.

For example, consider an individual location effect for binary Y_{it} where $m(x, \alpha) = F(x'\beta_0 + \alpha)$. Here the marginal effect will be

$$\mu_0 = D^{-1} \int [F(\tilde{x}'\beta^* + \alpha) - F(\bar{x}'\beta^* + \alpha)] Q^*(d\alpha).$$

The restrictions this binary choice model places on the conditional distribution of Y_{it} given X_i and α_i will be useful for bounding marginal effects, as further discussed below.

In this paper we focus on the discrete case where the support of X_i is a finite set. Thus, the events $X_{it} = \tilde{x}$ and $X_{it} = \bar{x}$ have positive probability and no smoothing is required. It would also be interesting to consider continuous X_{it}.

Linear fixed effect estimators are used in applied research to estimate marginal effects. For example, the linear probability model with fixed effects has been applied when Y_{it} is binary. Unfortunately, this estimator is not generally consistent for the marginal effect. There are two reasons for this. The first is the marginal effect is generally not identified, as further explained below. Second, the fixed effects estimator uses incorrect weighting.

To explain, we compare the limit of linear fixed effects estimators with the marginal effect μ_0. Suppose that X_i has finite support $\{X^1, ..., X^K\}$ and let $Q_k^*(\alpha)$ denote the CDF of the distribution of α conditional on $X_i = X^k$. Define

$$\mu_k = \int [m(\tilde{x}, \alpha) - m(\bar{x}, \alpha)] Q_k^*(d\alpha)/D, \quad \mathcal{P}_k = \Pr(X_i = X^k).$$

This μ_k is the marginal effect conditional on the entire time series $X_i = [X_{i1}, ..., X_{iT}]'$ being equal to X^k. By iterated expectations,

$$\mu_0 = \sum_{k=1}^{K} \mathcal{P}_k \mu_k. \tag{2}$$

We will compare this formula with the limit of linear fixed effects estimators.

An implication of the conditional mean model that is crucial for identification is

$$E[Y_{it}|X_i = X^k] = \int m(X_t^k, \alpha) Q_k^*(d\alpha). \tag{3}$$

This equation allows us to identify some of the μ_k from differences across time periods of identified conditional expectations.

To simplify the analysis of linear fixed effect estimators we focus on binary $X_{it} \in \{0, 1\}$. Consider $\hat{\beta}_w$ from least squares on

$$Y_{it} = X_{it}\beta + \gamma_i + v_{it}, (t = 1, ..., T; i = 1, ..., n),$$

where each γ_i is estimated. This is the usual within estimator, where for $\bar{X}_i = \sum_{t=1}^{T} X_{it}/T$,

$$\hat{\beta}_w = \frac{\sum_{i,t}(X_{it} - \bar{X}_i)Y_{it}}{\sum_{i,t}(X_{it} - \bar{X}_i)^2}.$$

Here the estimator of the marginal effect is just $\hat{\beta}_w$. To describe its limit, let $r^k = \#\{t : X_t^k = 1\}/T$ be the proportion of component of X^k that are equal to one and $\sigma_k^2 = r^k(1 - r^k)$ be the variance of a binomial with probability r^k.

THEOREM 1: *If equation (1) is satisfied, (X_i, Y_i) has finite second moments, and $\sum_{k=1}^{K} \mathcal{P}_k \sigma_k^2 > 0$, then*

$$\hat{\beta}_w \xrightarrow{p} \frac{\sum_{k=1}^{K} \mathcal{P}_k \sigma_k^2 \mu_k}{\sum_{k=1}^{K} \mathcal{P}_k \sigma_k^2}. \tag{4}$$

Comparing equations (2) and (4) we see that the linear fixed effects estimator converges to a weighted average of μ_k, weighted by σ_k^2, rather than the simple average in equation (2). The weights are never completely equal, so that the linear fixed effects estimator is not consistent for the marginal effect unless μ_k does not depend on k, i.e., unless the distribution of α given

$X_i = X^k$ does not vary with k (in its effect on μ_k). This amounts to exogeneity of α as far as the marginal effect goes, which is not very interesting.

One reason for inconsistency of $\hat{\beta}_w$ is that certain μ_k receive zero weight. For notational purposes let $X^1 = (0, ..., 0)'$ and $X^K = (1, ..., 1)'$ (where we implicitly assume that these are included in the support of X_i). Note that $\sigma_1^2 = \sigma_K^2 = 0$ so that μ_1 and μ_K are not included in the weighted average. The explanation for their absence is that μ_1 and μ_K are not identified. These are marginal effects conditional on X_i equal a vector of constants, where there are no changes over time to help identify the effect from equation (3).

Another reason for inconsistency of $\hat{\beta}_w$ is that for $T \geq 4$ the weights on μ_k will be different than the corresponding weights for μ_0. This is because r^k varies for $k \notin \{1, K\}$ except when $T = 2$ or $T = 3$.

This result is different from Hahn (2001), who found that $\hat{\beta}_w$ consistently estimates the marginal effect. The reason he obtained such a result is that he restricted the support of X_i to exclude both $(0, ..., 0)'$ or $(1, ..., 1)'$. Also, he only considered a case with $T = 2$. Thus, neither feature that causes inconsistency of $\hat{\beta}_w$ was present in his example. Thus, as noted by Hahn (2001), the conditions that lead to consistency of the linear fixed effects estimator in his example are quite special.

The inconsistency result is also different from Wooldridge (2005). There it is shown that if $b_i = m(1, \alpha_i) - m(0, \alpha_i)$ is mean independent of $X_{it} - \bar{X}_i$ for each t then linear fixed effects is consistent. The problem is that this independence assumption is very strong when X_{it} is discrete. Note that for $T = 2$, $X_{i2} - \bar{X}_i$ takes on the values 0 when $X_i = (1, 1)$ or $(0, 0)$, $-1/2$ when $X_i = (1, 0)$, and $1/2$ when $X_i = (0, 1)$. Thus mean independence of b_i and $X_{i2} - \bar{X}_i$ actually implies that $\mu_2 = \mu_3$ and that these are equal to the marginal effect conditional on $X_i \in \{X^1, X^4\}$. This is quite close to independence of b_i and X_i, which is not very interesting if we want to allow correlation between the regressors and the individual effect.

The result of Theorem 1 is related to Angrist (1998), who found that the probability limits of linear regression estimators are variance weighted average effects in cross sectional models with heterogenous effects. He focuses on estimation of averages of the identified effects.

The lack of identification of μ_1 and μ_K means the marginal effect is actually not identified. Therefore, no consistent estimator of it exists. Nevertheless, it is possible to find informative bounds for μ_0, as we show in the following sections.

We can correct the second reason for inconsistency of $\hat{\beta}_w$ by modifying the estimator. A simple way to do this is to estimate a different slope coefficient for each individual and then average. This estimator is obtained from averaging across individuals the least squares estimates of β_i in

$$Y_{it} = X_{it}\beta_i + \gamma_i + v_{it}, (t = 1, ..., T; i = 1, ..., n),$$

For $s_{xi}^2 = \sum_{t=1}^T (X_{it} - \bar{X}_i)^2$ and $n^* = \sum_{i=1}^n 1(s_{xi}^2 > 0)$, this estimator takes the form

$$\hat{\beta} = \frac{1}{n^*} \sum_{i=1}^n 1(s_{xi}^2 > 0) \frac{\sum_{t=1}^T (X_{it} - \bar{X}_i) Y_{it}}{s_{xi}^2}.$$

This is equivalent to running least squares in the model

$$Y_{it} = \beta_k X_{it} + \gamma_k + v_{it}, \tag{5}$$

for individuals with $X_i = X^k$, and averaging $\hat{\beta}_k$ over k weighted by the sample frequencies of X^k.

THEOREM 2: *If equation (1) is satisfied and (X_i, Y_i) have finite second moments then*

$$\hat{\beta} \xrightarrow{p} \mu_I = \sum_{k=2}^{K-1} \mathcal{P}_k^* \mu_k, \tag{6}$$

where $\mathcal{P}_k^* = \mathcal{P}_k / \sum_{k=2}^{K-1} \mathcal{P}_k$.

To see how big the inconsistency can be we consider a numerical example, where $X_{it} \in \{0, 1\}$ is i.i.d across i and t, $\Pr(X_{it} = 1) = p_X$, η_{it} is i.i.d. $N(0, 1)$,

$$Y_{it} = 1(X_{it} + \alpha_i + \eta_{it} > 0), \quad \alpha_i = \sqrt{T}(\bar{X}_i - p_X)/p_X(1 - p_X).$$

Here we consider the marginal effect for $\tilde{x} = 1, \bar{x} = 0, D = 1$, given by

$$\mu_0 = \int [\Phi(1 + \alpha) - \Phi(\alpha)] Q^*(d\alpha).$$

Table 1 and Figure 1 give numerical values for $\left[\lim(\hat{\beta}_w) - \mu_0\right]/\mu_0$ and $\left[\lim(\hat{\beta}) - \mu_0\right]/\mu_0$ for several values of T and p_X.

We find that the biases (inconsistencies) can be large in percentage terms. We also find that biases are largest when p_X is small. In this example, the inconsistency of fixed effects estimators of marginal effects seems to be largest when the regressor values are sparse. Also we find that differences between the limits of $\hat{\beta}$ and $\hat{\beta}_w$ are larger for larger T, which is to be expected due to the weights differing more for larger T.

The estimator $\hat{\beta}$ of the identified marginal effect μ_I can easily be extended to any discrete X_{it}. To describe the extension, let $\tilde{d}_{it} = 1(X_{it} = \tilde{x}), \bar{d}_{it} = 1(X_{it} = \bar{x})$, $\tilde{r}_i = \sum_{t=1}^T \tilde{d}_{it}/T, \bar{r}_i = \sum_{t=1}^T \bar{d}_{it}/T$, and $n^* = \sum_{i=1}^n 1(\tilde{r}_i > 0) 1(\bar{r}_i > 0)$. The estimator is given by

$$\hat{\beta} = \frac{1}{n^*} \sum_{i=1}^n 1(\tilde{r}_i > 0) 1(\bar{r}_i > 0) [\frac{\sum_{t=1}^T \tilde{d}_{it} Y_{it}}{T \tilde{r}_i} - \frac{\sum_{t=1}^T \bar{d}_{it} Y_{it}}{T \bar{r}_i}].$$

This estimator is the same as doing individual by individual least squares on a fully saturated model and then averaging the result. It will be identical to the previous $\hat{\beta}$ when X_{it} is binary.

It should be noted that $\hat{\beta}$ is not efficient for $T \geq 3$. The reason is that it is least squares over time, which does not account properly for time series heteroskedasticity or autocorrelation. An efficient estimator could be obtained by a minimum distance procedure, though that is complicated. Also, one would have only few observations to estimate needed weighting matrices, so its properties may not be great in small to medium sized samples. For these reasons we leave construction of an efficient estimator to future work.

To describe the limit of the estimator $\hat{\beta}$ in general, let $\mathcal{K}^* = \{k$:there is $\tilde{t}$ and $\bar{t}$ such that $X^k_{\tilde{t}} = \tilde{x}$ and $X^k_{\bar{t}} = \bar{x}\}$. This is the set of possible values for X_i where both $\tilde{x}$ and $\bar{x}$ occur for at least one time period, allowing identification of the marginal effect from differences. For all other values of k, either $\tilde{x}$ or $\bar{x}$ will be missing from the observations and the marginal effect will not be identified. In the next Section we will consider bounds for those effects.

Theorem 3: *If equation (1) is satisfied,* (X_i, Y_i) *have finite second moments and* $\sum_{k\in\mathcal{K}^*} \mathcal{P}_k > 0$, *then*

$$\hat{\beta} \xrightarrow{p} \mu_I = \sum_{k\in\mathcal{K}^*} \mathcal{P}^*_k \mu_k,$$

where $\mathcal{P}^*_k = \mathcal{P}_k / \sum_{k\in\mathcal{K}^*} \mathcal{P}_k$.

3 Bounds in the Conditional Mean Model

Although the marginal effect μ_0 is not identified it is straightforward to bound it. Also, as we will show below, these bounds can be quite informative, motivating the analysis that follows. Some additional notation is useful for describing the results. Let

$$\bar{m}^k_t = E[Y_{it}|X_i = X^k]/D$$

be the identified conditional expectations of each time period observation on Y_{it} conditional on the k^{th} support point. Also, let $\Delta(\alpha) = [m(\tilde{x}, \alpha) - m(\bar{x}, \alpha)]/D$. The next result gives identification and bound results for μ_k, which can then be used to obtain bounds for μ_0.

Lemma 4: *If there is* $\tilde{t}$ *and* $\bar{t}$ *such that* $X^k_{\tilde{t}} = \tilde{x}$ *and* $X^k_{\bar{t}} = \bar{x}$ *then*

$$\mu_k = \bar{m}^k_{\tilde{t}} - \bar{m}^k_{\bar{t}}.$$

Suppose that $B_\ell \leq m(x, \alpha)/D \leq B_u$. *If there is* $\tilde{t}$ *such that* $X^k_{\tilde{t}} = \tilde{x}$ *then*

$$\bar{m}^k_{\tilde{t}} - B_u \leq \mu_k \leq \bar{m}^k_{\tilde{t}} - B_\ell.$$

Also, if there is $\bar{t}_k$ *such that* $X^k_{\bar{t}} = \bar{x}$ *then*

$$B_\ell - \bar{m}^k_{\bar{t}} \leq \mu_k \leq B_u - \bar{m}^k_{\bar{t}}.$$

Suppose that $\Delta(\alpha)$ has the same sign for all α. Then if for some k there is $\tilde{t}$ and $\bar{t}$ such that $X_{\tilde{t}}^k = \tilde{x}$ and $X_{\bar{t}}^k = \bar{x}$, the sign of $\Delta(\alpha)$ is identified. Furthermore, if $\Delta(\alpha)$ is positive then the lower bounds may be replaced by zero and if $\Delta(\alpha)$ is negative then the upper bounds may be replaced by zero.

The bounds on each μ_k can be combined to obtain bounds for the marginal effect μ_0. Let

$$\begin{aligned}
\tilde{\mathcal{K}} &= \{k : \text{there is } \tilde{t} \text{ such that } X_{\tilde{t}}^k = \tilde{x} \text{ but no } \bar{t} \text{ such that } X_t^k = \bar{x}\}, \\
\bar{\mathcal{K}} &= \{k : \text{ there is } \bar{t} \text{ such that } X_{\tilde{t}}^k = \bar{x} \text{ but no } \tilde{t} \text{ such that } X_t^k = \tilde{x}\}.
\end{aligned}$$

Also, let $P^0 = \Pr(X_i : X_{it} \neq \tilde{x}$ and $X_{it} \neq \bar{x}$ for every $t)$. The following result is obtained by multiplying the k^{th} bound in Lemma 4 by $\mathcal{P}_k$ and summing.

THEOREM 5: *If $B_\ell \leq m(x,\alpha)/D \leq B_u$ then $\mu_\ell \leq \mu_0 \leq \mu_u$ for*

$$\begin{aligned}
\mu_\ell &= \mathcal{P}^0(B_\ell - B_u) + \sum_{k \in \tilde{K}} \mathcal{P}_k(\bar{m}_{\tilde{t}}^k - B_u) + \sum_{k \in \bar{K}} \mathcal{P}_k(B_\ell - \bar{m}_{\bar{t}}^k) + \sum_{k \in K^*} \mathcal{P}_k \mu_k, \\
\mu_u &= \mathcal{P}^0(B_u - B_\ell) + \sum_{k \in \tilde{K}} \mathcal{P}_k(\bar{m}_{\tilde{t}}^k - B_\ell) + \sum_{k \in \bar{K}} \mathcal{P}_k(B_u - \bar{m}_{\bar{t}}^k) + \sum_{k \in K^*} \mathcal{P}_k \mu_k.
\end{aligned}$$

If $\Delta(\alpha)$ has the same sign for all α and there is some k^* such that $X_{\tilde{t}}^{k^*} = \tilde{x}$ and $X_{\bar{t}}^{k^*} = \bar{x}$, the sign of μ_0 is identified, and if $\mu_0 > 0$ (< 0) then μ_ℓ (μ_u) can be replaced by $\sum_{k \in K^*} \mathcal{P}_k \mu_k$.

As an example, consider the binary X case where $X_{it} \in \{0,1\}$, $\tilde{x} = 1$, and $\bar{x} = 0$. Let X^K denote a $T \times 1$ unit vector and X^1 be the $T \times 1$ zero vector, assumed to lie in the support of X_i. Here the bounds will be

$$\begin{aligned}
\mu_\ell &= \mathcal{P}_K(\bar{m}_{\tilde{t}}^K - B_u) + \mathcal{P}_1(B_\ell - \bar{m}_{\bar{t}}^1) + \sum_{1<k<K} \mathcal{P}_k \mu_k, \\
\mu_u &= \mathcal{P}_K(\bar{m}_{\tilde{t}}^K - B_\ell) + \mathcal{P}_1(B_u - \bar{m}_{\bar{t}}^1) + \sum_{1<k<K} \mathcal{P}_k \mu_k.
\end{aligned} \tag{7}$$

It is interesting to ask how the bounds behave as T grows. If the bounds converge to μ_0 as T goes to infinity then μ_0 is identified for infinite T. If the bounds converge rapidly as T grows then one might hope to obtain tight bounds for T not very large. The following result gives a simple condition under which the bounds converge to μ_0 as T grows.

THEOREM 6: *Suppose that $B_\ell \leq m(x,\alpha)/D \leq B_u$ and $\overrightarrow{X}_i = (X_{i1}, X_{i2}, ...)$ is stationary and, conditional on α_i, the support of each X_{it} is the marginal support of X_{it} and $\overrightarrow{X}_i$ is ergodic. Then $\mu_\ell \longrightarrow \mu_0$ and $\mu_u \longrightarrow \mu_0$ as $T \longrightarrow \infty$.*

The rate at which the bounds converge in the general model is a complicated question. Here we will address it in an example and leave general treatment to another setting. The example we consider is that where $X_{it} \in \{0,1\}$.

THEOREM 7: *If $B_\ell \leq m(x,\alpha)/D \leq B_u$ and $\overrightarrow{X_i}$ is i.i.d. conditional on α_i then for $P(\alpha_i) = \Pr(X_{it} = 1|\alpha_i)$,*

$$\max\{|\mu_\ell - \mu_0|, |\mu_u - \mu_0|\} \leq (B_u - B_\ell)E[\{1 - P(\alpha_i)\}^T + P(\alpha_i)^T].$$

If there is $\varepsilon > 0$ such that $\varepsilon \leq P(\alpha_i) \leq 1 - \varepsilon$ for almost every α_i, then

$$\max\{|\mu_\ell - \mu_0|, |\mu_u - \mu_0|\} \leq (B_u - B_\ell)2(1 - \varepsilon)^T.$$

If $P(\alpha_i) = 1$ or $P(\alpha_i) = 0$ with positive probability either $\mu_\ell \nrightarrow \mu_0$ or $\mu_u \nrightarrow \mu_0$.

When $P(\alpha_i)$ is bounded away from zero and one the bounds will converge at an exponential rate. We conjecture that an analogous result could be shown in the general case above. Having $P(\alpha_i) = 1$ with positive probability violates a condition of Theorem 6, that the conditional support of X_{it} equals the marginal support. Theorem 7 shows that in this case the bounds may not shrink to the marginal effect.

The bounds may converge, but not exponentially fast, depending on $P(\alpha_i)$ and the distribution of α_i. For example, suppose that $X_{it} = 1(\alpha_i - \varepsilon_{it} > 0)$, $\alpha_i \sim N(0,1)$, $\varepsilon_{it} \sim N(0,1)$, with ε_{it} i.i.d. over t and independent of α_i. Then

$$\mathcal{P}_K = E[\Phi(\alpha_i)^T] = \int \Phi(\alpha)^T \phi(\alpha) d\alpha = [\frac{\Phi(\alpha)^{T+1}}{T+1}]\Big|_{-\infty}^{+\infty} = \frac{1}{T+1}.$$

In this example the bounds will converge at the slow rate $1/T$. More generally, the convergence rate will depend on the distribution of $P(\alpha_i)$.

It is interesting to note that the convergence rates we have derived so far depend only on the properties of the joint distribution of (X_i, α_i), and not on the properties of the conditional distribution of Y_i given (X_i, α_i). This feature of the problem is consistent with us placing no restrictions on $m(x,\alpha)$. In the next Section we find that the bounds and rates may be improved when the conditional distribution of Y_i given (X_{it}, α_i) is restricted.

4 Semiparametric Multinomial Choice

The bounds for marginal effects derived in the previous section did not use any functional form restrictions on the conditional distribution of Y_i given (X_i, α). If this distribution is restricted one may be able to tighten the bounds. To illustrate we consider a semiparametric multinomial choice model where the conditional distribution of Y_i given (X_i, α_i) is specified and the conditional distribution of α_i given X_i is uknown.

We assume that the vector Y_i of outcome variables can take J possible values $Y^1, \ldots, Y^J$. As before, we also assume that X_i has a discrete distribution and can take K possible values $X^1, \ldots, X^K$. Suppose that the conditional probability of Y_i given (X_i, α_i) is

$$\Pr(Y_i = Y^j | X_i = X^k, \alpha_i) = \mathcal{L}(Y^j | X^k, \alpha_i, \beta^*)$$

for some finite dimensional β^* and some known function $\mathcal{L}(Y|X, \alpha, \beta)$. Let Q_k^* denote the unknown conditional distribution of α_i given $X_i = X^k$. Let $\mathcal{P}_{jk}$ denote the conditional probability of $Y_i = Y^j$ given $X_i = X^k$. We then have

$$\mathcal{P}_{jk} = \int \mathcal{L}\left(Y^j | X^k, \alpha, \beta^*\right) Q_k^*\left(d\alpha\right), (j = 1, ..., J; k = 1, ..., K),$$

where $\mathcal{P}_{jk}$ is identified from the data and the right hand side are the probabilities predicted by the model. This model is semiparametric in having a likelihood $\mathcal{L}\left(Y^j | X^k, \alpha, \beta\right)$ that is parametric and conditional distributions $Q_k\left(\alpha\right)$ for the individual effect that are completely unspecified. In general the parameters of the model may be set identified, so the previous equation is satisfied by a set of values B that includes β^* and a set of distributions for Q_k that includes Q_k^* for $k = 1, ..., K$. We discuss identification of model parameters more in detail in the next Section. Here we will focus on bounds for the marginal effect when this model holds.

For example consider a binary choice model where $Y_{it} \in \{0, 1\}$, $Y_{i1}, ..., Y_{iT}$ are independent conditional on (X_i, α_i), and

$$\Pr(Y_{it} = 1 | X_i, \alpha_i, \beta^*) = F(X_{it}'\beta^* + \alpha_i)$$

for a known CDF $F(\cdot)$. Then each Y^j consists of a $T \times 1$ vector of zeros and ones, so with $J = 2^T$ possible values. Also,

$$\mathcal{L}\left(Y | X, \alpha, \beta\right) = \prod_{t=1}^{T} F(X_t'\beta + \alpha)^{Y_t}[1 - F(X_t'\beta + \alpha)]^{1 - Y_t}.$$

The observed conditional probabilities then satisfy

$$\mathcal{P}_{jk} = \int \left\{ \prod_{t=1}^{T} F(X_t^{k\prime}\beta^* + \alpha)^{Y_t^j}[1 - F(X_t^{k\prime}\beta^* + \alpha)]^{1 - Y_t^j} \right\} Q_k^*\left(d\alpha\right), (j = 1, ..., 2^T; k = 1, ..., K).$$

As discussed above, for the binary choice model the marginal effect of a change in X_{it} from $\bar{x}$ to $\tilde{x}$, conditional on $X_i = X^k$, is

$$\mu_k = D^{-1} \int [F\left(\tilde{x}'\beta^* + \alpha\right) - F\left(\bar{x}'\beta^* + \alpha\right)] Q_k^*(d\alpha), \tag{8}$$

for a distance D. This marginal effect is generally not identified. Bounds can be constructed using the results of Section 3 with $B_\ell = 0$ and $B_u = 1$, since $m(x, \alpha) = F(x'\beta^* + \alpha) \in [0, 1]$. Moreover, in this model the sign of $\Delta(\alpha) = D^{-1}[F(\tilde{x}'\beta^* + \alpha) - F(\bar{x}'\beta^* + \alpha)]$ does not change with α_i, so we can apply the result in Lemma 4 to reduce the size of the bounds. These bounds, however, are not tight because they do not fully exploit the structure of the model. Sharper bounds are given by

$$\begin{aligned} \underline{\mu}_k = \quad & \min_{\beta \in B, Q_k} D^{-1} \int [F(\tilde{x}'\beta + \alpha) - F(\bar{x}'\beta + \alpha)] Q_k(d\alpha) \\ & \text{s.t. } \mathcal{P}_{jk} = \int \mathcal{L}\left(Y^j | X^k, \alpha, \beta\right) Q_k(d\alpha) \ \ \forall j, \end{aligned} \tag{9}$$

and

$$\begin{aligned} \overline{\mu}_k = \quad & \max_{\beta \in B, Q_k} D^{-1} \int [F(\tilde{x}'\beta + \alpha) - F(\bar{x}'\beta + \alpha)] Q_k(d\alpha) \\ & \text{s.t. } \mathcal{P}_{jk} = \int \mathcal{L}\left(Y^j | X^k, \alpha, \beta\right) Q_k(d\alpha) \ \ \forall j. \end{aligned} \tag{10}$$

In the next Sections we will discuss how these bounds can be computed and estimated. Here we will consider how fast the bounds shrink as T grows.

First, note that since this model is a special case of (more restricted than) the conditional mean model, the bounds here will be sharper than bounds previously given. Therefore, the bounds here will converge at least as fast as the previous bounds. Imposing the structure here does improve convergence rates. In some cases one can obtain fast rates without any restrictions on the joint distribution of X_i and α_i.

We will consider carefully the logit model and leave other models to future work. The logit model is simpler than others because β^* is point identified. In other cases one would need to account for the bounds for β^*. To keep the notation simple we focus on the binary X case, $X_{it} \in \{0, 1\}$, where $\tilde{x} = 1$ and $\bar{x} = 0$. We find that the bounds shrink at rate T^{-r} for any finite r, without any restriction on the joint distribution of X_i and α_i.

Theorem 8: *For $k = 1$ or $k = K$ and for any $r > 0$, as $T \longrightarrow \infty$,*

$$\overline{\mu}_k - \underline{\mu}_k = O(T^{-r}).$$

Fixed effects maximum likelihood estimators (FEMLEs) are a common approach to estimate model parameters and marginal effects in multinomial panel models. Here we compare the probability limit of these estimators to the identified sets for the corresponding parameters. The FEMLE treats the realizations of the individual effects as parameters to be estimated. The corresponding population problem can be expressed as

$$\tilde{\beta} = \operatorname{argmax}_\beta \sum_{k=1}^{K} \mathcal{P}_k \sum_{j=1}^{J} \mathcal{P}_{jk} \log \mathcal{L}\left(Y^j | X^k, \alpha_{jk}(\beta), \beta\right), \tag{11}$$

where

$$\alpha_{jk}(\beta) = \operatorname{argmax}_a \log \mathcal{L}\left(Y^j | X^k, \alpha, \beta\right), \forall j, k. \tag{12}$$

Here, we first concentrate out the support points of the conditional distributions of α and then solve for the parameter β.

Fixed effects estimation therefore imposes that the estimate of Q_k has no more than J points of support. The distributions implicitly estimated by FE take the form

$$\tilde{Q}_{k\beta}(\alpha) = \begin{cases} \mathcal{P}_{jk}, & \text{for } \alpha = \alpha_{jk}(\beta); \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The following example illustrates this point using a simple two period model.

Example 1 *Consider a two-period binary choice model with binary regressor and symmetric CDF, i.e., $F(-x) = 1 - F(x)$. In this case the estimand of the fixed effects estimators are*

$$\alpha_{jk}(\beta) = \begin{cases} -\infty, & \text{if } Y^j = (0,0); \\ -\beta(X_1^k + X_2^k)/2, & \text{if } Y^j = (1,0) \text{ or } Y^j = (0,1); \\ \infty, & \text{if } Y^j = (1,1), \end{cases} \tag{14}$$

and the corresponding distribution for α has the form

$$\tilde{Q}_{k\beta}(\alpha) = \begin{cases} \Pr\{Y = (0,0)|X^k\}, & \text{if } \alpha = -\infty; \\ \Pr\{Y = (1,0)|X^k\} + \Pr\{Y = (0,1)|X^k\}, & \text{if } \alpha = -\beta(X_1^k + X_2^k)/2; \\ \Pr\{Y = (1,1)|X^k\}, & \text{if } \alpha = \infty. \end{cases} \tag{15}$$

This formulation of the problem is convenient to analyze the properties of the fixed effects estimators of marginal effects. Thus, for example, the fixed effects estimator of the marginal effect μ_k takes the form:

$$\tilde{\mu}_k(\beta) = D^{-1} \int [F(\tilde{x}'\beta + \alpha) - F(\bar{x}'\beta + \alpha)]\tilde{Q}_{k\beta}(\alpha). \tag{16}$$

This estimator is consistent for identified effects when X is binary for any symmetric CDF. This result is shown here analytically for the two-period case and through numerical examples for $T \geq 3$.

Theorem 9: *If $k \notin \{1, K\}$ and $F(-x) = 1 - F(x)$, then $\tilde{\mu}_k(\tilde{\beta}) \xrightarrow{p} \mu_k$.*

For not identified effects the fixed effects estimators are usually biased toward zero. To see this consider a logit model with binary regressor, $X^k = (0,0)$, $\bar{x} = 0$ and $\tilde{x} = 1$. Using that $\tilde{\beta} = 2\beta^*$ (Andersen, 1973) and $F'(x) = F(x)(1 - F(x)) \leq 1/4$, we have

$$\begin{aligned} \left|\tilde{\mu}_k(\tilde{\beta})\right| &= \left|F(\tilde{\beta}) - F(0)\right| [P(Y = (1,0)|X^k) + P(Y = (0,1)|X^k)] \\ &\leq \left|\tilde{\beta}/2\right| \int F(\alpha)(1-\alpha)Q_k(d\alpha) = \left|E[\beta^* F'(\bar{x}\beta^* + \alpha)|X = X^k]\right| \approx |\mu_k|. \end{aligned}$$

This conjecture is further explored numerically in the next section.

5 Calculating Population Bounds

We will begin our discussion of calculating bounds by considering bounds for the parameter β. Letting $Q \equiv (Q_1, \ldots, Q_K)$, we can write the individual log likelihood compactly as $L(Y_i, X_i; \beta, Q)$. Due to the usual argument based on Jensen's inequality, we can see that (β^*, Q^*) is such that

$$E[L(Y_i, X_i; \beta, Q)] \leq E[L(Y_i, X_i; \beta^*, Q^*)]$$

for every (β, Q). This implies that

$$\sup_Q E[L(Y_i, X_i; \beta, Q)] \leq \sup_Q E[L(Y_i, X_i; \beta^*, Q)]$$

for every β. Therefore, if we define B to be the set of β's that maximizes $\sup_Q E[L(Y_{i1}, Y_{i2}; \beta, Q)]$, i.e.,

$$B \equiv \left\{ \beta : \sup_Q E[L(Y_i, X_i; \beta, Q)] \geq \sup_Q E\left[L\left(Y_i, X_i; \beta', Q\right)\right], \quad \forall \beta' \right\}.$$

We can easily see that $\beta^* \in B$. In other words, β^* is set identified by the set B.

It follows from results of Lindsay (1995) that one need only search over discrete distributions for Q to find B. Note that

Lemma 10: *If the support $\mathbb{C}$ of α_i is compact and $\mathcal{L}\left(Y^j | X^k, \alpha, \beta\right)$ is continuous in α for each β, j, and k, then, for each $\beta \in B$ and k, a solution to*

$$\bar{Q}_{k\beta} = \arg\max_{Q_k} E[\ln \int \mathcal{L}\left(Y^j | X^k, \alpha, \beta\right) Q_k(d\alpha) | X_i = X^k]$$

exists that is a discrete distribution with at most J points of support, and $\int \mathcal{L}\left(Y^j | X^k, \alpha, \beta\right) \bar{Q}_{k\beta}(d\alpha) = \mathcal{P}_{jk}$, $\forall j, k$.

It is also true that bounds for the marginal effect can be found by searching over discrete distributions. We will focus on the upper bound $\overline{\mu}_k$; an analogous result holds for the lower bound $\underline{\mu}_k$.

Lemma 11: *If the support $\mathbb{C}$ of α_i is compact and $\mathcal{L}\left(Y^j | X^k, \alpha, \beta\right)$ is continuous in α for each β, j, and k, then, for each $\beta \in B$ and k, a solution to*

$$\bar{Q}_{k\beta} = \arg\max_{Q_k} D^{-1} \int [F(\tilde{x}'\beta + \alpha) - F(\bar{x}'\beta + \alpha)] Q_k(d\alpha) \text{ s.t. } \int \mathcal{L}\left(Y^j | X^k, \alpha, \beta\right) Q_k(d\alpha) = \mathcal{P}_{jk}$$

can be obtained from a discrete distribution with at most J points of support.

We carry out some numerical calculations to illustrate and complement the previous analytical results. We use the following binary choice model

$$Y_{it} = \mathbf{1}\{X_{it}\beta + \alpha_i + \varepsilon_{it} \geq 0\}, \tag{17}$$

with ε_{it} i.i.d. normal or logistic with zero mean and unit variance. The explanatory variable X_{it} is binary, independent across time periods with $p_X = \Pr\{X_{it} = 1\} = 0.5$. The unobserved individual effects α_i is correlated with the explanatory variable for each individual. In particular, we generate these effects as a mixture of a random component and the standardized individual sample mean of the regressor. The random part is independent of the regressors and follows a discretized standard normal distribution, as in Honoré and Tamer (2006). Thus, we have

$$\alpha_i = \alpha_{1i} + \alpha_{2i},$$

where

$$\Pr\{\alpha_{1i} = a_m\} = \begin{cases} \Phi\left(\frac{a_{m+1}+a_m}{2}\right), & \text{for } a_m = -3.0; \\ \Phi\left(\frac{a_{m+1}+a_m}{2}\right) - \Phi\left(\frac{a_m+a_{m-1}}{2}\right), & \text{for } a_m = -2.8, -2.6, ..., 2.8; \\ 1 - \Phi\left(\frac{a_m+a_{m-1}}{2}\right), & \text{for } a_m = 3.0. \end{cases}$$

and $\alpha_{2i} = \sqrt{T}(\bar{X} - p_X)/\sqrt{p_X(1-p_X)}$.

Identified sets for parameters and marginal effects are calculated for panels with 2, 3, and 4 periods based on the general conditional expectation model and semiparametric logit and probit models. For logit and probit models the sets are obtained using the linear programming algorithm of Honoré and Tamer (2006) for discrete regressors. Thus, for the parameter we have that $B = \{\beta : L(\beta) = 0\}$,[1] where

$$\begin{aligned} L(\beta) \quad = \quad & \min_{w_k, v_{jk}, \pi_{km}} \sum_{k=1}^{K} w_k + \sum_{j=1}^{J}\sum_{k=1}^{K} v_{jk} && (18) \\ & v_{jk} + \textstyle\sum_{m=1}^{M} \pi_{km}\mathcal{L}\left(Y^j|X^k, \alpha_m, \beta\right) = \mathcal{P}_{jk} && \forall j,k, \\ & w_k + \textstyle\sum_{m=1}^{M} \pi_{km} = 1 && \forall k, \\ & v_{jk} \geq 0, w_k \geq 0, \pi_{km} \geq 0 && \forall j,k,m. \end{aligned}$$

For marginal effects, we solve

$$\begin{aligned} \overline{\mu}_k/\underline{\mu}_k \quad = \quad & \max/\min_{\pi_{km}, \beta\in B} \sum_{m=1}^{M} \pi_{km}[F(\beta + \alpha_m) - F(\alpha_m)] \\ & \textstyle\sum_{m=1}^{M} \pi_{km}\mathcal{L}\left(Y^j|X^k, \alpha_m, \beta\right) = \mathcal{P}_{jk} \quad \forall j, \\ & \textstyle\sum_{m=1}^{M} \pi_{km} = 1, \pi_{km} \geq 0 \quad \forall j,m. \end{aligned}$$

The identified sets are also compared to the probability limits of linear and nonlinear fixed effects estimators.

[1] In calculating the identified sets, we search over a wide grid of support points for the mixing distribution that contains the points of support of α_i. In many cases the estimate of mixing distribution has points of support outside the range of true points of support of the true distribution.

Figure 2 shows identified sets for the index coefficient β in the logit model. The figures agree with the well-known result that the model parameter is point identified when $T \geq 2$, e.g., Andersen (1973). The fixed effect estimator is inconsistent and has a probability limit that is biased away from zero. For example, for $T = 2$ it coincides with the value $2\beta^*$ obtained by Andersen (1973). For $T > 2$, the proportionality $\tilde{\beta} = c\beta_0$ for some constant c breaks down.

Identified sets for marginal effects are plotted in Figures 3 - 7, together with the probability limits of fixed effects maximum likelihood estimators (Figures 4 - 6) and linear probability model estimators (Figure 7).[2] Figure 3 shows identified sets based on the general conditional mean model. The bounds of these sets are obtained using the general bounds (G-bound) for binary regressors in (7), and imposing the monotonicity restriction $\Delta(\alpha) > 0$ in Lemma 4 (GM-bound). In this example the monotonicity restriction has important identification content in reducing the size of the bounds.

Figures 4 - 6 show that marginal effects are point identified for individuals with switches in the value of the regressor, and fixed effects estimators are consistent for these effects. This numerical finding suggests that the consistency result for fixed effects estimators extends to more than two periods. Marginal effects for individuals without switches in the regressor are not point identified, unless $\beta^* = 0$, which also precludes point identification of the average effects. Fixed effects estimators are biased toward zero for the unidentified effects, and have probability limits that usually lie outside of the identified set. However, both the size of the identified sets and the asymptotic biases of the fixed effects estimators shrink very fast with the number of time periods. In Figure 7 we see that linear probability model estimators have probability limits that usually fall outside the identified set for the marginal effect.

For the probit, Figure 8 shows that the model parameter is not point identified, but the size of the identified set shrinks very fast with the number of time periods. The identified sets and limits of fixed effects estimators in Figures 9 - 13 are analogous to the results for logit.

6 Estimation

The population problem for the parameter β has the convenient linear programming formulation (18) when the regressors are discrete. To estimate B, Honoré and Tamer (2006) suggest solving the linear programming problem replacing the conditional probabilities $\mathcal{P}_{jk}$ by consistent sample

[2] We consider the version of the linear probability model that allows for individual specific slopes in addition to the fixed effects.

estimates P_{jk}. The proposed estimate of the identified set is $B_n = \{\beta : \hat{L}(\beta) = 0\}$ where

$$\begin{aligned}
\hat{L}(\beta) \quad = \quad & \min_{w_k, v_{jk}, \pi_{km}} \sum_{k=1}^{K} w_k + \sum_{j=1}^{J} \sum_{k=1}^{K} v_{jk} \\
& v_{jk} + \textstyle\sum_{m=1}^{M} \pi_{km} \mathcal{L}\left(Y^j | X^k, \alpha_m, \beta\right) = P_{jk} \quad \forall j, k, \\
& w_k + \textstyle\sum_{m=1}^{M} \pi_{km} = 1 \quad \forall k, \\
& v_{jk} \geq 0, w_k \geq 0, \pi_{km} \geq 0 \quad \forall j, k, m.
\end{aligned}$$

Here P_{jk} are the observed probabilities in the sample.

There are two important practical difficulties in the implementation of this approach for estimation. First, the solution for Q to the linear programming problem is very sensitive to the presence of empty cells, that is, when Y^j is not observed for some X^k. Then $P_{jk} = 0$ and $\hat{Q}_k$ is a degenerate distribution at $-\infty$. This issue is an artifact of the way the restrictions are formulated, which only allows for negative differences between model and true probabilities, together with the properties of the common specifications for the model probabilities, such as logit of probit, which only are zero at $-\infty$. We introduce a variation of a minimum distance procedure proposed by Honoré and Tamer (2006) that is less sensitive to the empty cell problem.

A second important drawback of the linear programming formulation, also shared by the minimum distance procedure, is that the solution for the identified set is generally an empty set because the minimum of the objective function is always positive. The source of this problem is sampling error in the estimated probabilities and model misspecification. This problem can be addressed by choosing B_n as the set of values of β for which the minimized objective function $\hat{L}(\beta)$ attains its minimum (instead of zero) up to a cut-off parameter. We apply this solution to the objective function of the minimum distance problem.

The modified minimum distance estimator that we propose is the solution to the following penalized weighted quadratic programming problem:

$$B_n = \left\{ \beta : \hat{T}(\beta) \leq \min_{\beta \in \mathbb{B}} \hat{T}(\beta) + \epsilon_n \right\}, \tag{19}$$

where $\epsilon_n \geq 0$ is a cut-off parameter that shrinks to zero as a function of the sample size as in Manski and Tamer (2002); and

$$\begin{aligned}
\hat{T}(\beta) \quad = \quad & \min_{z_{km}} \sum_{j,k} \left[\omega_{jk} \left(P_{jk} - \sum_{m=1}^{M} \pi_{km} \mathcal{L}\left(Y^j | X^k, \alpha_m, \beta \right) \right)^2 + \lambda_n \sum_{m=1}^{M} z_{km}^2 \right], \qquad (20) \\
& \text{s.t.} \quad \sum_{m=1}^{m} \pi_{km} = 1, \ \pi_{km} \geq 0, \ \forall j, k.
\end{aligned}$$

This formulation is less sensitive to the empty cell problem because it allows for positive and negative differences between model and observed probabilities. The weights ω_{jk} are chosen in order to have a chi-square type objective function and to increase the efficiency of the estimator by weighting more the sequences of X with higher sample frequencies. In particular, we set

$$\omega_{jk} = nP_k / \sum_{m=1}^{M} \tilde{\pi}_{km} \mathcal{L}\left(Y^j | X^k, \alpha_m, \tilde{\beta}\right),$$

where P_k is the observed probability of the sequence X^k in the sample and $(\tilde{\beta}, \{\tilde{\pi}_{km} : k = 1, ..., K; m = 1, ..., M\})$ are preliminary estimates of the parameters. These estimates can be obtained by setting $\omega_{jk} = nP_k$.

The penalty λ_n acts choosing a distribution among the set of discrete distributions with support contained in $\{\alpha_1, ..., \alpha_M\}$. This regularization solves the fundamental identification problem for Q_k, while keeping the computationally convenient quadratic programming formulation. In general there is an infinite number of solutions for Q_k to the population problem, one of them is a discrete distribution with no more than J support points by Lemma 10. Here, instead of searching for the solution with the minimal support, we search over discrete distributions with support points contained in a large partition of an interval of the real line. By making the partition fine enough we guarantee to cover the solutions to the problem with few support points, without having to find explicitly the location of those points.[3] The penalty favors distributions with large supports. Moreover, by setting $\lambda_n = o(1)$, the penalty does not affect the distribution of the objective function in large samples.

The solution to the penalized minimum distance problem cannot be directly used to obtain estimates for the marginal effects. The restrictions of the linear programs for these effects generally cannot be satisfied for any $\beta \in B_n$ due to sampling variation in the estimated probabilities and/or model misspecification. To make the problem feasible we replace the estimates of the conditional probabilities by the probabilities predicted by the model at the solution to the quadratic problem. These probabilities are consistent if the model is correctly specified, and equal to the probabilities predicted by the model at the solution to the quadratic problem by construction. To simplify the computation it is useful to note that we only need to solve the linear programming problem for the marginal effects that are not identified. For identified effects, we can use sample analogs of the results in Lemma 4 based on the recentered probabilities.

Another way to estimate B is by the the level set of the finite-sample profile likelihood

$$B_n = \left\{ \beta : \sup_{Q} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_i, X_i; \beta, Q\right) \geq \sup_{\beta} \sup_{Q} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_i, X_i; \beta, Q\right) - \epsilon_n \right\},$$

where $\epsilon_n > 0$ is a cut-off parameter that shrinks to zero as a function of the sample size, following Manski and Tamer (2002). Estimators for the bounds of the marginal effects defined above can

[3] Finding the explicit location of the support points is the main computational difficulty in the estimation of distribution of mixtures; see, e.g., Aitkin (1999).

be obtained by solving these problems with B_n in place of B.

Following Chernozhukov, Hahn, and Newey (2004) we can show consistency of this estimator under two conditions.

Assumption 1: (i) $\mathcal{L}\left(Y^j|X^k,\alpha,\beta\right)$ is continuous in (α,β) for all (j,k); (ii) $\beta^* \in \mathbb{B}$ for some compact $\mathbb{B}$; and (iii) α_i has a support contained in a compact set $\mathbb{C}$.

The other condition concerns the cut-off parameter.

Assumption 2: If B is a singleton, $B = \{\beta^*\}$, then $\epsilon_n = 0$. Otherwise, $\epsilon_n \propto n^{-1/2}a_n$ for some $a_n \to \infty$ and $n^{-1/2}a_n \to 0$.

We can now give a consistency result

Theorem 12: *If Assumptions 1 and 2 are satisfied*

$$d_H\left(B_n, B\right) = o_p(1),$$

where d_H is the Hausdorff distance between sets

$$d_H\left(B_n, B\right) = \max\left[\sup_{b_n\in B_n}\inf_{b\in B}|b_n - b|\,, \sup_{b\in B}\inf_{b_n\in B_n}|b_n - b|\right]$$

We can obtain a corresponding result for the marginal effect.

Corollary 13: *Let $\hat{\underline{\mu}}_k$ and $\hat{\overline{\mu}}_k$ denote the solutions to the programs (9) and (10) when B is replaced by B_n. If Assumptions 1 and 2 are satisfied then*

$$\hat{\underline{\mu}}_k \xrightarrow{p} \underline{\mu}_k \text{ and } \hat{\overline{\mu}}_k \xrightarrow{p} \overline{\mu}_k$$

7 Inference

Theorem 12 does not provide any practical guidance on the choice of the cut-off level ϵ_n. It is desirable that this choice be tied to inferential statements, which appear to pose special challenges in this setting. In this Section we propose to base inference on the inversion of the objective function of the quadratic program, embedding the previous semi-parametric likelihood in a more general nonparametric family. This approach provides conservative inferences about β and marginal effects.

From the proof of Theorem 12, it follows that the model-implied probabilities coincide with the true choice probabilities for any $\beta^* \in B$ and some (generally non-unique) pseudo-true Q^*:

$$\mathcal{P}_{jk} = \int_{\mathbb{C}} \mathcal{L}\left(Y^j|X^k,\alpha,\beta^*\right)Q_k^*(d\alpha) := \mathcal{L}_{jk}(\beta^*, Q_k^*), \forall j, k.$$

Let P_{jk} be the empirical probabilities. A chi-square type statistic evaluated at (β, Q) takes the form

$$T(\beta, Q) = n \sum_{j,k} P_k \frac{\left(P_{jk} - \mathcal{L}_{jk}(\beta, Q_k)\right)^2}{\mathcal{L}_{jk}(\beta, Q_k)}.$$

The quantity of especial interest is the profile statistic:

$$T(\beta) = n \sum_{j,k} P_k \frac{\left(P_{jk} - \mathcal{L}_{jk}(\beta, \hat{Q}_{k\beta})\right)^2}{\mathcal{L}_{jk}(\beta, \hat{Q}_{k\beta})}, \tag{21}$$

where $\hat{Q}_{k\beta}$ is the solution to the quadratic program (20) with the model parameter fixed to β. Since $\lambda_n \xrightarrow{p} 0$ as $n \to \infty$, with probability approaching to one $T(\beta) \leq T(\beta, Q)$ and the α-quantile of $T(\beta)$ is bounded from above by

$$c_\alpha(\beta) = \inf_c \left\{c : \Pr\{T(\beta, Q) \leq c\} \geq \alpha\right\}.$$

A conservative confidence interval for β^* is then given by

$$I_\alpha(\beta^*) = \{\beta : T(\beta) \leq c_\alpha(\beta)\}.$$

The upper bound of the quantile $c_\alpha(\beta)$ is asymptotically pivotal by the classical Pearson's argument $T(\beta^*, Q^*) \Rightarrow \chi^2(K(J-1))$, hence we have that $c_\alpha(\beta)$ can be consistently estimated by the $\alpha-$ quantile of a $\chi^2(K(J-1))$ variable, denoted as $\hat{c}_\alpha$. An approximate confidence region is then given by

$$\hat{I}_\alpha(\beta^*) = \{\beta : T(\beta) \leq \hat{c}_\alpha\}.$$

The preceding argument established the following result.

Theorem 14: *If Assumption 1 is satisfied then*

$$P\{\beta^* \in \hat{I}_\alpha(\beta^*)\} \to \bar{\alpha} \geq \alpha$$

as $n \to \infty$.

Theorem 14 also leads to a more precise choice of the cut-off level needed to insure consistent estimation in the previous section. One such choice is given by

$$\epsilon_n = \hat{c}_{\alpha_n} - \min_{\beta \in \mathbb{B}} T(\beta),$$

where the significance level α_n should tend to 1 such that the α_n -th quantile of $\chi^2(K(J-1))$ variable satisfies Assumption 2 as $n \to \infty$ slowly enough. This choice guarantees the estimating set B_n coincides with the desired confidence region of probability level α_n. In practice, α_n may be set equal to some conventional value such as .90 or .95.

Confidence regions for marginal effects can be formed as the union of the solutions to the linear programming problem for these effects for the values of the parameter in the confidence interval $\hat{I}_{\alpha}(\beta^{*})$. Computation can be greatly simplified if the marginal effects are monotone on the value of the parameter. In this case, which includes logit and probit models, the linear programs for the effects need only to be solved for values at the boundary of the confidence region for the parameter. The resulting confidence regions have coverage probability at least α in large samples by the continuous mapping theorem.

The previous projection method is computationally attractive because it typically involves repeating the two step estimation procedure only a few times, but it shares the problems common to objective function based inference procedures. In particular, the method can be conservative if the degree of over-identification of the model is high. Overidentification here is the difference between the dimension of the parameter and the degrees of freedom of the chi-square distribution (number of free probabilities), what determines the excess of degrees of freedom used above what is needed to test hypotheses about the parameter. More importantly, these procedures are very sensitive to model misspecification since the objective function increases with the difference between the true probabilities and the best approximating model probabilities. If the degree of misspecification is high enough the procedure can actually produce empty confidence regions. The reason is that the objective function-based tests are in fact omnibus tests for both model specification and the value of the parameters. The degree of overidentification has therefore two opposite effects on the confidence regions as it increases the size by raising the number of degrees of freedom of the test statistics, but also makes model misspecification more acute as the total number of free probabilities to fit becomes larger.

7.1 Bootstrap

An alternative to objective function inversion methods to make inference on the identified sets of interest is to use resampling techniques. If the outcome and regressors are discrete, nonparametric bootstrap corresponds to parametric bootstrap on the bivariate multinomial distribution for all the sequences of outcomes and regressors. Thus, we can construct bootstrap confidence regions for the identified sets of the parameters and marginal effects using the following procedure:

1. Draw bootstrap a dataset $\{X_i^{(r)}, Y_i^{(r)}\}_{i=1}^n$ from the observed bivariate multinomial frequencies $\{X_i, Y_i\}_{i=1}^n$.
2. Estimate the identified sets for the parameter $B_n^{(r)}$ and the corresponding marginal effects $[\hat{\underline{\mu}}_k^{(r)}, \hat{\overline{\mu}}_k^{(r)}]$ by solving the nonparametric MLE quadratic and linear programming problems.

3. Repeat the procedure R times.

4. Construct the α-level confidence regions as the smallest sets that fully contain a proportion α of the estimated regions for the parameters $\{B_n^{(r)}\}_{r=1}^R$ and marginal effects $\{[\underline{\hat{\mu}}_k^{(r)}, \hat{\overline{\mu}}_k^{(r)}]\}_{r=1}^R$.

This nonparametric bootstrap procedure is less sensitive to model misspecification since it does not impose the conditional model on the bootstrap data generating process (DGP). The confidence regions can therefore be interpreted as confidence regions for the best approximating model to the DGP. However, an important issue here is to show the consistency of bootstrap for the distribution of the estimators. The estimators of the model parameters and marginal effect are non regular and it is not clear if their distributions vary with perturbations of the DGP in a continuous way. We are not aware of any result on bootstrap validity for this problem or the related problem of estimation of finite mixture models.[4]

7.2 Perturbed Bootstrap

Dufour (2006) develops simulation methods to conduct inference in non regular cases where the estimators of the parameters of interest might have asymptotic distributions that depend on nuisance parameters in a discontinuous way, or even when they do not converge in distribution, see also Romano and Wolf (2000). These methods do not rely on point identification of the parameter of interest and can therefore be applied to set-identified models, see, e.g., Rytchkov (2006). The idea of this approach is to generate a class of distributions that covers the true DGP with probability one, and find the least favorable distribution for the estimators of interest within this class. The quantiles of this distribution can be used to construct confidence regions for the identified sets. We implement this method by a variation of the bootstrap described below that we denominate as *perturbed bootstrap* (Chernozhukov, 2007).

To describe how this method works, consider the general problem of making inference on a parameter θ based on a sample statistic T_n with distribution $G_n(t, F)$ under the DGP $F \in \mathcal{F}$. The set $\mathcal{F}$ is a class of distribution functions restricted to have compact support. The goal is to estimate the distribution of the statistic under the true F_0, i.e., to find $G_n(t, F_0)$. The method proceeds by constructing a confidence region $CR_{1-\gamma_n}(F_0)$ that contains the true DGP F_0 with probability approaching to one, i.e., $\gamma_n \to 0$, and such that, as $n \to \infty$,

$$d(CR_{1-\gamma_n}(F_0), F_0) := \inf_{F \in CR_{1-\gamma_n}(F_0)} d_K(F, F_0) \xrightarrow{p} 0, \tag{22}$$

[4]Feng and McCulloch (1996) conjecture the validity of bootstrap for the distribution of the likelihood ratio test for the number of components of the mixture distribution and provide some numerical evidence. See also the monograph on finite mixture models by McLachlan and Peel (2000).

where d_K is the sup (Kolmogorov) distance defined by $d_K(F,G) := \sup_t |F(t) - G(t)|$. The least favorable distributions for $G_n(t,F_0)$ are given by

$$\underline{G}_n(t,F_0)/\overline{G}_n(t,F_0) = \inf / \sup_{F \in CR_{\gamma_n}(F_0)} G_n(t,F). \tag{23}$$

Romano and Wolf (2000) show that the $(\alpha - \gamma_n)/2$ quantile of $\underline{G}_n(t,F_0)$ and the $1-(\alpha-\gamma_n)/2$ quantile of $\overline{G}_n(t,F_0)$ can be used to form valid confidence regions of level $1-\alpha$. Moreover, if the test statistic is efficient for the parameter, then these confidence regions are as efficient asymptotically as the confidence regions that use the true sampling distribution $G_n(t,F_0)$ provided that $d_K(\underline{G}_n(t,F_0), G_n(t,F_0)) \xrightarrow{p} 0$ and $d_K(\overline{G}_n(t,F_0), G_n(t,F_0)) \xrightarrow{p} 0$.

For panel data models with discrete outcomes and regressors, this inference approach can be implemented using this procedure (perturbed bootstrap):

1. Draw a potential DGP from the observed bivariate multinomial obtained from $\{X_i, Y_i\}_{i=1}^n$.

2. Test that the observed sample is consistent with the potential DGP with high probability. This step can be carried out by checking that the observed dataset passes a chi-square test with small level γ_n (e.g., set $\gamma_n = .01$). Note that since we are not imposing the conditional model the chi-square distribution has $JK - 1$ degrees of freedom under the hypothesis that the observed distribution comes from the potential DGP.

3. Repeat steps 1 and 2 until a DGP, DGP_p, passes the test.

4. Estimate the distribution of the estimator by nonparametric bootstrap from DGP_p (see the previous subsection for details on implementation).

5. Repeat the steps (1) to (4) for $p = 1, ..., P$.

6. Obtain
$$\hat{\underline{G}}(t,F_0)/\hat{\overline{G}}(t,F_0) = \min / \max\{\hat{G}(t, DGP_1), ..., \hat{G}(t, DGP_P)\}.$$

7. Construct a $1-\alpha$ confidence region for the parameter of interest as
$$CR_\alpha(\theta) = \{\underline{\theta}, \overline{\theta}\}$$
where $\underline{\theta}$ is the $(\alpha - \gamma_n)/2$ quantile of $\hat{\underline{G}}(t,F_0)$ and $\overline{\theta}$ is the $1-(\alpha-\gamma_n)/2$ quantile of $\hat{\overline{G}}(t,F_0)$.

8 Empirical Example

We now turn to an an empirical application of our methods to a binary choice panel model of female labor force participation. It is based on a sample of married women in the National Longitudinal Survey of Youth 1979 (NLSY79). We focus on the relationship between participation and the presence of young children in the years 1990, 1992, 1994, and 1996. The NLSY79 data set is convenient to apply our methods because it provides a relatively homogenous sample of women between 25 and 33 year-old in 1990, what reduces the extent of other potential confounding factors that may affect the participation decision, such as the age profile, and that are difficult to incorporate in our methods. Other studies that estimate similar models of participation in panel data include Heckman and MaCurdy (1980), Heckman and MaCurdy (1982), Chamberlain (1984), Hyslop (1999), Chay and Hyslop (2000), Carrasco (2001), Carro (2007), and Fernández-Val (2008).

The sample consists of 1,587 married women. Only women continuously married, not students or in the active forces, and with complete information on the relevant variables in the entire sample period are selected from the survey. Descriptive statistics for the sample are shown in Table 2. The labor force participation variable (LFP) is an indicator that takes the value one if the woman employment status is "in the labor force" according to the CPS definition, and zero otherwise. The fertility variable ($kids$) indicates whether the woman has any child less than 3 year-old. We focus on very young preschool children as most empirical studies find that their presence have the strongest impact on the mother participation decision. LFP is stable across the years considered, whereas $kids$ initially increases to peak in 1994 and drops sharply in the last year of the sample. The proportion of women that change fertility status grows steadily with the number of time periods of the panel, but there are still 40% of the women in the sample for which the effect of fertility is not identified after 4 periods.

The empirical specification we use is similar to Chamberlain (1984). In particular, we estimate the following equation

$$LFP_{it} = \mathbf{1}\left\{\beta \cdot kids_{it} + \alpha_i + \epsilon_{it} \geq 0\right\}, \tag{24}$$

where α_i is an individual specific effect. The parameters of interest are the marginal effects of fertility on participation for different groups of individuals including the entire population. These effects are estimated using the general conditional expectation model and semiparametric logit and probit models described in Sections 3 and 4, together with linear and nonlinear fixed effects estimators. Analytical and Jackknife large-T bias corrections are also considered, and conditional fixed effects estimates are reported for the logit model.[5] The estimates from the

[5]The analytical corrections use the estimators of the bias based on expected quantities in Fernández-Val (2008).

general model impose monotonicity of the effects. For the semiparametric estimators, we choose a penalty $\lambda_n = 1/\ln n$ and iterate the quadratic program 3 times, what makes the estimates insensitive to the penalty and the weighting. We search over discrete distributions with 23 support points at $\{-\infty, -4, -3.6, ..., 3.6, 4, \infty\}$ in the quadratic problem, and with 163 support points at $\{-\infty, -8, -7.9, ..., 7.9, 8, \infty\}$ in the linear programming problems. The estimates are based on panels of 2, 3, and 4 time periods, all of them starting in 1990.

Tables 3 to 5 report estimates of the model parameters and marginal effects for 2, 3, and 4 period panels, together with 95% confidence regions obtained using the procedures described in the previous Section. For the general model these regions are constructed using the normal approximation ($N - CI$) and nonparametric bootstrap with 200 repetitions ($B - CI$). For the logit and probit models, the confidence regions are obtained by inversion of the objective function or projection method ($P - CI$), nonparametric bootstrap with 200 repetitions ($B - CI$), and perturbed bootstrap ($PB - CI$) with $\beta_n = .01$, 100 DGP's, and 200 bootstrap repetitions for each DGP. For the fixed effects estimators, the confidence regions are based on the asymptotic normal approximation. The semiparametric estimates are shown for $\epsilon_n = 0$, which is for the solution that gives the minimum value in the quadratic problem.[6]

Overall, we find that the estimates and confidence regions based on the general model are too wide to provide informative evidence about the relationship between participation and fertility for the entire population. The semiparametric estimates seem to offer a good compromise between producing more accurate results without adding too much structure to the model. Thus, these estimates are always inside the confidence regions of the general model and do not suffer of important efficiency losses relative to the more restrictive fixed effects estimates. Another salient feature of the results is that the misspecification problem of the projection method clearly shows up in this application. Thus, this procedure gives empty confidence regions for panels of 3 and 4 periods. Note that in this case, where we only have one parameter and binary outcome and regressor, the degree of over-identification is 11, 55, and 239 for the 2, 3, and 4 period panels, respectively.

9 Possible Extensions

Our analysis is yet confined to models with only discrete explanatory variables. It would be interesting to extend the analysis to models with continuous explanatory variables. It may be possible to come up with a sieve-type modification. We expect to obtain a consistent estimator

The Jackknife bias correction uses the procedure described in Hahn and Newey (2004).

[6]For the logit model the parameter β is identified and this choice of ϵ_n is justified by Theorem 12. For the probit model the reported estimate is only guaranteed to be contained in the identified set with probability approaching to one.

of the bound by applying the semiparametric method combined with increasing number of partitions of the support of the explanatory variables, but we do not yet have any proof. Empirical likelihood based methods should work in a straightforward manner if the panel model of interest is characterized by a set of moment restrictions instead of a likelihood. We may be able to improve the finite-sample property of our confidence region by using Bartlett type corrections.

10 Appendix: Proofs

Proof of Theorem 1: By eq. (3),

$$\begin{aligned}\sum_t (X_t^k - r^k)E[Y_{it}|X_i = X^k] &= Tr^k(1-r^k)\int m(1,\alpha)Q_k^*(d\alpha) \\ &\quad + T(1-r^k)(-r^k)\int m(0,\alpha)Q_k^*(d\alpha) = T\sigma_k^2\mu_k. \end{aligned} \tag{25}$$

Note also that $\bar{X}_i = r^k$ when $X_i = X^k$. Then by the law of large numbers,

$$\begin{aligned}\sum_{i,t}(X_{it}-\bar{X}_i)^2/n \xrightarrow{p} E[\sum_t (X_{it}-\bar{X}_i)^2] &= \sum_k \mathcal{P}_k \sum_t (X_t^k - r^k)^2 = \sum_k \mathcal{P}_k T\sigma_k^2, \\ \sum_{i,t}(X_{it}-\bar{X}_i)Y_{it}/n \xrightarrow{p} E[\sum_t (X_{it}-\bar{X}_i)Y_{it}] &= \sum_k \mathcal{P}_k \sum_t (X_t^k - r^k)E[Y_{it}|X_i = X^k] \\ &= \sum_k \mathcal{P}_k T\sigma_k^2\mu_k. \end{aligned}$$

Dividing and applying the continuous mapping theorem gives the result. Q.E.D.

Proof of Theorem 2: Note that $\sum_{t=1}^T (X_t^k - \bar{X}^k)^2 = Tr^k(1-r^k) = T\sigma_k^2 > 0$ for all $2 \leq k \leq K-1$, so by eq. (25) and the law of large numbers,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n 1(s_{xi}^2 &> 0)\frac{\sum_{t=1}^T (X_{it}-\bar{X}_i)Y_{it}}{s_{xi}^2} \xrightarrow{p} E[1(s_{xi}^2>0)\frac{\sum_{t=1}^T (X_{it}-\bar{X}_i)Y_{it}}{s_{xi}^2}] \\ &= E[1(s_{xi}^2>0)\frac{\sum_{t=1}^T (X_{it}-\bar{X}_i)E[Y_{it}|X_i]}{s_{xi}^2}] = \sum_{k=2}^{K-1}\mathcal{P}_k \frac{\sum_{t=1}^T (X_t^k - r^k)E[Y_{it}|X_i = X^k]}{T\sigma_k^2} \\ &= \sum_{k=2}^{K-1}\mathcal{P}_k \frac{T\sigma_k^2\mu_k}{T\sigma_k^2} = \sum_{k=2}^{K-1}\mathcal{P}_k\mu_k, \\ \frac{1}{n}\sum_{i=1}^n 1(s_{xi}^2 &> 0) \xrightarrow{p} E[1(s_{xi}^2>0)] = \sum_{k=2}^{K-1}\mathcal{P}_k. \end{aligned}$$

Dividing and applying the continuous mapping theorem gives the result. Q.E.D.

Proof of Theorem 3: The set of X_i where $\tilde{r}_i > 0$ and $\bar{r}_i > 0$ coincides with the set for which $X_i = X^k$ for $k \in \mathcal{K}^*$. On this set it will be the case that $\tilde{r}_i$ and $\bar{r}_i$ are bounded away from zero.

Note also that for $\tilde{t}$ such that $X^k_{\tilde{t}} = \tilde{x}$ we have $E[Y_{i\tilde{t}}|X_i = X^k] = \int m(\tilde{x}, \alpha)Q^*_k(d\alpha)$. Therefore, for $\tilde{r}^k = \#\{t : X^k_t = \tilde{x}\}/T$ and $\bar{r}^k = \#\{t : X^k_t = \bar{x}\}/T$, by the law of large numbers,

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} 1(\tilde{r}_i &> 0)1(\bar{r}_i > 0)\{\frac{\sum_{t=1}^{T}\tilde{d}_{it}Y_{it}}{T\tilde{r}_i} - \frac{\sum_{t=1}^{T}\bar{d}_{it}Y_{it}}{T\bar{r}_i}\}/D \\
&\xrightarrow{p} E[1(\tilde{r}_i > 0)1(\bar{r}_i > 0)\{\frac{\sum_{t=1}^{T}\tilde{d}_{it}Y_{it}}{T\tilde{r}_i} - \frac{\sum_{t=1}^{T}\bar{d}_{it}Y_{it}}{T\bar{r}_i}\}]/D \\
&= E[1(\tilde{r}_i > 0)1(\bar{r}_i > 0)\{\frac{\sum_{t=1}^{T}\tilde{d}_{it}E[Y_{it}|X_i]}{T\tilde{r}_i} - \frac{\sum_{t=1}^{T}\bar{d}_{it}E[Y_{it}|X_i]}{T\bar{r}_i}\}]/D \\
&= \sum_{k\in\mathcal{K}^*}\mathcal{P}_k\{\frac{T\tilde{r}^k\int m(\tilde{x},\alpha)Q^*_k(d\alpha)}{T\tilde{r}^k} - \frac{T\bar{r}^k\int m(\bar{x},\alpha)Q^*_k(d\alpha)}{T\bar{r}^k}\}/D = \sum_{k\in\mathcal{K}^*}\mathcal{P}_k\mu_k, \\
\frac{1}{n}\sum_{i=1}^{n} 1(\tilde{r}_i &> 0)1(\bar{r}_i > 0) \xrightarrow{p} E[1(\tilde{r}_i > 0)1(\bar{r}_i > 0)] = \sum_{k\in\mathcal{K}^*}\mathcal{P}_k.
\end{aligned}
$$

Dividing and applying the continuous mapping theorem gives the result. Q.E.D.

Proof of Lemma 4: As before let $Q^*_k(\alpha)$ denote the conditional CDF of α given $X_i = X^k$. Note that

$$\bar{m}^k_t = \frac{E[Y_{it}|X_i = X^k]}{D} = \frac{\int m(X^k_t, \alpha)Q^*_k(d\alpha)}{D}.$$

Also we have

$$\mu_k = \int \Delta(\alpha)Q^*_k(d\alpha) = \frac{\int m(\tilde{x},\alpha)Q^*_k(d\alpha)}{D} - \frac{\int m(\bar{x},\alpha)Q^*_k(d\alpha)}{D}.$$

Then if there is $\tilde{t}$ and $\bar{t}$ such that $X^k_{\tilde{t}} = \tilde{x}$ and $X^k_{\bar{t}} = \bar{x}$

$$\bar{m}^k_{\tilde{t}} - \bar{m}^k_{\bar{t}} = \frac{\int m(\tilde{x},\alpha)Q^*_k(d\alpha)}{D} - \frac{\int m(\bar{x},\alpha)Q^*_k(d\alpha)}{D} = \mu_k.$$

Also, if $B_\ell \le m(x,\alpha)/D \le B_u$, then for each k,

$$B_\ell \le \frac{\int m(\tilde{x},\alpha)Q^*_k(d\alpha)}{D} \le B_u, -B_u \le -\frac{\int m(\bar{x},\alpha)Q^*_k(d\alpha)}{D} \le -B_\ell$$

Then if there is $\tilde{t}$ such that $X^k_{\tilde{t}} = \tilde{x}$ we have

$$\bar{m}^k_{\tilde{t}} - B_u = \frac{\int m(\tilde{x},\alpha)Q^*_k(d\alpha)}{D} - B_u \le \mu_k \le \frac{\int m(\tilde{x},\alpha)Q^*_k(d\alpha)}{D} - B_\ell = \bar{m}^k_{\tilde{t}} - B_\ell.$$

The second inequality in the statement of the theorem follows similarly.

Next, if $\Delta(\alpha)$ has the same sign for all α and if for some k^* there is $\tilde{t}$ and $\bar{t}$ such that $X^{k^*}_{\tilde{t}} = \tilde{x}$ and $X^{k^*}_{\bar{t}} = \bar{x}$, then $sgn(\Delta(\alpha)) = sgn(\mu_{k^*})$. Furthermore, since $sgn(\mu_k) = sgn(\mu_{k^*})$ is then known for all k, if it is positive the lower bounds, which are nonpositive, can be replaced by zero, while if it is negative the upper bounds, which are nonnegative, can be replaced by zero. Q.E.D.

Proof of Theorem 5: See text.

Proof of Theorem 6: Let $Z_{iT} = \min\{\sum_{t=1}^{T} 1(X_{it} = \tilde{x})/T, \sum_{t=1}^{T} 1(X_{it} = \bar{x})/T\}$. Note that if $Z_{iT} > 0$ then $1(A_{iT}) = 1$ for the event A_{iT} that there exists $\tilde{t}$ such that $X_{i\tilde{t}} = \tilde{x}$ and $X_{i\bar{t}} = \bar{x}$. By the ergodic theorem and continuity of the minimum, conditional on α_i we have $Z_{iT} \xrightarrow{as} b(\alpha_i) = \min\{\Pr(X_{it} = \tilde{x}|\alpha_i), \Pr(X_{it} = \bar{x}|\alpha_i)\} > 0$. Therefore $\Pr(A_{iT}|\alpha_i) \geq \Pr(Z_{iT} > 0|\alpha_i) \longrightarrow 1$ for almost all α_i. It then follows by the dominated convergence theorem that

$$\Pr(A_{iT}) = E[\Pr(A_{iT}|\alpha_i)] \longrightarrow 1.$$

Also note that $\Pr(A_{iT}) = 1 - \mathcal{P}^0 - \sum_{k \in \tilde{K}} \mathcal{P}_k - \sum_{k \in \bar{K}} \mathcal{P}_k$, so that

$$|\mu_\ell - \mu_0| \leq (B_u - B_\ell)(\mathcal{P}^0 + \sum_{k \in \tilde{K}} \mathcal{P}_k + \sum_{k \in \bar{K}} \mathcal{P}_k) \longrightarrow 0.Q.E.D.$$

Proof of Theorem 7: Let $\mathcal{P}_1$ and $\mathcal{P}_K$ be as in equation (7). Since $X_{i1}, \ldots, X_{iT}$ are i.i.d. conditional on α_i we have

$$\begin{aligned} \mathcal{P}_1 &= \Pr(X_{i1} = \cdots = X_{iT} = 0) = E[\Pr(X_{i1} = \cdots = X_{iT} = 0|\alpha_i)] \\ &= E[\Pi_{t=1}^{T} \Pr(X_{it} = 0|\alpha_i)] = E[\{1 - P(\alpha_i)\}^T]. \\ \mathcal{P}_K &= E[P(\alpha_i)^T]. \end{aligned}$$

The first bound then follows as in (7). The second bound then follows from $P(\alpha_i) \leq 1 - \varepsilon$ and $1 - P(\alpha_i) \leq 1 - \varepsilon$. Now suppose that $P(\alpha_i) = 1$ with positive probability. Then

$$\mathcal{P}_K \geq E[1(P(\alpha_i) = 1) \cdot P(\alpha_i)^T] = \Pr(P(\alpha_i) = 1) > 0.$$

Therefore, for all T the probability $\mathcal{P}_K$ is bounded away from zero, and hence $\mu_\ell \nrightarrow \mu_0$ or $\mu_u \nrightarrow \mu_0$.Q.E.D.

Proof of Theorem 8: The size of the identified set for the marginal effect is

$$\overline{\mu}_k - \underline{\mu}_k = \max_{Q_k \in \mathcal{Q}_{k\beta}, \beta \in B} D^{-1} \int [F(\beta + \alpha) - F(\alpha)] Q_k(d\alpha) - \min_{Q_k \in \mathcal{Q}_{k\beta}, \beta \in B} D^{-1} \int [F(\beta + \alpha) - F(\alpha)] Q_k(d\alpha),$$

where $\mathcal{Q}_{k\beta} = \{Q_k : \int \mathcal{L}\left(Y^j|X^k, \alpha, \beta\right) Q_k(d\alpha) = \mathcal{P}_{jk}, \quad j = 1, ..., J\}$. The feasible set of distributions $\mathcal{Q}_{k\beta}$ can be further characterized in this case. Let $F_T(\beta, \alpha) := (1, F(X_1^k \beta + \alpha), ..., F(X_T^k \beta + \alpha))$ and $\mathcal{F}_J(\beta, \alpha)$ denote the $J \times 1$ power vector of $F_T(\beta, \alpha)$ including all the different products of the elements of $F_T(\beta, \alpha)$, i.e.,

$$\mathcal{F}_J(\beta, \alpha) = (1, ..., F(X_T^k \beta + \alpha), F(X_1^k \beta + \alpha) F(X_2^k \beta + \alpha), ..., \prod_{t=1}^{T} F(X_t^k \beta + \alpha)).$$

Note that $\mathcal{L}\left(Y^j|X^k,\alpha,\beta\right)=\prod_{t=1}^{T}F(X_t^k\beta+\alpha)^{Y_t^j}\{1-F(X_t^k\beta+\alpha)\}^{1-Y_t^j}$, so the model probabilities are linear combinations of the elements of $\mathcal{F}_J(\beta,\alpha)$. Therefore, for $\Pi_k=(\mathcal{P}_{1k},...,\mathcal{P}_{Jk})$ we have $\mathcal{Q}_{k\beta}=\{Q_k:\mathcal{A}_J\int\mathcal{F}_J(\beta,\alpha)Q_k\left(d\alpha\right)=\Pi_k\}$, where $\mathcal{A}_J$ is a $J\times J$ matrix of known constants. The matrix $\mathcal{A}_J$ is nonsingular, so we have:

$$\mathcal{Q}_{k\beta}=\left\{Q_k:\int\mathcal{F}_J(\beta,\alpha)Q_k\left(d\alpha\right)=M_k\right\},$$

where the $J\times 1$ vector $M_k=\mathcal{A}_J^{-1}\Pi_k$ is identified from the data.

Now we turn to the analysis of the size of the identified sets. We focus on the case where $k=1$, i.e., X^k is a vector of zeros, and a similar argument applies to $k=K$. For $k=1$ we have that $F(X_t^k\beta+\alpha)=F(\alpha)$ for all t, so the power vector has only $T+1$ different elements given by $(1,F(\alpha),...,F(\alpha)^T)$. The feasible set simplifies to:

$$\mathcal{Q}_{k\beta}=\left\{Q_k:\int F(\alpha)^tQ_k\left(d\alpha\right)=M_{tk},\ t=0,...,T\right\},$$

where the moments M_{kt} are identified by the data. Here $\int F(\alpha)Q_k\left(d\alpha\right)=M_{1k}$ is fixed in $\mathcal{Q}_{k\beta}$, so the size of the identified set is given by:

$$\overline{\mu}_k-\underline{\mu}_k=\max_{Q_k\in\mathcal{Q}_{k\beta},\beta\in B}D^{-1}\int F\left(\beta+\alpha\right)Q_k(d\alpha)-\min_{Q_k\in\mathcal{Q}_{k\beta},\beta\in B}D^{-1}\int F\left(\beta+\alpha\right)Q_k(d\alpha).$$

By a change of variable $Z=F(\alpha)$, we can express the previous problem in a form that is related to a Hausdorff truncated moment problem:

$$\overline{\mu}_k-\underline{\mu}_k=\max_{G_k\in\mathcal{G}_{k\beta},\beta\in B}D^{-1}\int_0^1h_\beta(z)G_k(dz)-\max_{G_k\in\mathcal{G}_{k\beta},\beta\in B}D^{-1}\int_0^1h_\beta(z)G_k(dz),\tag{26}$$

where $\mathcal{G}_{k\beta}=\{G_k:\int_0^1z^tG_k(dz)=M_{tk},\quad t=0,...,T\}$, $h_\beta(z)=F(\beta+F^{-1}(z))$, and F^{-1} is the inverse of F.

If the objective function is r times continuously differentiable, $h_\beta\in\mathcal{C}^r[0,1]$, with uniformly bounded r-th derivative, $\|h_\beta^r(z)\|_\infty\leq\bar{h}_\beta^r$, then we can decompose h_β using standard approximation theory techniques as

$$h_\beta(z)=P_\beta(z,T)+R_\beta(z,T),\tag{27}$$

where $P_\beta(z,T)$ is the T-degree best polynomial approximation to h_β and $R_\beta(z,T)$ is the remainder term of the approximation, see, e.g., Judd (1998) Chap. 3. By Jackson's Theorem the remainder term is uniformly bounded by

$$\|R_\beta(z,T)\|_\infty\leq\frac{(T-r)!}{T!}\left(\frac{\pi}{4}\right)^r\bar{h}_\beta^r=O\left(T^{-r}\right),\tag{28}$$

as $T\to\infty$, and this is the best possible uniform rate of approximation by a T-degree polynomial.

Next, note that for any $G_k \in \mathcal{G}_{k\beta}$ we have that $\int_0^1 P_\beta(z,T)G_k(dz)$ is fixed, since the first T moments of Z are fixed at $\mathcal{G}_{k\beta}$. Moreover, $\int_0^1 P_\beta(z,T)G_k(dz)$ is fixed at B if the parameter is point identified, $B = \{\beta^*\}$. Then, we have

$$\overline{\mu}_k - \underline{\mu}_k = \max_{G_k \in \mathcal{G}_{k\beta}} \int_0^1 R_{\beta^*}(z,T)G_k(dx) - \min_{G_k \in \mathcal{G}_{k\beta}} \int_0^1 R_{\beta^*}(z,T)G_k(dx) \leq 2\bar{h}^r_{\beta^*} = O\left(T^{-r}\right). \quad (29)$$

To complete the proof, we need to check the continuous differentiability condition and the point identification of the parameter for the logit model. Point identification follows from Chamberlain (1992). For differentiability, note that for the logit model

$$h_\beta(z) = \frac{ze^\beta}{1-(1-e^\beta)z}, \quad (30)$$

with derivatives

$$h^r_\beta(z) = r!\frac{e^\beta(1-e^\beta)^{r-1}}{[1-(1-e^\beta)z]^r}. \quad (31)$$

These derivatives are uniformly bounded by $\bar{h}^r_\beta = r!\ e^{|\beta|}(e^{|\beta|}-1|)^{r-1} < \infty$ for any finite r. Q.E.D.

Proof of Theorem 9: Consider the case where $X^k = (0,1)$, a similar argument applies to $X^k = (1,0)$. By Lemma 4 we have that the marginal effect μ_k is identified by

$$\mu_k = P\{Y=(0,1)|X=(0,1)\} - P\{Y=(1,0)|X=(0,1)\}. \quad (32)$$

The probability limit of the fixed effects estimator for this marginal effect is

$$\tilde{\mu}_k(\tilde{\beta}) = [2F(\tilde{\beta}/2)-1][P\{Y=(0,1)|X=(0,1)\} + P\{Y=(1,0)|X=(0,1)\}]. \quad (33)$$

The condition for consistency $\tilde{\mu}_k(\tilde{\beta}) = \mu_k$ is therefore

$$F(\tilde{\beta}/2) = \frac{P\{Y=(0,1)|X=(0,1)\}}{P\{Y=(0,1)|X=(0,1)\} + P\{Y=(1,0)|X=(0,1)\}}, \quad (34)$$

but this is precisely the first order condition of the program (11). This result follows, after some algebra, using that $P\{Y=(1,0)|X=(1,0)\} = P\{Y=(0,1)|X=(0,1)\}$ and $P\{Y=(0,1)|X=(1,0)\} = P\{Y=(1,0)|X=(0,1)\}$. Q.E.D.

Proof of Lemma 10: Let the vector of conditional choice probabilities for $(Y^1, ..., Y^J)$ be

$$\mathcal{L}_k(\beta,\alpha) \equiv \left(\mathcal{L}_{1k}(\beta,\alpha), ..., \mathcal{L}_{Jk}(\beta,\alpha)\right)'.$$

Let $\Gamma_k(\beta) \equiv \{\mathcal{L}_k(\beta,\alpha) : \alpha \in \mathbb{C}\}$. Note that, for each $\beta \in B$, $\Gamma_k(\beta)$ is a closed and bounded set due to compactness of $\mathbb{C}$. Now, let $\mathcal{M}_k(\beta)$ denote the convex hull of $\Gamma_k(\beta)$. By Lindsay (1995, Theorem 18, p. 112), it follows that there exists a unique $\bar{\mathcal{L}}_k(\beta)$ on the boundary of $\mathcal{M}_k(\beta)$ that maximizes $\sum_{j=1}^{J}\mathcal{P}_{jk}\log(l_{jk})$ over all $(l_{1k},...,l_{Jk})\in\mathcal{M}_k(\beta)$. By Lindsay (1995, Theorem 21, p. 116), the solution $\bar{\mathcal{L}}_k(\beta)$ can be represented as

$$\left(\int\mathcal{L}_{1k}(\beta,\alpha)\,\bar{Q}_k(d\alpha),...,\int\mathcal{L}_{Jk}(\beta,\alpha)\,\bar{Q}_{k\beta}(d\alpha)\right)',$$

where $\bar{Q}_{k\beta}$ has no more than J points of support. Also, by $\beta\in B$, we have that $\arg\max_{(l_{1k},...,l_{Jk})\in\mathcal{M}_k(\beta)}\sum_{j=1}^{J}\mathcal{P}_{jk}$ satisfies $l_{jk}=\mathcal{P}_{jk}$. Q.E.D.

Proof of Lemma 11: For $\beta\in B$, let $\mathcal{Q}_{k\beta}=\{Q_k:\int\mathcal{L}\left(Y^j|X^k,\alpha,\beta\right)Q_k(d\alpha)=\mathcal{P}_{jk},\quad j=1,...,J\}$. Let $Q_{k\beta}\in\mathcal{Q}_{k\beta}$ denote some maximizing value such that

$$\overline{\mu}_{k\beta}=D^{-1}\int_{\mathbb{C}}[F\left(\tilde{x}'\beta+\alpha\right)-F\left(\bar{x}'\beta+\alpha\right)]Q_{k\beta}(d\alpha).$$

Note that, for any $\epsilon>0$ we can find a distribution $\bar{Q}_{k\beta}^{M}\in\mathcal{Q}_{k\beta}$ with a large number $M\gg J$ of support points $(\alpha_1,...,\alpha_M)$ such that

$$\overline{\mu}_{k\beta}-\epsilon<D^{-1}\int_{\mathbb{C}}[F\left(\tilde{x}'\beta+\alpha\right)-F\left(\bar{x}'\beta+\alpha\right)]\bar{Q}_{k\beta}^{M}(d\alpha)\leq\overline{\mu}_{k\beta}.$$

Our goal is to show that given such $\bar{Q}_{k\beta}^{M}$ it suffices to allocate its mass over only at most J support points. Indeed, consider the problem of allocating $(\pi_{1k},...,\pi_{Mk})$ among $(\alpha_1,...,\alpha_M)$ in order to solve

$$\max_{(\pi_{1k},...,\pi_{Mk})}\sum_{m=1}^{M}[F\left(\tilde{x}'\beta+\alpha_m\right)-F\left(\bar{x}'\beta+\alpha_m\right)]\pi_{mk}$$

subject to the constraints:

$$\begin{aligned}&\pi_{mk}\geq 0,\quad m=1,\ldots,M\\&\sum_{m=1}^{M}\pi_{mk}\mathcal{L}\left(Y^j|X^k,\alpha_m,\beta\right)=\mathcal{P}_{jk},\ j=1,...,J,\\&\sum_{m=1}^{M}\pi_{mk}=1.\end{aligned}$$

This a linear program of the form

$$\max_{\pi\in\mathbb{R}^M}c'\pi\quad\text{such that}\quad\pi\geq 0,\quad A\pi=b,\quad 1'\pi=1,$$

and any basic feasible solution to this program has M active constraints, of which at most $rank\,(A)+1$ can be equality constraints. This means that at least $M-\text{rank}(A)-1$ of active constraints are the form $\pi_{mk}=0$.[7] Hence a basic solution to this linear programming problem

[7] See, e.g., Theorem 2.3 and Definition 2.9 (ii) in Bertsimas and Tsitsiklis (1997).

will have at least $M-J$ zeroes, that is at most J strictly positive π_{mk}'s.[8] Thus, we have shown that given the original $\bar{Q}_{k\beta}^{M}$ with $M \gg J$ points of support there exists a distribution $\bar{Q}_{k\beta}^{L} \in \mathcal{Q}_{k\beta}$ with just J points of support such that

$$\overline{\mu}_{k\beta}-\epsilon < D^{-1}\int_{\mathbb{C}}[F\left(\tilde{x}'\beta+\alpha\right)-F\left(\bar{x}'\beta+\alpha\right)]\bar{Q}_{k\beta}^{M}\left(d\alpha\right) \leq D^{-1}\int_{\mathbb{C}}[F\left(\tilde{x}'\beta+\alpha\right)-F\left(\bar{x}'\beta+\alpha\right)]\bar{Q}_{k\beta}^{L}\left(d\alpha\right) \leq \overline{\mu}_{k\beta}.$$

This construction works for every $\epsilon > 0$.

The final claim is that there exists a distribution $\bar{Q}_{k\beta}^{L} \in \mathcal{Q}_{k\beta}$ with J points of support $(\alpha_1, ..., \alpha_J)$ such that

$$\overline{\mu}_{k\beta} = D^{-1}\int_{\mathbb{C}}[F\left(\tilde{x}'\beta+\alpha\right)-F\left(\bar{x}'\beta+\alpha\right)]\bar{Q}_{k\beta}^{L}\left(d\alpha\right).$$

Suppose otherwise, then it must be that

$$\overline{\mu}_{k\beta} > \overline{\mu}_{k\beta}-\epsilon \geq D^{-1}\int_{\mathbb{C}}[F\left(\tilde{x}'\beta+\alpha\right)-F\left(\bar{x}'\beta+\alpha\right)]\bar{Q}_{k\beta}^{L}\left(d\alpha\right),$$

for some $\epsilon > 0$ and for *all* $\bar{Q}_{k\beta}^{L}$ with J points of support. This immediately gives a contradiction to the previous step where we have shown that, for any $\epsilon > 0$, $\overline{\mu}_{k\beta}$ and the right hand side can be brought close to each other by strictly less than ϵ. Q.E.D.

Some Lemmas are useful for proving Theorem 12. For the proof of Theorem 12 we will assume for simplicity of notation that the regressor only takes one value $X^k = (x_1, x_2)$ and drop the dependence on k. We will also assume a two-period binary choice model with individual location effect. The proof for the general case follows by an identical argument, but the notation is more cumbersome.

The first Lemma establishes uniform consistency of $\frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right)$, as is useful for showing consistency of B_n.

LEMMA A1: *If Assumption 1 is satisfied then for $\mathbb{Q}$ equal to the collection of distributions with support contained in a compact set $\mathbb{C}$.*

$$\sup_{\beta\in\mathbb{B},Q\in\mathbb{Q}}\left|\frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) - E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right]\right| = O_{p^*}\left(\frac{1}{\sqrt{n}}\right)$$

[8] Note that $rank(A) \leq J-1$, since $\sum_{j=1}^{J}\mathcal{L}\left(Y^j|X^k,\alpha,\beta\right) = 1$. The exact rank of A depends on the sequence X^k, the parameter β, the function F and T. For $T=2$ and X binary, for example, $rank(A) = J-2 = 2$ when $x_1 = x_2$, $\beta = 0$, or F is the logistic distribution; whereas $rank(A) = J-1 = 3$ for $X_1^k \neq X_2^k$, $\beta \neq 0$, and F is any continuous distribution different from the logistic.

Proof: Note that

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1},Y_{i2};\beta,Q\right) \\
= \;& \left[\frac{1}{n}\sum_{i=1}^{n} Y_{i1}Y_{i2}\right]\cdot\log\left(\int F\left(x_1'\beta+\alpha\right)F\left(x_2'\beta+\alpha\right)Q\left(d\alpha\right)\right) \\
& +\left[\frac{1}{n}\sum_{i=1}^{n} Y_{i1}\left(1-Y_{i2}\right)\right]\cdot\log\left(\int F\left(x_1'\beta+\alpha\right)\left(1-F\left(x_2'\beta+\alpha\right)\right)Q\left(d\alpha\right)\right) \\
& +\left[\frac{1}{n}\sum_{i=1}^{n} \left(1-Y_{i1}\right)Y_{i2}\right]\cdot\log\left(\int \left(1-F\left(x_1'\beta+\alpha\right)\right)F\left(x_2'\beta+\alpha\right)Q\left(d\alpha\right)\right) \\
& +\left[\frac{1}{n}\sum_{i=1}^{n} \left(1-Y_{i1}\right)\left(1-Y_{i2}\right)\right]\cdot\log\left(\int \left(1-F\left(x_1'\beta+\alpha\right)\right)\left(1-F\left(x_2'\beta+\alpha\right)\right)Q\left(d\alpha\right)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
& E\left[L\left(Y_{i1},Y_{i2};\beta,Q\right)\right] \\
= \;& E\left[Y_{i1}Y_{i2}\right]\cdot\log\left(\int F\left(x_1'\beta+\alpha\right)F\left(x_2'\beta+\alpha\right)Q\left(d\alpha\right)\right) \\
& +E\left[Y_{i1}\left(1-Y_{i2}\right)\right]\cdot\log\left(\int F\left(x_1'\beta+\alpha\right)\left(1-F\left(x_2'\beta+\alpha\right)\right)Q\left(d\alpha\right)\right) \\
& +E\left[\left(1-Y_{i1}\right)Y_{i2}\right]\cdot\log\left(\int \left(1-F\left(x_1'\beta+\alpha\right)\right)F\left(x_2'\beta+\alpha\right)Q\left(d\alpha\right)\right) \\
& +E\left[\left(1-Y_{i1}\right)\left(1-Y_{i2}\right)\right]\cdot\log\left(\int \left(1-F\left(x_1'\beta+\alpha\right)\right)\left(1-F\left(x_2'\beta+\alpha\right)\right)Q\left(d\alpha\right)\right)
\end{aligned}
$$

Further note that $\frac{1}{n}\sum_{i=1}^{n} Y_{i1}Y_{i2} = E\left[Y_{i1}Y_{i2}\right]+O_p\left(\frac{1}{\sqrt{n}}\right)$, etc. Therefore, the requisite uniform convergence with rate $O_p\left(\frac{1}{\sqrt{n}}\right)$

$$
\Delta_n = \sup_{\beta\in\mathbb{B},Q\in\mathbb{Q}}\left|\frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1},Y_{i2};\beta,Q\right)-E\left[L\left(Y_{i1},Y_{i2};\beta,Q\right)\right]\right| = O_p\left(\frac{1}{\sqrt{n}}\right)
$$

follows, provided

$$
\begin{aligned}
& \left|\log\left(\int F\left(x_1'\beta+\alpha\right)F\left(x_2'\beta+\alpha\right)Q\left(d\alpha\right)\right)\right|, \quad \left|\log\left(\int F\left(x_1'\beta+\alpha\right)\left(1-F\left(x_2'\beta+\alpha\right)\right)Q\left(d\alpha\right)\right)\right|, \\
& \left|\log\left(\int \left(1-F\left(x_1'\beta+\alpha\right)\right)F\left(x_2'\beta+\alpha\right)Q\left(d\alpha\right)\right)\right|, \quad \left|\log\left(\int \left(1-F\left(x_1'\beta+\alpha\right)\right)\left(1-F\left(x_2'\beta+\alpha\right)\right)Q\left(d\alpha\right)\right)\right|
\end{aligned}
$$

are bounded, which in turn is implied by Assumption 1. Q.E.D.

From Lemma A1, we obtain one-sided uniform convergence:

Lemma A2: *If Assumption 1 is satisfied then*

$$
\sup_{\beta\in\mathbb{B}}\left|\sup_{Q\in\mathbb{Q}}\frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1},Y_{i2};\beta,Q\right)-\sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1},Y_{i2};\beta,Q\right)\right]\right| = O_{p^*}\left(\frac{1}{\sqrt{n}}\right)
$$

Proof: Define

$$Q^*(\beta) \in \arg\sup_{Q\in\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right), \;\; Q^{\#}(\beta) \in \arg\sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right].$$

By definition of $Q^*(\beta)$ and $Q^{\#}(\beta)$, we have uniformly in β and for all n,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} L(Y_{i1}, Y_{i2}; \beta, Q^{\#}(\beta)) - E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q^{\#}(\beta)\right)\right] \\
&\leq \;\; \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q^*(\beta)\right) - E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q^{\#}(\beta)\right)\right] \\
&\leq \;\; \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q^*(\beta)\right) - E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q^*(\beta)\right)\right]
\end{aligned}$$

Hence

$$\left|\frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q^*(\beta)\right) - E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q^{\#}(\beta)\right)\right]\right| \leq 2\Delta_n = O_{p^*}\left(\frac{1}{\sqrt{n}}\right)$$

uniformly in β, where Δ_n was defined in (35). Because $\Delta_n = O_p\left(\frac{1}{\sqrt{n}}\right)$, we obtain the desired result. Q.E.D.

Lemma A3: *If Assumption 1 is satisfied then* $\max_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right]$ *is continuous in* β.

Proof: The problem

$$\max_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right]$$

can be rewritten as

$$\max_{\substack{(\alpha_1,\ldots,\alpha_J)\in\mathbb{C}\\(\pi_1,\ldots,\pi_J)\in\mathbb{S}}} \sum_{j=1}^{J} \mathcal{P}_j \log\left[\sum_{m=1}^{J} \mathcal{L}\left(Y^j | X, \alpha_m, \beta\right)\pi_m\right],$$

where $J = 4$, $\mathcal{P}_j = \Pr(Y_i = Y^j)$ and $\mathbb{S}$ denotes the unit simplex in $\mathbb{R}^J$. Here, $(\alpha_1, \ldots, \alpha_j)$ and $(\pi_1, \ldots, \pi_J)$ characterize a discrete distribution with no more than J points of support. Because the objective function is continuous in $(\beta, \alpha_1, \ldots, \alpha_J, p_1, \ldots, p_J)$, and because $\mathbb{C}\times\mathbb{S}$ is compact, we can apply the Theorem of the Maximum (e.g. Stokey and Lucas 1989, Theorem 3.6), and obtain the desired conclusion. Q.E.D.

Proof of Theorem 12: If B is a singleton the result follows by the uniform convergence of the profile objective function from Lemma A2 and the continuity of the limit objective function from Lemma A3. The proof for case where B is not a singleton consists of two parts.

Part 1: The first part of the proof modifies slightly the argument of Manski and Tamer (2002) for the present context. Define

$$\begin{aligned}
\bar{L}_n^* &\equiv \sup_{\beta\in\mathbb{B}} \sup_{Q\in\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right), \\
L_n^* &\equiv \inf_{\beta\in B} \sup_{Q\in\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right), \\
L^* &\equiv \sup_{\beta\in\mathbb{B}} \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] = \sup_{\beta\in B} \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right], \\
\Delta_n &\equiv \sup_{\beta\in\mathbb{B}, Q\in\mathbb{Q}} \left| \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) - E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right|.
\end{aligned} \tag{35}$$

Note that $\sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right]$ is constant over B by definition, which implies that

$$L^* = \inf_{\beta\in B} \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right]$$

Therefore, we obtain

$$\begin{aligned}
\left|L_n^* - L^*\right| &= \left| \inf_{\beta\in B} \sup_{Q\in\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) - \inf_{\beta\in B} \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right| \\
&\leq \sup_{\beta\in B} \left| \sup_{Q\in\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) - \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right| \\
&\leq \sup_{\beta\in\mathbb{B}, Q\in\mathbb{Q}} \left| \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) - E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right| = \Delta_n
\end{aligned}$$

Also note that

$$\left|\bar{L}_n^* - L^*\right| = \left| \sup_{\beta\in\mathbb{B}} \sup_{Q\in\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) - \sup_{\beta\in\mathbb{B}} \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right| \leq \Delta_n$$

It follows that

$$\left|\bar{L}_n^* - L_n^*\right| \leq \left|\bar{L}_n^* - L^*\right| + \left|L_n^* - L^*\right| \leq \Delta_n + \Delta_n = 2\Delta_n.$$

Suppose now that $b \in B$. Note that

$$\bar{L}_n^* - \sup_{Q\in\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; b, Q\right) \leq \bar{L}_n^* - \inf_{\beta\in B} \sup_{Q\in\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) = \bar{L}_n^* - L_n^*$$

Therefore, if $\epsilon_n > \bar{L}_n^* - L_n^*$, then we have $\bar{L}_n^* - \sup_{Q\in\mathbb{Q}} \frac{1}{n}\sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; b, Q\right) \leq \epsilon_n$, or

$$b \in B_n$$

by definition of B_n. In other words, $\epsilon_n > \bar{L}_n^* - L_n^*$, then $\epsilon_n > \bar{L}_n^* - L_n^*$, $\inf_{b_n\in B_n} |b_n - b| = 0$. Because the choice of b was arbitrary, we can conclude that

$$\sup_{b\in B} \inf_{b_n\in B_n} |b_n - b| = 0$$

if $\epsilon_n > \bar{L}_n^* - L_n^*$. Because $\epsilon_n > 2\Delta_n$ with probability converging to one due to Lemma A2 and the choice of ϵ_n, it follows that $\sup_{b\in B} \inf_{b_n\in B_n} |b_n - b| = 0$ with probability converging to one.[9]

Part 2: Define

$$B(\epsilon) \equiv \left\{ \beta : L^* - \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \leq \epsilon \right\}$$

It suffices to show that $B_n \subseteq B(\epsilon)$ with probability converging to one. This is because it would imply $\inf_{b\in B} |b_n - b| < \delta(\epsilon)$ for $(b_n \in B_n)$, which implies

$$\sup_{b_n\in B_n} \inf_{b\in B} |b_n - b| < \delta(\epsilon),$$

with probability converging to one. Here $\delta(\epsilon)$ that can be made arbitrarily small by making ϵ sufficiently small by continuity of $\sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right]$ in β, which was established in Lemma A3. This would prove that $\sup_{b_n\in B_n} \inf_{b\in B} |b_n - b| = o_p(1)$.

It remains to show that, for any $\epsilon > 0$, we have $B_n \subseteq B(\epsilon)$ with probability converging to one. For this purpose it suffices to show that

$$\sup_{\beta\in B_n} \left[L^* - \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right] \leq \epsilon.$$

Note that

$$\begin{aligned}
& \left| \sup_{\beta\in B_n} \left(L^* - \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right) - \sup_{\beta\in B_n} \left(\bar{L}_n^* - \sup_{Q\in\mathbb{Q}} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) \right) \right| \\
\leq \; & \sup_{\beta\in B_n} \left| \left(L^* - \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right) - \left(\bar{L}_n^* - \sup_{Q\in\mathbb{Q}} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) \right) \right| \\
\leq \; & \left| L^* - \bar{L}_n^* \right| + \sup_{\beta\in B_n} \left| \sup_{Q\in\mathbb{Q}} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) - \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right| \\
\leq \; & 2\Delta_n.
\end{aligned}$$

By definition of the level set B_n, we have

$$\sup_{\beta\in B_n} \left[\bar{L}_n^* - \sup_{Q\in\mathbb{Q}} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_{i1}, Y_{i2}; \beta, Q\right) \right] \leq \epsilon_n.$$

It follows that

$$\sup_{\beta\in B_n} \left[L^* - \sup_{Q\in\mathbb{Q}} E\left[L\left(Y_{i1}, Y_{i2}; \beta, Q\right)\right] \right] \leq \epsilon_n + 2\Delta_n$$

[9] The "probability" here actually means the inner probability. We ignore such measure theoretic subtlety in this paper.

By Lemma A1 and choice of ϵ_n, we have $\epsilon_n + 2\Delta_n < \epsilon$ with probability converging to one, which shows the requisite claim. Q.E.D.

Proof of Corollary 13: The results follows from Theorem 12 and the continuous mapping theorem. Q.E.D.

References

[1] Aitkin, M, (1999), "A general maximum likelihood analysis of variance components in generalized linear models," *Biometrics* 55 (1), 117–128.

[2] Alvarez, J., and M. Arellano (2003), "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," *Econometrica* 71, 1121-1159.

[3] Angrist, J. D. (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica* 66, 249–288.

[4] Andersen, E. (1973), *Conditional Inference and Models for Measuring.* Copenhagen: Mentalhygiejnisk Forlag.

[5] Bertsimas, D., and Tsitsiklis, J. N. (1997), *Introduction to Linear Optimization*, Athena Scientific, Belmont, Massachusetts.

[6] Blundell, R. and J.L. Powell (2003), "Endogeneity in Nonparametric and Semiparametric Regression Models," in M. Dewatripont, L. P. Hansen and S. J. Turnsovsky (eds.) *Advances in Economics and Econometrics*, Cambridge: Cambridge University Press.

[7] Browning, M. and J. Carro (2007), "Heterogeneity and Microeconometrics Modeling," in Blundell, R., W.K. Newey, T. Persson (eds.), *Advances in Theory and Econometrics, Vol. 3*, Cambridge: Cambridge University Press.

[8] Carro, J. M. (2007), "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects," *Journal of Econometrics* 140(2), pp 503-528.

[9] Carrasco, R. (2001), "Binary Choice With Binary Endogenous Regressors in Panel Data: Estimating the Effect of Fertility on Female Labor Participation," *Journal of Business and Economic Statistics* 19(4), 385-394.

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

[11] CHAMBERLAIN, G. (1984), "Panel Data," in Z. GRILICHES AND M. INTRILIGATOR eds *Handbook of Econometrics*. Amsterdam: North-Holland.

[12] CHAMBERLAIN, G. (1992), "Binary Response Models for Panel Data: Identification and Information," *unpublished manuscript.*

[13] CHAY, K. Y., AND D. R. HYSLOP (2000), "Identification and Estimation of Dynamic Binary Response Panel Data Models: Empirical Evidence using Alternative Approaches," unpublished manuscript, University of California at Berkeley.

[14] CHERNOZHUKOV, V. (2007), "Course Materials for 14.385 Nonlinear Econometric Analysis, Fall 2007," MIT OpenCourseWare (http://ocw.mit.edu), MIT.

[15] CHERNOZHUKOV, V., J.HAHN, AND W.K.NEWEY (2004), "Bound Analysis in Panel Models with Correlated Random Effects," *unpublished manuscript.*

[16] DUFOUR, J.-M. (2006), "Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics," *Journal of Econometrics* 133, 443–477.

[17] FENG, Z., AND C. E. MCCULLOCH (1992), "Using Bootstrap Likelihood Ratios in Finite Mixture Models," *Journal of the Royal Statistical Society. Series B (Methodological)* 58(3), 609–617.

[18] FERNÁNDEZ-VAL, I. (2008), "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models," *unpublished manuscript.*

[19] GRAHAM, B.W. AND J.L. POWELL (2008), "Semiparametric Identification and Estimation of Correlated Random Coefficient Models for Panel Data," *unpublished manuscript.*

[20] HAHN, J. (2001), "Comment: Binary Regressors in Nonlinear Panel-Data Models with Fixed Effects," *Journal of Business and Economic Statistics* 19, 16-17.

[21] HAHN, J., AND G. KUERSTEINER (2002), "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when Both n and T Are Large," *Econometrica* 70, 1639-1657.

[22] HAHN, J., AND W. NEWEY (2004), "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica* 72, 1295-1319.

[23] HAHN, J., AND G. KUERSTEINER (2007), "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects, "*unpublished manuscript.*

[24] HECKMAN, J. J., AND T. E. MACURDY (1980), "A Life Cycle Model of Female Labor Supply," *Review of Economic Studies* 47, 47-74.

[25] HECKMAN, J. J., AND T. E. MACURDY (1982), "Corrigendum on: A Life Cycle Model of Female Labor Supply," *Review of Economic Studies* 49, 659-660.

[26] HONORÉ, B.E., AND E. TAMER (2006), "Bounds on Parameters in Dynamic Discrete Choice Models", *Econometrica* 74(3), 611-629.

[27] HYSLOP, D. R. (1999), "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica* 67(6), 1255-1294.

[28] JUDD, K. L. (1998), *Numerical Methods in Economics.* MIT Press, Cambridge, MA.

[29] LINDSAY, B.G. (1995), *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5, IMS: Hayward.

[30] MANSKI, C.F., AND E. TAMER (2002), "Inference on Regressions with Interval Data on a Regressor or Outcome,"*Econometrica* 70, 519 - 546.

[31] MCLACHLAN, G., AND D. PEEL (2000), *Finite mixture models.* Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

[32] ROMANO, J. P., AND M. WOLF, (2000), "Finite sample nonparametric inference and large sample efficiency," *Annals of Statistics*, 28(3), 756–778.

[33] RYTCHKOV, O.(2006), "Filtering Out Expected Dividends and Expected Returns," *unpublished manuscript.*

[34] STOKEY, N.L., AND R.E. LUCAS (1989), *Recursive Methods in Economic Dynamics*, Harvard University Press: Cambridge.

[35] WOOLDRIDGE, J.M. (2002), "*Econometric Analysis of Cross Section and Panel Data*," Cambridge, MA: MIT Press.

[36] WOOLDRIDGE, J.M. (2005), "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *Review of Economics and Statistics* 87, 385–390.

[37] WOUTERSEN, T.(2002), "Robustness Against Incidental Parameters," *unpublished manuscript.*

Table 1: Biases of linear probability model estimators in percentage of marginal effect (average probability of the response in parenthesis)

	p_X					
	0.5		0.1		0.9	
T	β_w	β	β_w	β	β_w	β
2	34.63	34.63	-91.20	-91.20	-31.07	-31.07
	(0.60)		(0.45)		(0.77)	
4	12.77	9.91	-61.52	-59.77	20.52	25.32
	(0.61)		(0.47)		(0.75)	
8	5.76	0.74	-33.16	-20.40	19.90	30.38
	(0.62)		(0.49)		(0.74)	

Notes: probit model with a single binary regressor with parameter equal to one. The individual effect is the standardized mean of the regressor. β_w is the probability limit of the linear fixed effects estimator with constant slopes and β is the probability limit of the average of the linear fixed effects estimators with individual specific slopes.

Table 2: Descriptive Statistics for NLSY79 sample
(n = 1,587)

Variable	Mean	Changes (%)
LFP1990	0.75	
LFP1992	0.74	0.17
LFP1994	0.75	0.28
LFP1996	0.76	0.35
kids1990	0.38	
kids1992	0.35	0.31
kids1994	0.45	0.51
kids1996	0.21	0.60

Notes: LFP - 1 if woman is in the labor force, 0 otherwise; kid - number of children of age less than 3. Changes (%) measures the proportion of women who change status between 1990 and the year corresponding to the row.

Table 3: LFP and Fertility (T = 2, n = 1,587)

		General CEF Model			Semiparametric Model							Linear
	$P(X^k)$	Est.	95% N-CI	95% B-CI*	Est.	95% LR	95% B-CI*	95% PB-CI**	FE	FE-BC	CMLE	FE
						Logit						
β					-.36	(-.75, .02)	(-.56, -.16)	(-.88, .08)	-.78	-.36	-.39	
									(-1.11, -.46)	(-.67, -.05,)	(-.70, -.08)	
μ(0,0)	.48	[-.81, 0]	(-.83, 0)	(-.84, 0)	[-.06, -.04]	(-.17, .00)	(-.11, -.02)	(-.22, .01)	-.05			
μ(0,1)	.14	-.12	(-.20, -.04)	(-.18, -.06)	-.06	(-.11, .00)	(-.09, -.03)	(-.16, .01)	-.07			
μ(1,0)	.17	-.03	(-.10, .05)	(-.08, .03)	-.07	(-.13, .00)	(-.10, -.03)	(-.15, .01)	-.07			
μ(1,1)	.21	[-.38, 0]	(-.42, 0)	(-.44, 0)	[-.07, -.05]	(-.15, .00)	(-.11, -.02)	(-.20, .01)	-.05			
μ		[-.49, -.02]	(-.53, .00)	(-.52, -.01)	[-.06, -.05]	(-.15, .00)	(-.11, -.02)	(-.19, .01)	-.06	-.04	-.05	-.07
									(-.08, -.04)	(-.06, -.02)	(-.07, -.02)	(-.11, -.03)
						Probit						
β					-.41	(-.85, .03)	(-.64, -.20)	(-1.06, .10)	-.88	-.51		
									(-1.24, -.52)	(-.86, -.16)		
μ(0,0)	.48	[-.81, 0]	(-.83, 0)	(-.84, 0)	[-.08, -.04]	(-.20, .00)	(-.14, -.02)	(-.24, .02)	-.05			
μ(0,1)	.14	-.12	(-.20, -.04)	(-.18, -.06)	-.07	(-.12, .00)	(-.11, -.03)	(-.17, .01)	-.07			
μ(1,0)	.17	-.03	(-.10, .05)	(-.08, .03)	-.07	(-.13, .01)	(-.10, -.04)	(-.15, .02)	-.07			
μ(1,1)	.21	[-.38, 0]	(-.42, 0)	(-.44, 0)	[-.07, -.05]	(-.16, .01)	(-.12, -.02)	(-.19, .02)	-.06			
μ		[-.49, -.02]	(-.53, .00)	(-.52, -.01)	[-.07, -.05]	(-.17, .00)	(-.12, -.03)	(-.19, .02)	-.06	-.05		-.07
									(-.08, -.04)	(-.07, -.02)		(-.11, -.03)

Notes: Dependent variable is labor force participation indicator; regressor is a fertility indicator that takes the value 1 if the woman has a child less than 3 years old. Time periods: 1990 and 1992. Source: NLSY79. *Based on 200 bootstrap repetitions. **Based on 100 DGP's.
9/7/2008

Table 4: LFP and Fertility (T = 3, n = 1,587)

		General CEF Model			Semiparametric Model								Linear
	$P(X^k)$	Est.	95% N-CI	95% B-CI*	Est.	95% LR	95% B-CI*	95% PB-CI**	FE	FE-BC	Jack.	CMLE	FE
							Logit						
β					-.42	-	(-.51, -.24)	(-.74, -.12)	-.71	-.46	-.38	-.46	
									(-.90, -.52)	(-.64, -.28)	(-.70, -.05)	(-.65, -.28)	
μ(0,0,0)	.4	[-.81, 0]	(-.83, 0)	(-.81, 0)	[-.06, -.06]	-	(-.08, -.03)	(-.14, -.02)	-.06				
μ(0,0,1)	.08	-.12	(-.21, -.04)	(-.20, -.06)	-.07	-	(-.09, -.03)	(-.14, -.01)	-.07				
μ(0,1,0)	.06	-.1	(-.20, .01)	(-.17, .01)	-.08	-	(-.10, -.04)	(-.17, -.02)	-.09				
μ(1,0,0)	.14	-.06	(-.14, .01)	(-.10, .01)	-.08	-	(-.10, -.05)	(-.15, -.02)	-.09				
μ(0,1,1)	.08	-.18	(-.26, -.09)	(-.23, -.09)	-.07	-	(-.10, -.04)	(-.16, -.02)	-.08				
μ(1,0,1)	.03	.02	(-.16, .20)	(-.02, .22)	-.08	-	(-.11, -.04)	(-.19, -.02)	-.09				
μ(1,1,0)	.12	-.04	(-.12, .04)	(-.15, -.01)	-.07	-	(-.10, -.04)	(-.15, -.02)	-.08				
μ(1,1,1)	.09	[-.41, 0]	(-.46, 0)	(-.42, 0)	[-.08, -.07]	-	(-.10, -.04)	(-.16, -.02)	-.08				
μ		[-.40, -.04]	(-.46, .00)	(-.41, -.02)	[-.07, -.07]	-	(-.09, -.04)	(-.13, -.02)	-.08	-.07	-.09	-.07	-.08
									(-.09, -.06)	(-.09, -.05)	(-.11, -.07)	(-.09, -.05)	(-.11, -.06)
							Probit						
β					-.46	-	(-.61, -.30)	(-.73, -.16)	-.78	-.55	-.38		
									(-.99, -.57)	(-.75, -.35)	(-.58, -.18)		
μ(0,0,0)	.4	[-.81, 0]	(-.83, 0)	(-.81, 0)	[-.08, -.06]	-	(-.11, -.04)	(-.15, -.02)	-.06				
μ(0,0,1)	.08	-.12	(-.21, -.04)	(-.20, -.06)	-.08	-	(-.11, -.04)	(-.16, -.02)	-.07				
μ(0,1,0)	.06	-.1	(-.20, .01)	(-.17, .01)	-.08	-	(-.11, -.05)	(-.16, -.03)	-.08				
μ(1,0,0)	.14	-.06	(-.14, .01)	(-.10, .01)	-.09	-	(-.11, -.05)	(-.15, -.02)	-.09				
μ(0,1,1)	.08	-.18	(-.26, -.09)	(-.23, -.09)	-.08	-	(-.11, -.05)	(-.17, -.02)	-.08				
μ(1,0,1)	.03	.02	(-.16, .20)	(-.02, .22)	-.08	-	(-.11, -.05)	(-.22, -.02)	-.09				
μ(1,1,0)	.12	-.04	(-.12, .04)	(-.15, -.01)	-.07	-	(-.11, -.05)	(-.15, -.02)	-.09				
μ(1,1,1)	.09	[-.41, 0]	(-.46, 0)	(-.42, 0)	[-.09, -.07]	-	(-.12, -.05)	(-.17, -.02)	-.08				
μ		[-.40, -.04]	(-.46, .00)	(-.41, -.02)	[-.08, -.07]	-	(-.11, -.05)	(-.14, -.03)	-.08	-.07	-.09		-.08
									(-.09, -.06)	(-.09, -.05)	(-.11, -.07)		(-.11, -.06)

Notes: Dependent variable is labor force participation indicator; regressor is a fertility indicator that takes the value 1 if the woman has a child less than 3 years old. Time periods: 1990, 1992, and 1994. Source: NLSY79. *200 boostraps repetitions. **Based on 100 DGP's.
9/7/2008

Table 5: LFP and Fertility (T = 4, n = 1,587)

		General CEF Model			Semiparametric Model							Linear
	$P(X^k)$	Est.	95% N-CI	95% B-CI	Est	95% LR	95% B-CI	FE	FE-BC	Jack.	CMLE	FE
							Logit					
β					-.4	-	(-.42, -.23)	-.65	-.48	-.44	-.48	
								(-.79, -.50)	(-.62, -.34)	(-.69, -.19)	(-.62, -.34)	
μ(0,0,0,0)	.37	[-.80, 0]	(-.82, 0)	(-.80, 0)	[-.07, -.07]	-	(-.07, -.04)	-.08				
μ(0,0,0,1)	.04	-.2	(-.33, -.07)	(-.30, -.09)	-.05	-	(-.07, -.02)	-.06				
μ(0,0,1,0)	.03	-.04	(-.17, .10)	(-.22, .03)	-.07	-	(-.09, -.03)	-.08				
μ(0,1,0,0)	.05	-.09	(-.20, .02)	(-.14, .03)	-.06	-	(-.08, -.03)	-.08				
μ(1,0,0,0)	.13	-.08	(-.15, -.01)	(-.11, .00)	-.08	-	(-.09, -.05)	-.1				
μ(0,0,1,1)	.05	-.14	(-.23, -.05)	(-.25, -.07)	-.07	-	(-.08, -.03)	-.08				
μ(0,1,0,1)	.01	-.09	(-.31, .13)	(-.40, .05)	-.1	-	(-.13, -.05)	-.12				
μ(1,0,0,1)	.01	-.06	(-.29, .18)	(-.25, .14)	-.08	-	(-.11, -.03)	-.1				
μ(0,1,1,0)	.04	-.06	(-.16, .05)	(-.23, -.05)	-.08	-	(-.10, -.04)	-.1				
μ(1,0,1,0)	.01	.03	(-.17, .22)	(-.12, .22)	-.05	-	(-.08, -.01)	-.07				
μ(1,1,0,0)	.11	-.08	(-.15, -.00)	(-.13, .01)	-.08	-	(-.09, -.04)	-.09				
μ(0,1,1,1)	.04	-.24	(-.35, -.13)	(-.25, -.03)	-.07	-	(-.08, -.01)	-.09				
μ(1,0,1,1)	.01	.08	(-.17, .32)	(-.05, .33)	-.08	-	(-.09, -.04)	-.1				
μ(1,1,0,1)	.01	-.12	(-.35, .10)	(-.37, .06)	-.07	-	(-.11, -.03)	-.08				
μ(1,1,1,0)	.06	-.05	(-.17, .06)	(-.11, .10)	-.08	-	(-.09, -.02)	-.1				
μ(1,1,1,1)	.04	[-0.41, 0]	(-.47, 0)	(-.48, 0)	[-.08, -.08]	-	(-.09, -.04)	-.09				
μ		[-.36, -.05]	(-.44, .01)	(-.36, -.03)	[-.07, -.07]	-	(-.08, -.04)	-.08	-.08	-.09	-.08	-.09
								(-.10, -.07)	(-.10, -.06)	(-.11, -.07)	(-.10, -.06)	(-.11, -.07)
							Probit					
β					-.44	-	(-.47, -.26)	-.71	-.55	-.46		
								(-.86, -.56)	(-.70, -.39)	(-.61, -.31)		
μ(0,0,0,0)	.37	[-.80, 0]	(-.82, 0)	(-.80, 0)	[-.08, -.07]	-	(-.08, -.04)	-.08				
μ(0,0,0,1)	.04	-.2	(-.33, -.07)	(-.30, -.09)	-.06	-	(-.08, -.02)	-.05				
μ(0,0,1,0)	.03	-.04	(-.17, .10)	(-.22, .03)	-.07	-	(-.09, -.03)	-.08				
μ(0,1,0,0)	.05	-.09	(-.20, .02)	(-.14, .03)	-.06	-	(-.08, -.03)	-.08				
μ(1,0,0,0)	.13	-.08	(-.15, -.01)	(-.11, .00)	-.08	-	(-.09, -.04)	-.1				
μ(0,0,1,1)	.05	-.14	(-.23, -.05)	(-.25, -.07)	-.07	-	(-.08, -.04)	-.08				
μ(0,1,0,1)	.01	-.09	(-.31, .13)	(-.40, .05)	-.11	-	(-.12, -.05)	-.12				
μ(1,0,0,1)	.01	-.06	(-.29, .18)	(-.25, .14)	-.08	-	(-.11, -.03)	-.1				
μ(0,1,1,0)	.04	-.06	(-.16, .05)	(-.23, -.05)	-.08	-	(-.10, -.04)	-.1				
μ(1,0,1,0)	.01	.03	(-.17, .22)	(-.12, .22)	-.06	-	(-.08, -.02)	-.07				
μ(1,1,0,0)	.11	-.08	(-.15, -.00)	(-.13, .01)	-.08	-	(-.09, -.05)	-.09				
μ(0,1,1,1)	.04	-.24	(-.35, -.13)	(-.25, -.03)	-.08	-	(-.10, -.04)	-.09				
μ(1,0,1,1)	.01	.08	(-.17, .32)	(-.05, .33)	-.08	-	(-.12, -.03)	-.1				
μ(1,1,0,1)	.01	-.12	(-.35, .10)	(-.37, .06)	-.07	-	(-.09, -.03)	-.08				
μ(1,1,1,0)	.06	-.05	(-.17, .06)	(-.11, .10)	-.08	-	(-.10, -.05)	-.1				
μ(1,1,1,1)	.04	[-0.41, 0]	(-.47, 0)	(-.48, 0)	[-.08, -.08]	-	(-.10, -.04)	-.09				
μ		[-.36, -.05]	(-.44, .01)	(-.36, -.03)	[-.08, -.07]	-	(-.08, -.04)	-.08	-.08	-.09		-.09
								(-.10, -.07)	(-.10, -.06)	(-.11, -.07)		(-.11, -.07)

Notes: Dependent variable is labor force participation indicator; regressor is a fertility indicator that takes the value 1 if the woman has a child less than 3 years old. Time periods: 1990, 1992, 1994, and 1996. Source: NLSY79.

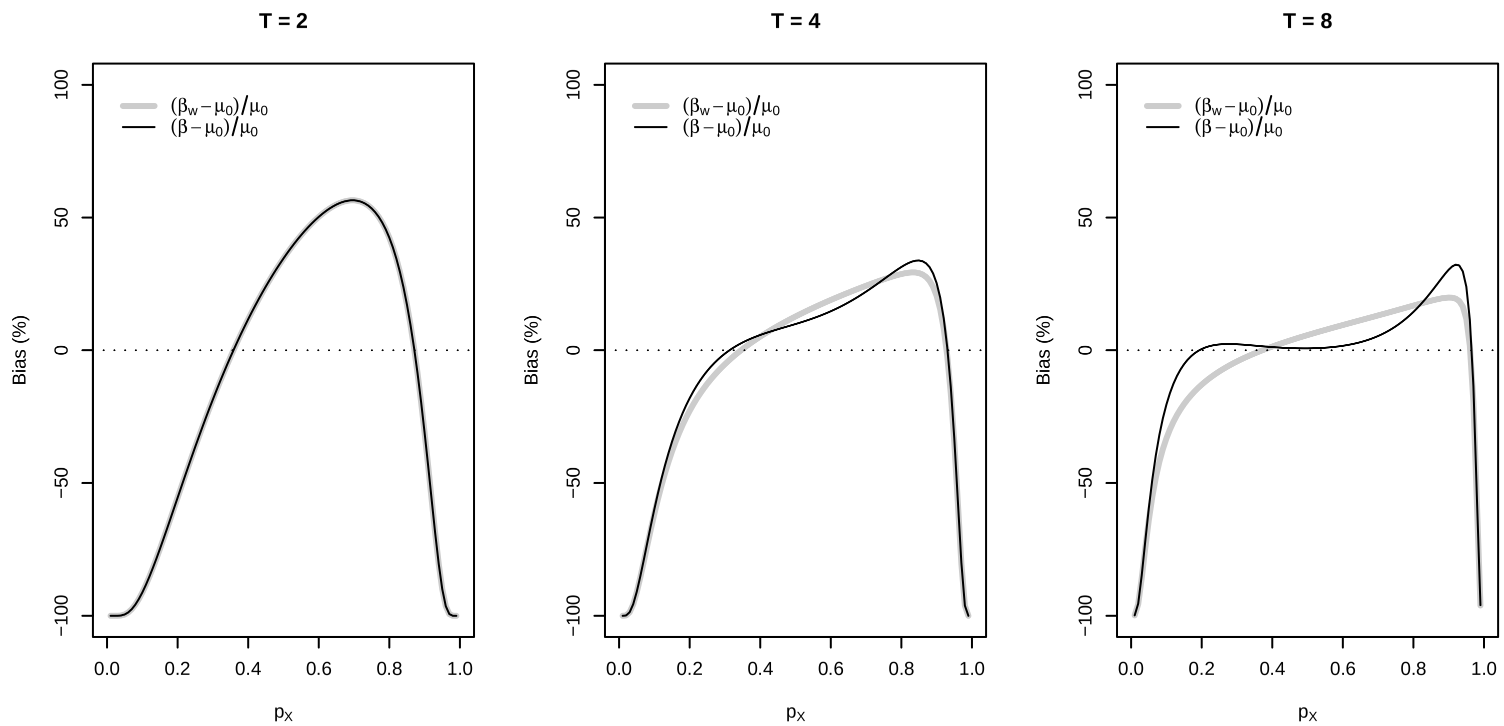

Figure 1: Biases of linear probability model estimators in percentage of marginal effect. Probit model with a single binary regressor with parameter equal to one and individual location effect. Individual effect is the standardized individual mean of the regressor.

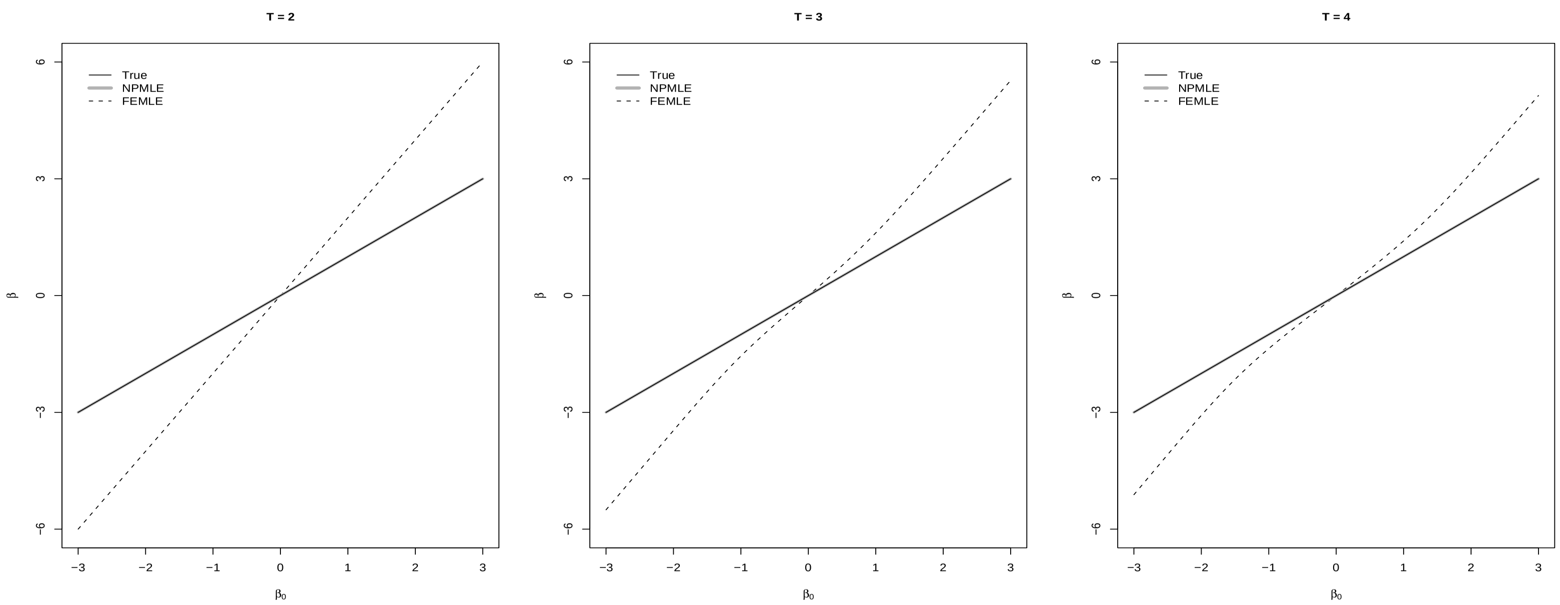

Figure 2: Logit model: Nonparametric MLE identification sets for model parameter β_0 and probability limits of fixed effects estimators.

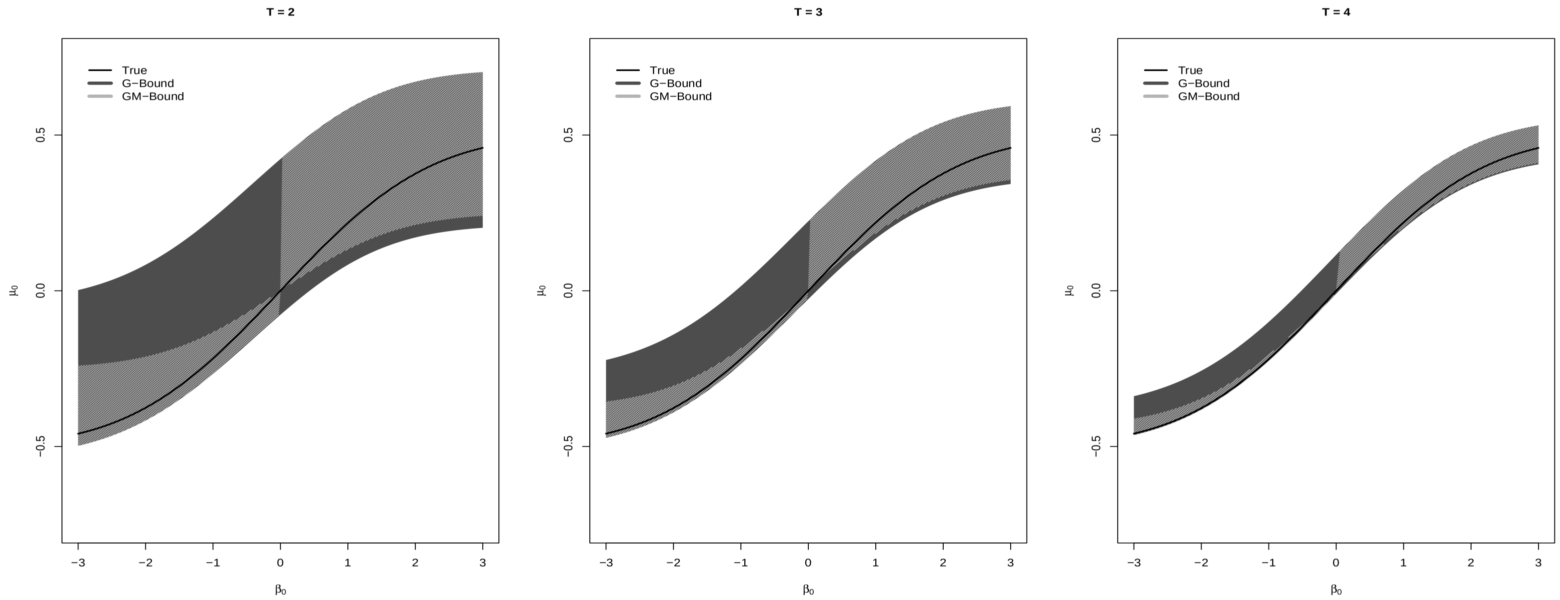

Figure 3: Logit model: Identification sets for average marginal effects μ_0 based on general model. G-bounds are obtained using equation (7) and GM-bounds impose monotonicity of the marginal effects.

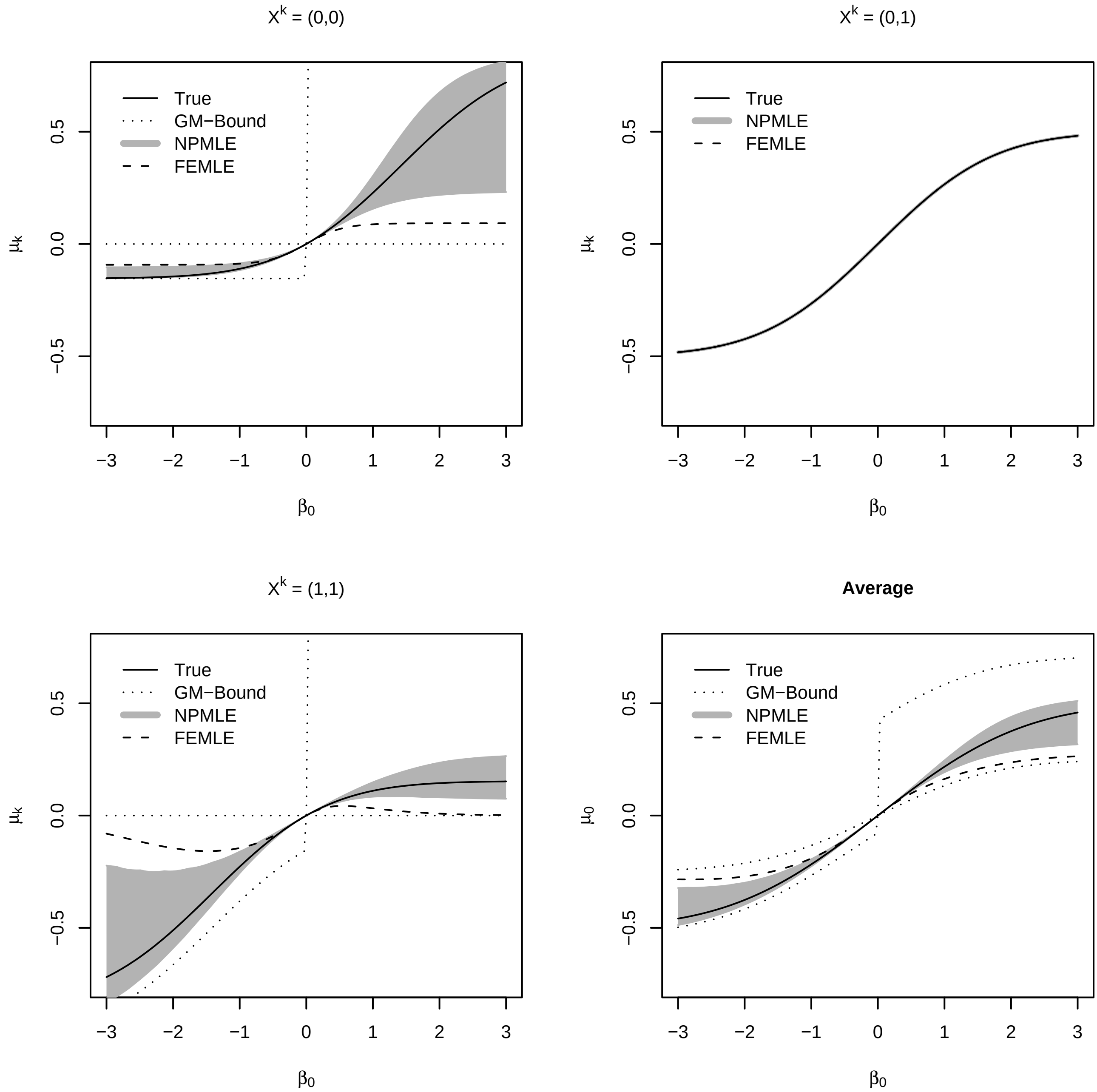

Figure 4: Logit model (T = 2). Identification sets for marginal effects and probability limits of fixed effects estimators.

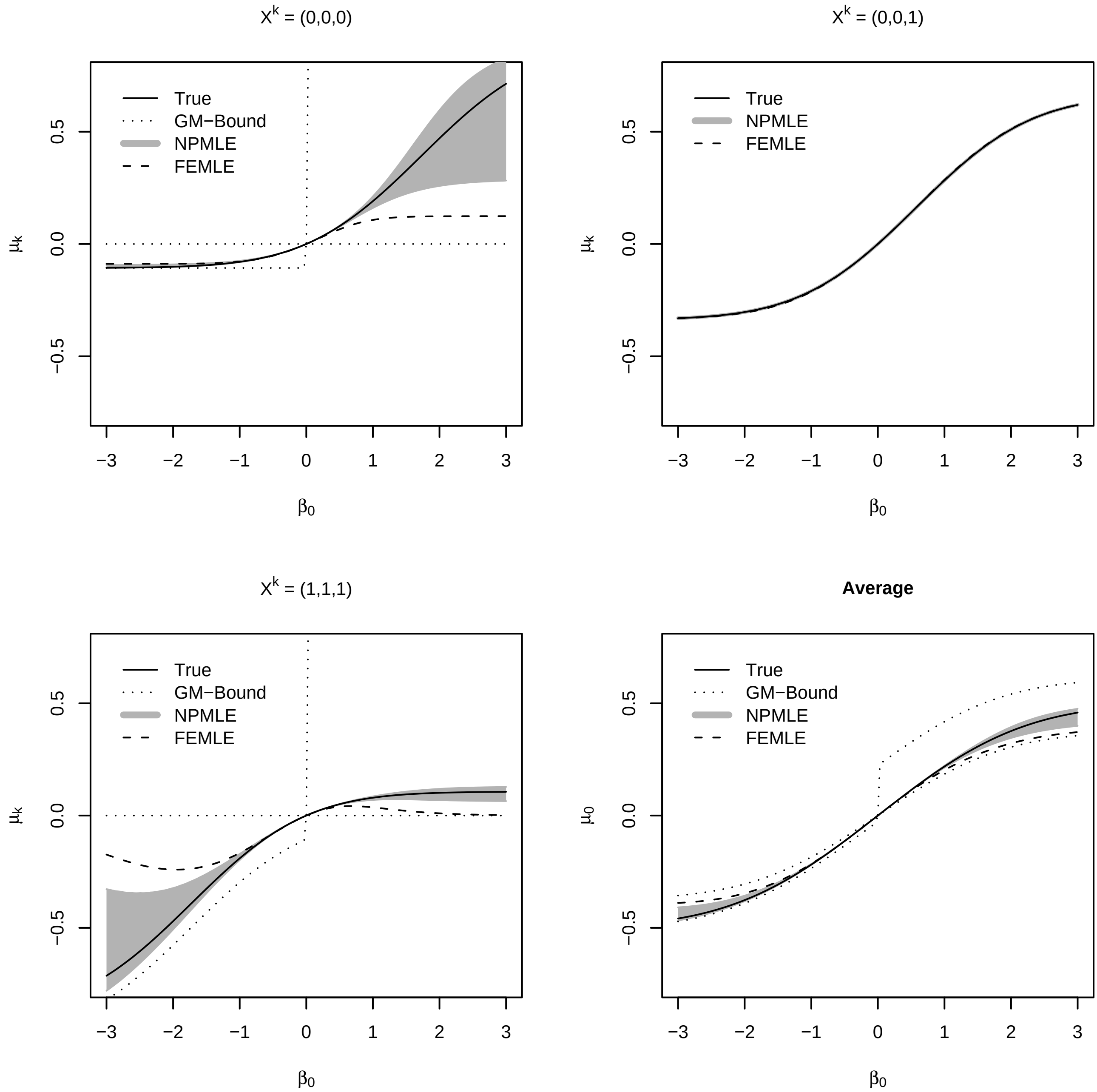

Figure 5: Logit model (T = 3). Identification sets for marginal effects and probability limits of fixed effects estimators.

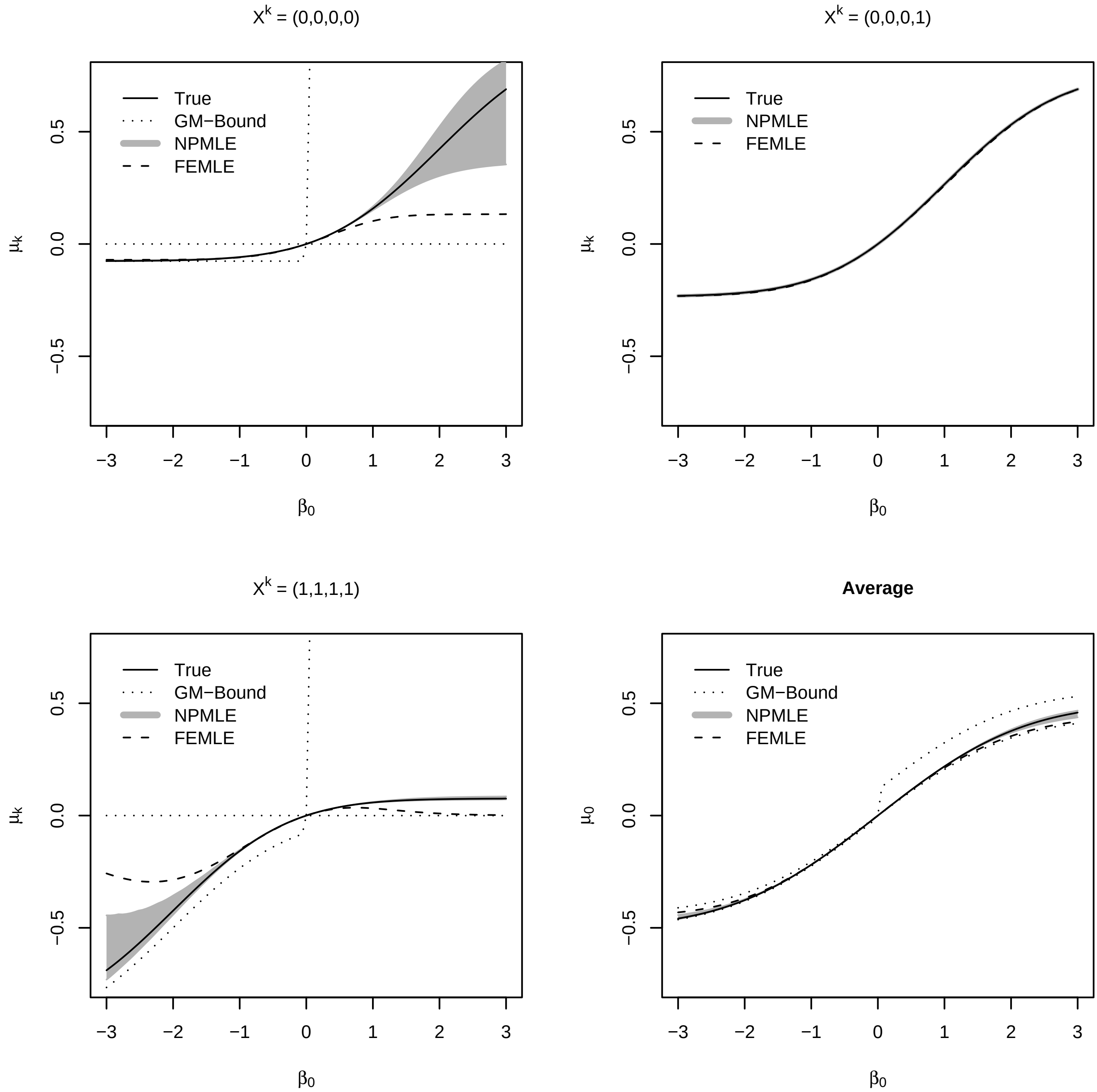

Figure 6: Logit model (T = 4). Identification sets for marginal effects and probability limits of fixed effects estimators.

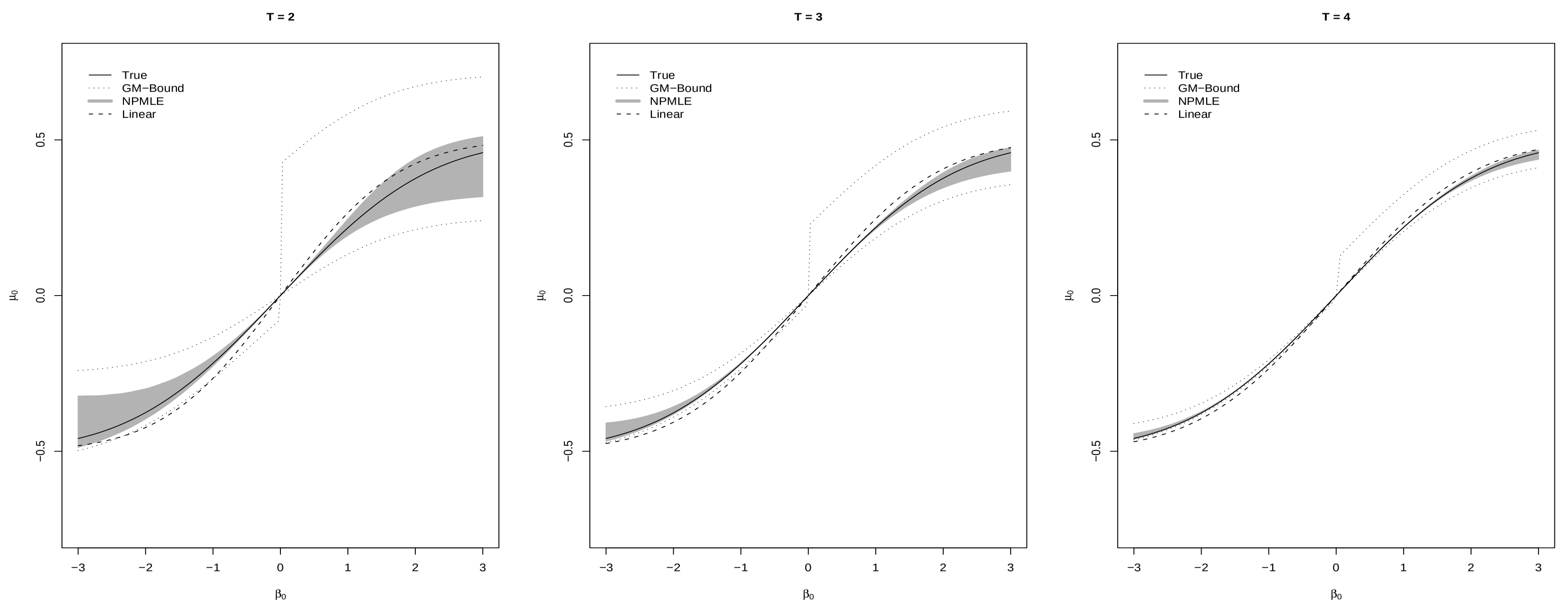

Figure 7: Logit model: Identification sets for average marginal effects and probability limits of linear model estimators.

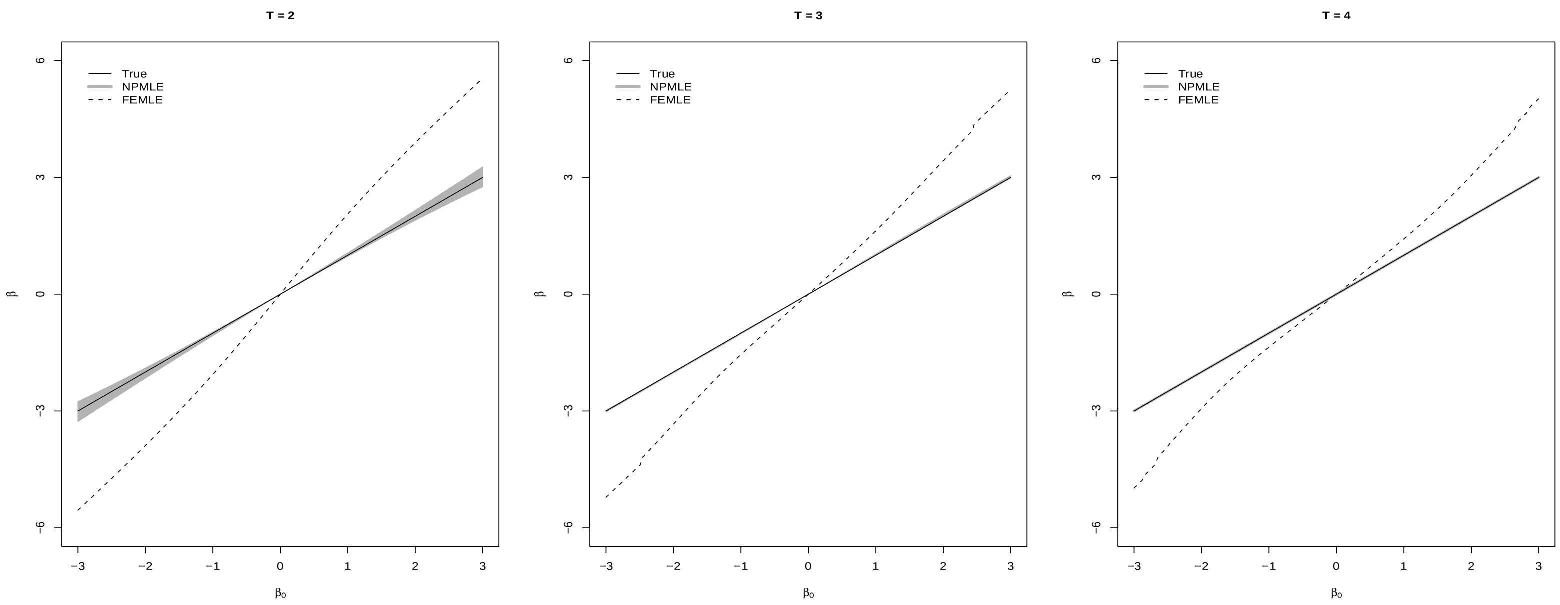

Figure 8: Probit model: Nonparametric MLE identification sets for model parameter β_0 and probability limits of fixed effects estimators.

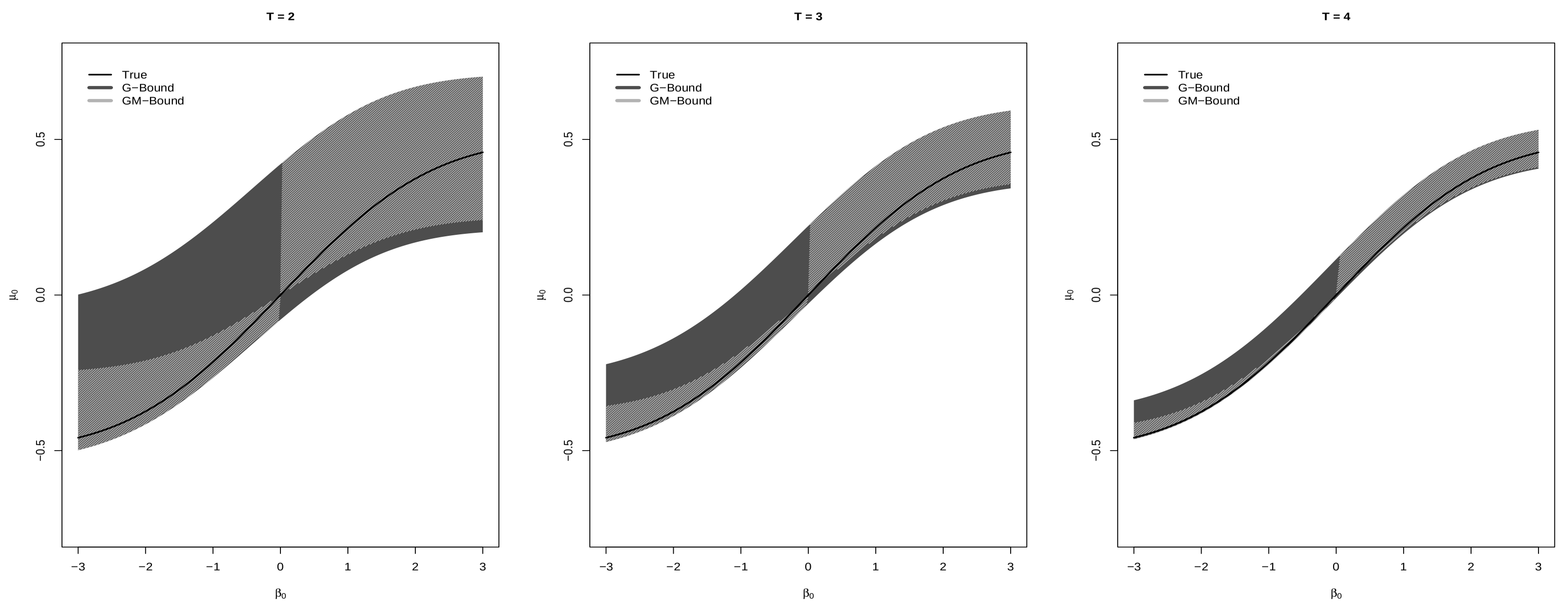

Figure 9: Probit model: Identification sets for average marginal effects μ_0 based on general model. G-bounds are obtained using equation (7) and GM-bounds impose monotonicity of the marginal effects.

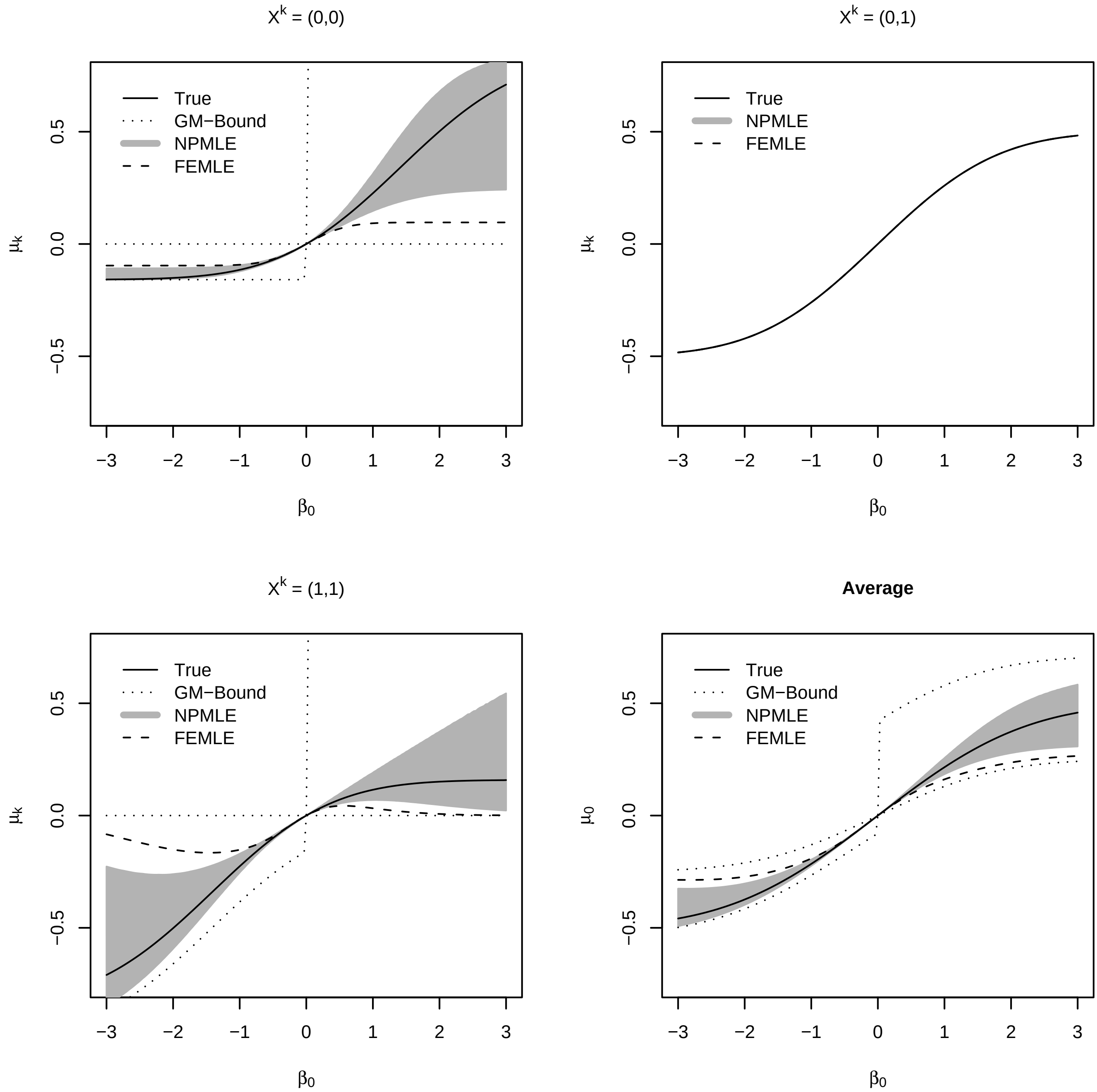

Figure 10: Probit model (T = 2). Identification sets for marginal effects and probability limits of fixed effects estimators.

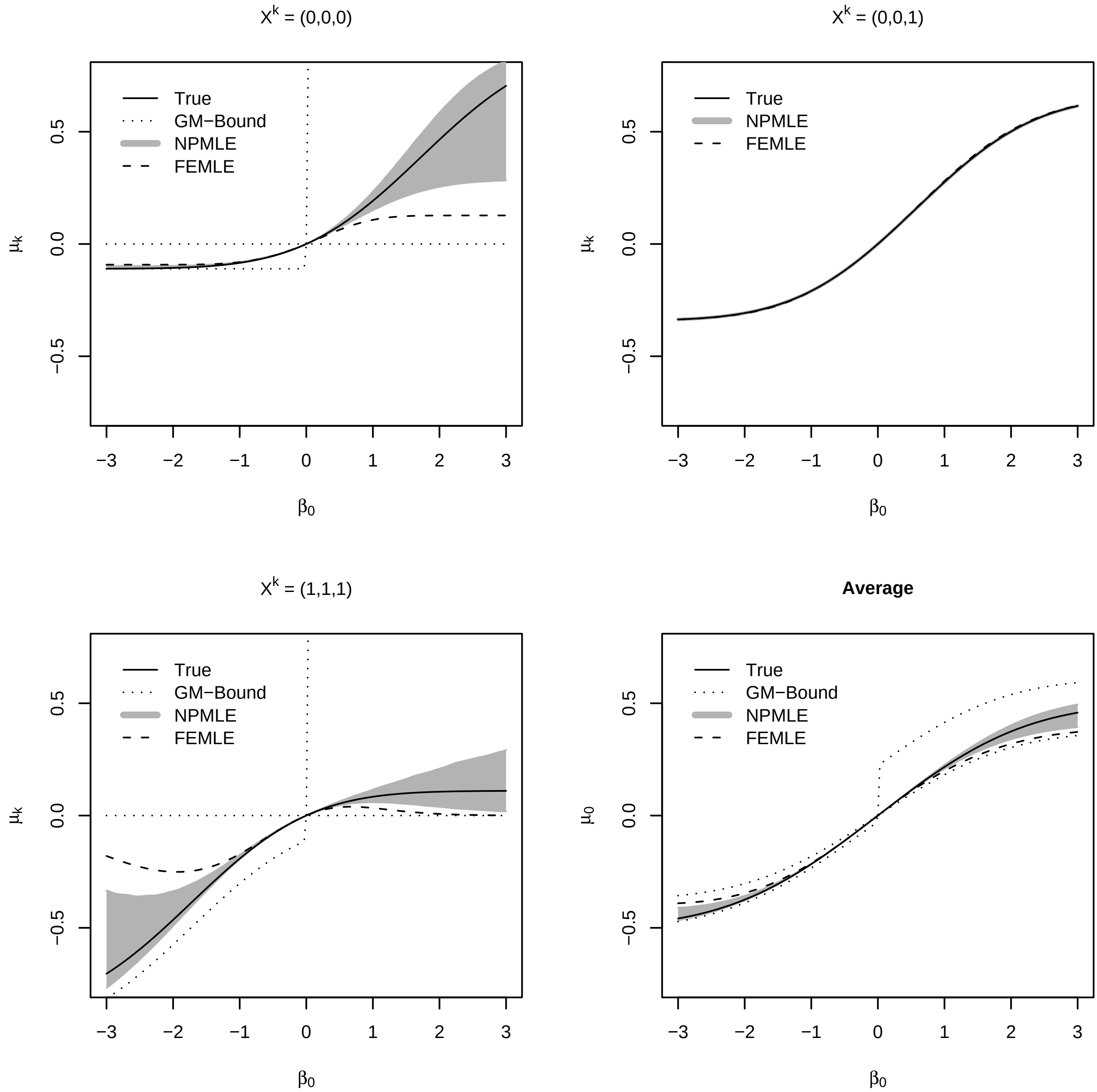

Figure 11: Probit model (T = 3). Identification sets for marginal effects and probability limits of fixed effects estimators.

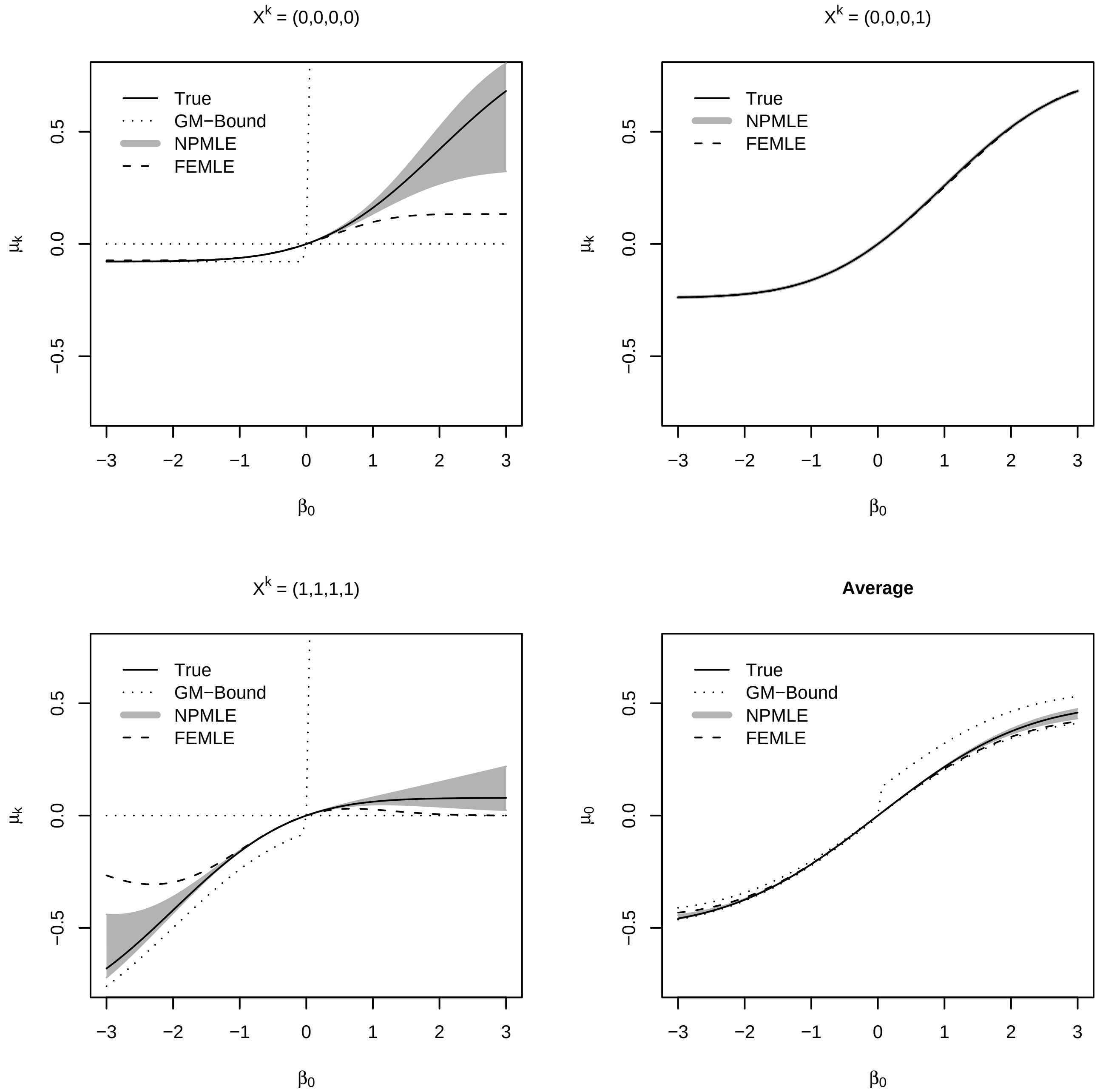

Figure 12: Probit model (T = 4). Identification sets for marginal effects and probability limits of fixed effects estimators.

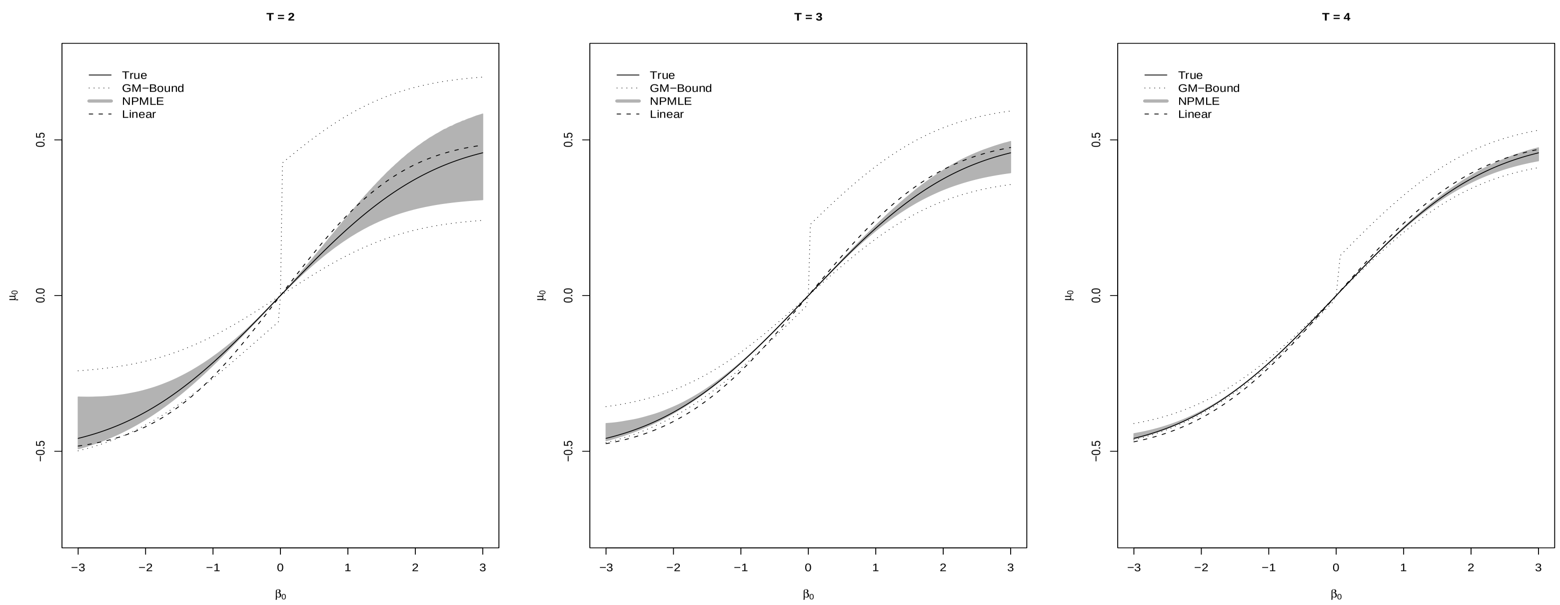

Figure 13: Probit model: Identification sets for average marginal effects and probability limits of linear model estimators.