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

Victor Chernozhukov
MIT

Iván Fernández-Val
BU

Jinyong Hahn
UCLA

Whitney Newey
MIT

February 4, 2009

[1]First version of May 2007. We thank J. Angrist, B. Graham, and seminar participants of Brown University, CEMFI, CEMMAP Microeconometrics: Measurement Matters Conference, CEMMAP Inference in Partially Identified Models with Applications Conference, CIREQ Inference with Incomplete Models Conference, Georgetown, Harvard/MIT, MIT, UC Berkeley, USC, 2007 WISE Panel Data Conference, and 2009 Winter Econometric Society Meetings 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 propose two novel inference methods for parameters defined as solutions to linear and nonlinear programs such as marginal effects in multinomial choice models. We show that these methods produce uniformly valid confidence regions in large samples. We give an empirical illustration.

1 Introduction

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 variable. Our paper gives identification and estimation results for marginal effects in panel data under time stationarity and discrete regressors.

It is sometimes thought that marginal effects can be estimated using linear fixed effects, as shown by Hahn (2001) in an example and Wooldridge (2005) under strong independence conditions. It turns out that the situation is more complicated. The marginal effect may not be identified. Furthermore, with a binary regressor, the linear fixed effects estimator 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, extending a result of Chamberlain (1982). We also derive nonparametric bounds for the marginal effect when it is not identified and when regressors are either exogenous or predetermined conditional on individual effects.

The nonparametric bounds are quite simple to compute and to use for inference but can be quite wide when T is small. We also consider bounds in semiparametric multinomial choice models where the form of the conditional probability given regressors and individual effects is specified. We find that the semiparametric bounds can be quite tight in binary choice models with additive heterogeneity.

We also give theorems showing that the bounds can tighten quickly as T grows. We find that the nonparametric bounds tighten exponetially fast when conditional probabilities of certain regressor values are bounded away from zero. We also find that in a semiparametric logit model the bounds tighten nearly that fast without any restriction on the distribution of regressors.

These results suggest how the bounds can be used in practice. For large T the nonparametric bounds may provide useful information. For small T, bounds in semiparametric models may be quite tight. Also, the tightness of semiparametric bounds 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.

We also develop estimation and inference methods for semiparametric multinomial choice models. The inferential problem is rather challenging. Indeed, the programs that characterize the population bounds on model parameters and marginal effects are very difficult to use for

inference, since the data-dependent constraints are often infeasible in finite samples or under misspecification, which produces empty set estimates and confidence regions. We overcome these difficulties by projecting these data-dependent constraints onto the model space, thus producing an always feasible data-dependent constraint set. We then propose linear and nonlinear programming methods that use these new modified constraints. Our inference procedures have the appealing justification of targeting the true model under correct specification and targeting the best approximating model under incorrect specification. We develop two novel inferential procedures, one called *modified projection* and another *perturbed bootstrap*, that produce uniformly valid inference in large samples. These methods may be of substantial independent interest.

This paper builds on 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 instead focus on marginal effects and give results on the rate of convergence of bounds as T grows. Moreover, the identification results in Honoré and Tamer (2006) and Chernozhukov, Hahn, and Newey (2004) characterize the bounds via linear and non-linear programs, and thus, for the reasons we stated above, they cannot be immediately used for practical estimation and inference. We propose new methods for estimation and inference, which are practical and which can be of interest in other problems, and we illustrate them with an empirical application.

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. Chamberlain (1982) gives conditions for consistent estimation of marginal effects in linear correlated random coefficient models. Graham and Powell (2008) extend the analysis of Chamberlain (1982) by relaxing some of the regularity conditions in models with continuous regressors.

In semiparametric binary choice models 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, nonlinear 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

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

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

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

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

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

This model has been studied by Chamberlain (1980, 1984, 1992), Hahn and Newey (2004), and others. The familiar linear model $E[Y_{it} \mid X_i, \alpha_i] = X_{it}'\beta^* + \alpha_i$ is also included as a special case of equation (1).

For binary $X_{it} \in \{0, 1\}$ the model of equation (1) reduces to the correlated random coefficients model of Chamberlain (1982). For other X_{it} with finite support that does not vary with t it is a multiple regression version of that 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. We consider identification without the strict exogeneity assumption below. 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 $m(x, \alpha) = F(x'\beta^* + \alpha)$. 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 shown by Chamberlain (1982) for binary X_{it}. Second, the fixed effects estimator uses incorrect weighting.

To explain, we compare the limit of the usual linear fixed effects estimator 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

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

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 the linear fixed effect estimator 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 $r^k = \#\{t : X_t^k = 1\}/T$ 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*

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

This result is similar to Angrist (1998) who found that, in a treatment effects model in cross section data, the partially linear slope estimator is a variance weighted average effect. 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 how μ_k varies with k is restricted. Imposing restrictions on how μ_k varies with k amounts to restricting the conditional distribution of α_i given X_i, which we are not doing in this paper.

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). Nonidentification of these effects was pointed out by Chamberlain (1982).

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. Hahn (2001) restricted the support of X_i to exclude both $(0, ..., 0)'$ or $(1, ..., 1)'$ and only considered a case with $T = 2$. Thus, neither feature that causes inconsistency of $\hat{\beta}_w$ was present in that example. As noted by Hahn (2001), the conditions that lead to consistency of the linear fixed effects estimator in his example are quite special.

Theorem 1 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 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, when $m(x, \alpha)$ is bounded there are informative bounds for μ_0, as we show below.

The second reason for inconsistency of $\hat{\beta}_w$ can be corrected by modifying the estimator. In the binary X_{it} case Chamberlain (1982) gave a consistent estimator for the identified effect $\mu_I = \sum_{k=2}^{K-1} \mathcal{P}_k \mu_k / \sum_{k=2}^{K-1} \mathcal{P}_k$. The 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.

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 extends Chamberlain's (1982) estimator to the case where X_{it} is not binary.

To describe the limit of the estimator $\hat{\beta}$ in general, let $\mathcal{K}^* = \{k$:there is $\tilde{t}$ and $\bar{t}$ such that $X_{\tilde{t}}^k = \tilde{x}$ and $X_{\bar{t}}^k = \bar{x}\}$. 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 2: *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 $\mathcal{P}_k^* = \mathcal{P}_k / \sum_{k\in\mathcal{K}^*} \mathcal{P}_k$.

Here $\hat{\beta}$ is not an efficient estimator of μ_I for $T \geq 3$, because $\hat{\beta}$ 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 see how big the inconsistency of the linear estimators 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. $N(0, 1)$,

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

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 $(\beta_w - \mu_0)/\mu_0$ and $(\beta - \mu_0)/\mu_0$ for several values of T and p_X, where $\beta_w = \text{plim } \hat{\beta}_w$ and $\beta = \text{plim } \hat{\beta}$.

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.

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

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

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 3: *Suppose that equation (1) is satisfied. If there is $\tilde{t}$ and $\bar{t}$ such that $X_{\tilde{t}}^k = \tilde{x}$ and $X_{\bar{t}}^k = \bar{x}$ then*

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

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

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

Also, if there is $\bar{t}_k$ such that $X_{\bar{t}}^k = \bar{x}$ then

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

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

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

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

THEOREM 4*: If equation (1) is satisfied and $B_\ell \leq m(x,\alpha)/D \leq B_u$ then $\mu_\ell \leq \mu_0 \leq \mu_u$ for*

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

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

An estimator can be constructed by replacing the probabilities by sample proportions $P_k = \sum_i 1(X_i = X^k)/n$ and $P^0 = 1 - \sum_{k\in\tilde{\mathcal{K}}} P_k - \sum_{k\in\bar{\mathcal{K}}} P_k - \sum_{k\in\mathcal{K}^*} P_k$, and each $\bar{m}_t^k$ by

$$\hat{m}_t^k = 1(n^k > 0) \sum_{i=1}^n 1(X_i = X^k) Y_{it}/n^k, n^k = \sum_{i=1}^n 1(X_i = X^k).$$

Estimators of the upper and lower bound respectively are given by

$$\begin{aligned}
\hat{\mu}_\ell &= P^0(B_\ell - B_u) + \sum_{k\in\tilde{\mathcal{K}}} P_k(\hat{m}_t^k - B_u) + \sum_{k\in\bar{\mathcal{K}}} P_k(B_\ell - \hat{m}_t^k) + (n^*/n)\hat{\beta}, \\
\hat{\mu}_u &= P^0(B_u - B_\ell) + \sum_{k\in\tilde{\mathcal{K}}} P_k(\hat{m}_t^k - B_\ell) + \sum_{k\in\bar{\mathcal{K}}} P_k(B_u - \hat{m}_t^k) + (n^*/n)\hat{\beta}.
\end{aligned}$$

The bounds $\hat{\mu}_\ell$ and $\hat{\mu}_u$ will be jointly asymptotically normal with variance matrix that can be estimated in the usual way, so that set inference can be carried out as described in Chernozhukov, Hong, and Tamer (2007), or Beresteanu and Molinari (2008).

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

$$\begin{aligned}
\mu_\ell &= \mathcal{P}_K(\bar{m}_t^K - B_u) + \mathcal{P}_1(B_\ell - \bar{m}_t^1) + \sum_{1<k<K} \mathcal{P}_k\mu_k, \\
\mu_u &= \mathcal{P}_K(\bar{m}_t^K - B_\ell) + \mathcal{P}_1(B_u - \bar{m}_t^1) + \sum_{1<k<K} \mathcal{P}_k\mu_k.
\end{aligned} \tag{6}$$

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 5: *If equation (1) is satisfied, $B_\ell \leq m(x,\alpha)/D \leq B_u$ $\overrightarrow{X}_i = (X_{i1}, X_{i2}, ...)$ 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 $T \longrightarrow \infty$.*

This result gives conditions for identification as T grows, generalizing a result of Chamberlain (1982) for binary X_{it}. In addition, it shows that the bounds derived above shrink to the marginal effect as T grows. 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 6: *If equation (1) is satisfied,* $B_\ell \leq m(x, \alpha)/D \leq B_u$ *and* $\overrightarrow{X_i}$ *is stationary and Markov of order* J *conditional on* α_i *then for* $p_i^1 = \Pr(X_{it} = 0 | X_{i,t-1} = \cdots = X_{i,t-J} = 0, \alpha_i)$ *and* $p_i^K = \Pr(X_{it} = 1 | X_{i,t-1} = \cdots = X_{i,t-J} = 1, \alpha_i)$

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

If there is $\varepsilon > 0$ *such that* $p_i^1 \leq 1 - \varepsilon$ *and* $p_i^K \leq 1 - \varepsilon$ *then*

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

If there is a set $\mathcal{A}$ *of* α_i *such that* $\Pr(\mathcal{A}) > 0$ *and either* $\Pr(X_{i1} = \cdots = X_{iJ} = 0 \mid \alpha_i) > 0$ *for* $\alpha_i \in \mathcal{A}$ *and* $p_i^1 = 1$ *for all* $\alpha_i \in \mathcal{A}$, *or* $\Pr(X_{i1} = \cdots = X_{iJ} = 1 \mid \alpha_i) > 0$ *for* $\alpha_i \in \mathcal{A}$ *and* $p_i^K = 1$, *then* $\mu_\ell \nrightarrow \mu_0$, *or* $\mu_u \nrightarrow \mu_0$.

When the conditional probabilities that X_{it} is zero or one are bounded away from one the bounds will converge at an exponential rate. We conjecture that an analogous result could be shown for general X_{it}. The conditions that imply that one of the bounds does not converge violates a hypothesis of Theorem 5, that the conditional support of X_{it} equals the marginal support. Theorem 6 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 α_i i.i.d. over i, and 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_i^1 and p_i^K.

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 Section 5 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 Predetermined Regressors

The previous bound analysis can be extended to cases where the regressor X_{it} is just predetermined instead of strictly exogenous. These cases cover, for example, dynamic panel models where X_{it} includes lags of Y_{it}. To describe this extension let $X_i(t) = [X_{i1}, ..., X_{it}]'$ and suppose that

$$E[Y_{it}|X_i(t), \alpha_i] = m(X_{it}, \alpha_i), \ (t = 1, ..., T). \tag{7}$$

For example, this includes the heterogenous, dynamic binary choice model of Browning and Carro (2007), where $Y_{it} \in \{0, 1\}$ and $X_{it} = Y_{i,t-1}$.

As before, the marginal effect is given by $\mu_0 = \int [m(\tilde{x}, \alpha) - m(\bar{x}, \alpha)]Q^*(d\alpha)/D$ for two different possible values $\tilde{x}$ and $\bar{x}$ of the regressors and a distance D. Also, as before, the marginal effect will have an identified component and an unidentified component. The key implication that is used to obtain the identified component is

$$E[Y_{it}|X_i(t) = X(t)] = \int m(X_t(t), \alpha)Q^*(d\alpha|X_i(t) = X(t)), \tag{8}$$

where $X(t) = [X_1, ..., X_t]'$.

Bounds are obtained by partitioning the set of possible X_i into subsets that can make use of the above key implication and a subset where bounds on $m(x, \alpha)/D$ are applied. The key implication applies to subsets of the form $\mathcal{X}^t(x) = \{X : \ X_t = x, \ X_s \neq x \ \forall s < t\}$, that is a set of possible X_i vectors that have x as the t^{th} component and not as any previous components. The bound applies to the same subset as before, that where x never appears, given by $\bar{\mathcal{X}}(x) = \{X : \ X_t \neq x \ \forall t\}$. Together the union of $\mathcal{X}^t(x)$ over all t and $\bar{\mathcal{X}}(x)$ constitute a partition of possible X vectors. Let $\bar{\mathcal{P}}(x) = \Pr(X_i \in \bar{\mathcal{X}}(x))$ be the probability that none of the components of X_i is equal to x and

$$\delta_0 = E[\sum_{t=1}^{T} \{1(X_i \in \mathcal{X}^t(\tilde{x})) - 1(X_i \in \mathcal{X}^t(\bar{x}))\}Y_{it}]/D.$$

Then the key implication and iterated expectations give

THEOREM 7*: If equation (7) is satisfied and* $B_\ell \leq m(x, \alpha)/D \leq B_u$ *then* $\mu_\ell \leq \mu_0 \leq \mu_u$ *for*

$$\mu_\ell = \delta_0 + B_\ell \bar{\mathcal{P}}(\tilde{x}) - B_u \bar{\mathcal{P}}(\bar{x}), \quad \mu_u = \delta_0 + B_u \bar{\mathcal{P}}(\tilde{x}) - B_\ell \bar{\mathcal{P}}(\bar{x}). \tag{9}$$

As previously, estimates of these bounds can be formed from sample analogs. Let $\bar{P}(x) = \sum_{i=1}^{n} 1(X_i \in \bar{\mathcal{X}}(x))/n$ and

$$\hat{\delta} = \sum_{i=1}^{n} \sum_{t=1}^{T} [1(X_i \in \mathcal{X}^t(\tilde{x})) - 1(X_i \in \mathcal{X}^t(\bar{x}))]Y_{it}/(nD).$$

The estimates of the bounds are given by

$$\hat{\mu}_\ell = \hat{\delta} + B_\ell \bar{P}(\tilde{x}) - B_u \bar{P}(\bar{x}), \quad \hat{\mu}_u = \hat{\delta} + B_u \bar{P}(\tilde{x}) - B_\ell \bar{P}(\bar{x}).$$

Inference using these bounds can be carried out analogously to the strictly exogenous case.

An important example is binary $Y_{it} \in \{0, 1\}$ where $X_{it} = Y_{i,t-1}$. Here $B_u = 1$ and $B_\ell = 0$, so the marginal effect is

$$\mu_0 = \int [\Pr(Y_{it} = 1 | Y_{i,t-1} = 1, \alpha) - \Pr(Y_{it} = 1 | Y_{i,t-1} = 0, \alpha)] Q^*(d\alpha),$$

i.e., the effect of the lagged $Y_{i,t-1}$ on the probability that $Y_{it} = 1$, holding α_i constant, averaged over α_i. In this sense the bounds provide an approximate solution to the problem considered by Feller (1943) and Heckman (1981) of evaluating duration dependence in the presence of unobserved heterogeneity. In this example the bounds estimates are

$$\hat{\mu}_\ell = \hat{\delta} - \bar{P}(0), \quad \hat{\mu}_u = \hat{\delta} + \bar{P}(1). \tag{10}$$

The width of the bounds is $\bar{P}(0) + \bar{P}(1)$, so although these bounds may not be very informative in short panels, in long panels, where $\bar{P}(0) + \bar{P}(1)$ is small, they will be.

Theorems 5 and 6 on convergence of the bounds as T grows apply to μ_ℓ and μ_u from equation (9), since the bounds have a similar structure and the convergence results explicitly allow for dependence over time of X_{it} conditional on α_i. For example, for $Y_{it} \in \{0, 1\}$ and $X_{it} = Y_{i,t-1}$, equation (7) implies that Y_{it} is Markov conditional on α_i with $J = 1$. Theorem 5 then shows that the bounds converge to the marginal effect as T grows if $0 < \Pr(Y_{it} = 1 | \alpha_i) < 1$ with probability one. Theorem 6 also gives the rate at which the bounds converge, e.g. that will be exponential if $\Pr(Y_{it} = 1 | Y_{i,t-1} = 1, \alpha_i)$ and $\Pr(Y_{it} = 0 | Y_{i,t-1} = 0, \alpha_i)$ are bounded away from one.

It appears that, unlike the strictly exogenous case, there is only one way to estimate the identified component δ_0. In this sense the estimators given here for the bounds should be asymptotically efficient, so there should be no gain in trying to account for heteroskedasticity and autcorrelation over time. Also, it does not appear possible to obtain tighter bounds when monotonicity holds, because the partition is different for $\tilde{x}$ and $\bar{x}$

5 Semiparametric Multinomial Choice

The bounds for marginal effects derived in the previous sections did not use any functional form restrictions on the conditional distribution of Y_i given (X_i, α_i). 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 unknown.

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

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

for some finite dimensional β^* and some known function $\mathcal{L}$. 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

$$\mathcal{P}_{jk} = \int \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta^*\right) Q_k^*\left(d\alpha\right), (j = 1, ..., J; k = 1, ..., K), \tag{11}$$

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 $\mathcal{L}$ that is parametric and conditional distributions Q_k^* 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 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

$$\Pr(Y_{it} = 1 \mid X_i, \alpha_i, \beta) = F(X_{it}'\beta + \alpha_i) \tag{12}$$

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,

$$\mathcal{L}\left(Y_i \mid X_i, \alpha_i, \beta\right) = \prod_{t=1}^{T} F(X_{it}'\beta + \alpha_i)^{Y_{it}}[1 - F(X_{it}'\beta + \alpha_i)]^{1-Y_{it}}.$$

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

$$\mu_k = D^{-1} \int [F\left(\tilde{x}'\beta^* + \alpha\right) - F\left(\bar{x}'\beta^* + \alpha\right)] Q_k^*(d\alpha), \tag{13}$$

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, so we can apply the result in Lemma 3 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

$$\begin{array}{rl} \underline{\mu}_k = & \min_{\beta \in B, Q_k} D^{-1} \int [F(\tilde{x}'\beta + \alpha) - F(\bar{x}'\beta + \alpha)] Q_k(d\alpha) \\ & \text{s.t. } \mathcal{P}_{jk} = \int \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right) Q_k(d\alpha) \ \forall j, \end{array} \tag{14}$$

and

$$\begin{array}{rl} \overline{\mu}_k = & \max_{\beta \in B, Q_k} D^{-1} \int [F(\tilde{x}'\beta + \alpha) - F(\bar{x}'\beta + \alpha)] Q_k(d\alpha) \\ & \text{s.t. } \mathcal{P}_{jk} = \int \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right) Q_k(d\alpha) \ \forall j. \end{array} \tag{15}$$

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 the 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

$$\tilde{\beta} = \text{argmax}_\beta \sum_{k=1}^{K} \mathcal{P}_k \sum_{j=1}^{J} \mathcal{P}_{jk} \log \mathcal{L}\left(Y^j \mid X^k, \alpha_{jk}(\beta), \beta\right), \tag{16}$$

where

$$\alpha_{jk}(\beta) = \text{argmax}_\alpha \log \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right), \forall j, k. \tag{17}$$

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

$$\tilde{Q}_{k\beta}(\alpha) = \begin{cases} \mathcal{P}_{jk}, & \text{for } \alpha = \alpha_{jk}(\beta); \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

The following example illustrates this point using a simple two period model. Consider a two-period binary choice model with binary regressor and strictly increasing and symmetric CDF, i.e., $F(-x) = 1 - F(x)$. In this case the estimand of the fixed effects estimators are

$$\alpha_{jk}(\beta) = \begin{cases} -\infty, & \text{if } Y^j = (0,0); \\ -\beta(X_1^k + X_2^k)/2, & \text{if } Y^j = (1,0) \text{ or } Y^j = (0,1); \\ \infty, & \text{if } Y^j = (1,1), \end{cases} \tag{19}$$

and the corresponding distribution for α has the form

$$\tilde{Q}_{k\beta}(\alpha) = \begin{cases} \Pr\{Y = (0,0) \mid X^k\}, & \text{if } \alpha = -\infty; \\ \Pr\{Y = (1,0) \mid X^k\} + \Pr\{Y = (0,1) \mid X^k\}, & \text{if } \alpha = -\beta(X_1^k + X_2^k)/2; \\ \Pr\{Y = (1,1) \mid X^k\}, & \text{if } \alpha = \infty. \end{cases} \tag{20}$$

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

$$\tilde{\mu}_k(\beta) = D^{-1} \int [F(\tilde{x}'\beta + \alpha) - F(\bar{x}'\beta + \alpha)] \tilde{Q}_{k\beta}(\alpha). \tag{21}$$

The average of these estimates across individuals with identified effects is consistent for the identified effect μ_I when X is binary. This result is shown here analytically for the two-period case and through numerical examples for $T \geq 3$.

THEOREM 9: *If* $F'(x) > 0$, $F(-x) = 1 - F(x)$, *and* $\sum_{k=1}^{K-1} \mathcal{P}_k > 0$, *then, for* $\mathcal{P}_k^* = \mathcal{P}_k / \sum_{k=2}^{K-1} \mathcal{P}_k$,

$$\tilde{\mu}_I = \sum_{k=2}^{K-1} \mathcal{P}_k^* \tilde{\mu}_k(\tilde{\beta}) = \mu_I.$$

For not identified effects the nonlinear fixed effects estimators are usually biased toward zero, introducing bias of the same direction in the fixed effect estimator of the average effect μ_0 if there are individuals with not identified effects in the population. 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

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

This conjecture is further explored numerically in the next section.

6 Characterization and Computation of Population Bounds

6.1 Identification Sets and Extremal Distributions

We will begin our discussion of calculating bounds by considering bounds for the parameter β. Let $\mathcal{L}_{jk}(\beta, Q_k) := \int \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right) Q_k(d\alpha)$ and $Q := (Q_1, \ldots, Q_K)$. For the subsequent inferential analysis, it is convenient to consider a quadratic loss function

$$T(\beta, Q; \mathcal{P}) = \sum_{j,k} \omega_{jk}(\mathcal{P}) \left(\mathcal{P}_{jk} - \mathcal{L}_{jk}(\beta, Q_k)\right)^2, \tag{22}$$

where $\omega_{jk}(\mathcal{P})$ are positive weights. By the definition of the model in (11), we can see that (β^*, Q^*) is such that

$$T(\beta, Q; \mathcal{P}) \geq T(\beta^*, Q^*; \mathcal{P}) = 0,$$

for every (β, Q). For $T(\beta; \mathcal{P}) := \inf_Q T(\beta, Q; \mathcal{P})$, this implies that

$$T(\beta; \mathcal{P}) \geq T(\beta^*; \mathcal{P}) = 0,$$

for every β. Let B be the set of β's that minimizes $T(\beta; \mathcal{P})$, i.e.,

$$B := \{\beta : T(\beta; \mathcal{P}) = 0\}\,.$$

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

It follows from the following lemma that one needs only to search over discrete distributions for Q to find B. Note that

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

$$Q_{k\beta} = \arg\min_{Q_k} \sum_{j=1}^{J} \omega_{jk}(\mathcal{P}) \left(\mathcal{P}_{jk} - \mathcal{L}_{jk}(\beta, Q_k)\right)^2$$

exists that is a discrete distribution with at most J points of support, and $\mathcal{L}_{jk}(\beta, Q_{k\beta}) = \mathcal{P}_{jk}$, $\forall j, k$.

Another important result is that the bounds for marginal effects can be also found by searching over discrete distributions with few points of support. We will focus on the upper bound $\overline{\mu}_k$ defined in (15); an analogous result holds for the lower bound $\underline{\mu}_k$ in (14).

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

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

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

6.2 Numerical Examples

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

$$Y_{it} = \mathbf{1}\{X_{it}\beta^* + \alpha_i + \varepsilon_{it} \geq 0\}, \tag{23}$$

with ε_{it} i.i.d. over t normal or logistic with zero mean and unit variance. The explanatory variable X_{it} is binary and i.i.d. over t with $p_X = \Pr\{X_{it} = 1\} = 0.5$. The unobserved individual effect α_i is correlated with the explanatory variable for each individual. In particular, we generate this effect 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

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

and $\alpha_{2i} = \sqrt{T}(\bar{X}_i - p_X)/\sqrt{p_X(1-p_X)}$ with $\bar{X}_i = \sum_{t=1}^{T} X_{it}/T$.

Identified sets for parameters and marginal effects are calculated for panels with 2, 3, and 4 periods based on the conditional mean model of Section 2 and semiparametric logit and probit models. For logit and probit models the sets are obtained using a linear programming algorithm for discrete regressors, as in Honoré and Tamer (2006). Thus, for the parameter we have that

$B = \{\beta : L(\beta) = 0\}$, where

$$
\begin{aligned}
L(\beta) \quad = \quad & \min_{w_k, v_{jk}, \pi_{km}} \sum_{k=1}^{K} w_k + \sum_{j=1}^{J}\sum_{k=1}^{K} v_{jk} \qquad (24) \\
& v_{jk} + \textstyle\sum_{m=1}^{M} \pi_{km} \mathcal{L}\left(Y^j \mid X^k, \alpha_m, \beta\right) = \mathcal{P}_{jk} \quad \forall j, k, \\
& w_k + \textstyle\sum_{m=1}^{M} \pi_{km} = 1 \quad \forall k, \\
& v_{jk} \geq 0, w_k \geq 0, \pi_{km} \geq 0 \quad \forall j, k, m.
\end{aligned}
$$

For marginal effects, see also Chernozhukov, Hahn, and Newey (2004), we solve

$$
\begin{aligned}
\overline{\mu}_k / \underline{\mu}_k \quad = \quad & \max / \min_{\pi_{km}, \beta \in B} \sum_{m=1}^{M} \pi_{km} [F(\tilde{x}'\beta + \alpha_m) - F(x'\beta + \alpha_m)] \qquad (25) \\
& \textstyle\sum_{m=1}^{M} \pi_{km} \mathcal{L}\left(Y^j \mid X^k, \alpha_m, \beta\right) = \mathcal{P}_{jk} \quad \forall j, \\
& \textstyle\sum_{m=1}^{M} \pi_{km} = 1, \pi_{km} \geq 0 \quad \forall j, m.
\end{aligned}
$$

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

Figure 2 shows identified sets for the slope 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 $\tilde{\beta} = c\beta^*$ 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).[1] 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 (6), and imposing the monotonicity restriction on $\Delta(\alpha)$ in Lemma 3 (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 nonlinear fixed effects estimators are consistent for these effects. This numerical finding suggests that the consistency result for nonlinear fixed effects estimators extends to more than two periods. Unless $\beta^* = 0$, marginal effects for individuals without switches in the regressor are not point identified, which also precludes point identification of the average effect. Nonlinear 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

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

the size of the identified sets and the asymptotic biases of these 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.

7 Estimation

7.1 Minimum Distance Estimator

In multinomial models with discrete regressors the complete description of the DGP is provided by the parameter vector

$$(\Pi', \Pi^{X'})', \ \Pi = (\Pi_{jk}, j = 1, ..., J, k = 1, ..., K)'^X = (\Pi_k, k = 1, ..., K)',$$

where

$$\Pi_{jk} = \Pr(Y = Y^j | X = X^k), \quad \Pi_k = \Pr(X = X^k).$$

We denote the true value of this parameter vector by $(\mathcal{P}', \mathcal{P}^{X'})'$, and the nonparametric empirical estimates by $(P', P^{X'})'$. As it is common in regression analysis, we condition on the observed distribution of X by setting the true value of the probabilities of X to the empirical ones, that is,

$$\Pi^X = P^X, \ \mathcal{P}^X = P^X.$$

Having fixed the distribution of X, the DGP is completely described by the conditional choice probabilities Π.

Our minimum distance estimator is the solution to the following quadratic problem:

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

where $\mathbb{B}$ is the parameter space, ϵ_n is a positive cut-off parameter that shrinks to zero with the sample size, as in Chernozhukov, Hong, and Tamer (2007), and

$$T(\beta; P) = \min_{Q=(Q_1,...,Q_k)\in\mathbb{Q}} \sum_{j,k} \omega_{jk}(P) \left[P_{jk} - \int_{\mathbb{C}} \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right) Q_k(d\alpha) \right]^2,$$

where $\mathbb{Q}$ is the set of conditional distributions for α with J points of support for each covariate value index k, that is, for $\mathbb{S}$ the unit simplex in $\mathbb{R}^J$ and $\delta_{\alpha_{km}}$ the Dirac delta function at α_{km},

$$\mathbb{Q} = \left\{ Q := (Q_1, ..., Q_k) : Q_k(d\alpha) = \sum_{m=1}^{J} \pi_{km} \delta_{\alpha_{km}}(\alpha) d\alpha, (\alpha_{k1}, \ldots, \alpha_{kJ}) \in \mathbb{C}, (\pi_{k1}, \ldots, \pi_{kJ}) \in \mathbb{S}, \forall k \right\}.$$

Here we make use of Lemma 10 that tells us that we can obtain a maximizing solution for Q_k as a discrete distribution with at most J points of support for each k. Alternatively, we can write more explicitly

$$T(\beta; P) = \min_{\substack{\alpha_k=(\alpha_{k1},...,\alpha_{kJ})\in\mathbb{C},\forall k \\ \pi_k=(\pi_{k1},...,\pi_{kJ})\in\mathbb{S},\forall k}} \sum_{j,k} \omega_{jk}(P) \left[P_{jk} - \sum_{m=1}^{J} \pi_{km} \mathcal{L}\left(Y^j \mid X^k, \alpha_{km}, \beta\right) \right]^2. \tag{26}$$

In the appendix we give a computational algorithm to solve this problem.

For estimation and inference it is important to allow for the possibility that the postulated model is not perfectly specified, but still provides a good approximation to the true DGP. In this case, when the conditional choice probabilities are misspecified, B_n estimates the identified set for the parameter of the best approximating model to the true DGP with respect to a chi-square distance. This model is obtained by projecting the true DGP $\mathcal{P}$ onto Ξ, the space of conditional choice probabilities that are compatible with the model. In particular, the projection $\mathcal{P}^*$ corresponds to the solution of the minimum distance problem:

$$\mathcal{P}^* = \Pi^*(\mathcal{P}) \in \arg\min_{\Pi\in\Xi} W(\Pi, \mathcal{P}), \quad W(\Pi, \mathcal{P}) = \sum_{j,k} w_{jk}(\mathcal{P})(\mathcal{P}_{jk} - \Pi_{jk})^2, \tag{27}$$

where

$$\begin{aligned} \Xi \;=\; & \{\Pi : \Pi_{jk} = \sum_{m=1}^{J} \pi_{km} \mathcal{L}(Y^j \mid X^k, \alpha_{km}, \beta), \\ & (\alpha_{k1}, \ldots, \alpha_{kJ}) \in \mathbb{C}, (\pi_{k1}, \ldots, \pi_{kJ}) \in \mathbb{S}, \beta \in \mathbb{B}, \forall (j,k)\}. \end{aligned}$$

To simplify the exposition, we will assume throughout that $\mathcal{P}^*$ is unique. Of course, when $\mathcal{P} \in \Xi$, then $\mathcal{P}^* = \mathcal{P}$ and the assumption holds trivially.[2] The identified set for the parameter of the best approximating model is

$$B^* = \left\{ \beta \in \mathbb{B} : \exists Q \in \mathbb{Q} \;\text{ s.t. }\; \int \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right) dQ_k(\alpha) = \mathcal{P}^*_{jk}, \forall (j,k) \right\},$$

i.e., the values of the parameter β that are compatible with the projected DGP $\mathcal{P}^* = (\mathcal{P}^*_{jk}, j = 1, ..., J, k = 1, ..., K)$. Under correct specification of the semiparametric model, we have that $\mathcal{P}^* = \mathcal{P}$ and $B^* = B$.

We shall use the following assumptions.

[2]Otherwise, the assumption can be justified using a genericity argument similar to that presented in Newey (1986), see Appendix. For non-generic values, we can simply select one element of the projection using an additional complete ordering criterion, and work with the resulting approximating model. In practice, we never encountered a non-generic value.

ASSUMPTION 1: *(i) The function F defined in (12) is continuous in (α, β), so that the conditional choice probabilities $\mathcal{L}_{jk}(\alpha, \beta) = \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right)$ are also continuous for all (j, k); (ii) $B^* \subseteq \mathbb{B}$ for some compact set $\mathbb{B}$; (iii) α_i has a support contained in a compact set $\mathbb{C}$; and (iv) the weights $\omega_{jk}(P)$ are continuous in P at $\mathcal{P}$, and $0 < \omega_{jk}(\mathcal{P}) < \infty$ for all (j, k).*

Assumption 1(i) holds for commonly used semiparametric models such as logit and probit models. The condition 1(iv) about the weights is satisfied by the chi-square weights $\omega_{jk}(\mathcal{P}) = \mathcal{P}_k/\mathcal{P}_{jk}$ if $\mathcal{P}_{jk} > 0$, $\forall(j, k)$.

In some results, we also employ the following assumption.

ASSUMPTION 2: *Every $\beta^* \in B^*$ is regular at $\mathcal{P}$ in the sense that, for any sequence $\Pi_n \to \mathcal{P}$, there exists a sequence $\beta_n \in \arg\min_{\beta\in\mathbb{B}} T(\beta, \Pi_n)$ such that $\beta_n \to \beta^*$.*

In a variety of cases the assumption of regularity appears to be a good one. First of all, the assumption holds under point identification, as in the logit model, by the standard consistency argument for maximum likelihood/minimum distance estimators. Second, for probit and other similar models, we can argue that this assumption can also be expected to hold when the true distribution of the individual effect α_i is absolutely continuous, with the exception perhaps of very non-regular parameter spaces and non-generic situations.

To explain the last point, it is convenient to consider a correctly specified model for simplicity. Let the vector of model conditional choice probabilities for $(Y^1, ..., Y^J)$ be $\mathcal{L}_k(\alpha, \beta) \equiv (\mathcal{L}_{1k}(\alpha, \beta), ..., \mathcal{L}_{Jk}(\alpha, \beta))'$. Let $\Gamma_k(\beta) \equiv \{\mathcal{L}_k(\alpha, \beta) : \alpha \in \mathbb{C}\}$ and let $\mathcal{M}_k(\beta)$ be the convex hull of $\Gamma_k(\beta)$. In the case of probit the specification is non-trivial in the sense that $\mathcal{M}_k(\beta)$ possesses a non-empty interior with respect to the J dimensional simplex. For every $\beta^* \in B$ and some Q_k^*, we have that $\mathcal{L}_{jk}(\beta^*, Q^*) = \mathcal{P}_{jk}$ for all (j, k), that is, $(\mathcal{P}_{1k}, ..., \mathcal{P}_{Jk}) \in \mathcal{M}_k(\beta^*)$ for all k. Moreover, under absolute continuity of the true Q^* we must have $(\mathcal{P}_{1k}, ..., \mathcal{P}_{Jk}) \in$ interior $\mathcal{M}_k(\beta_0)$ for all k, where $\beta_0 \in B$ is the true value of β. Next, for any β^* in the neighborhood of β_0, we must have $(\mathcal{P}_{1k}, ..., \mathcal{P}_{Jk}) \in$ interior $\mathcal{M}_k(\beta^*)$ for all k, and so on. In order for a point β^* to be located on the boundary of B we must have that $(\mathcal{P}_{1k}, ..., \mathcal{P}_{Jk}) \in \partial\mathcal{M}_k(\beta^*)$ for some k. Thus, if the identified set has a dense interior, which we verified numerically in a variety of examples for the probit model, then each point in the identified set must be regular. Indeed, take first a point β^* in the interior of B. Then, for any sequence $\Pi_n \to \mathcal{P}$, we must have $(\Pi_{1k}, ..., \Pi_{Jk}) \in \mathcal{M}_k(\beta^*)$ for all k for large n, so that $T(\beta^*; \Pi_n) = 0$ for large n. Thus, there is a sequence of points β_n in $\arg\min_{\beta\in\mathbb{B}} T(\beta; \Pi_n)$ converging to β^*. Now take the point β^{**} on the boundary of B, then for each $\epsilon > 0$, there is β^* in the interior such $\|\beta^* - \beta^{**}\| \leq \epsilon/2$ and such that there is a sequence of points β_n in $\arg\min_{\beta\in\mathbb{B}} T(\beta; \Pi_n)$ and a finite number $n(\epsilon)$ such that for all $n \geq n(\epsilon)$, $\|\beta^* - \beta_n\| \leq \epsilon/2$. Thus, for all $n \geq n(\epsilon)$, $\|\beta^{**} - \beta_n\| \leq \epsilon$. Since $\epsilon > 0$ is arbitrary, it follows that β^{**} is regular.

We can now give a consistency result for the quadratic estimator.

THEOREM 12: *If Assumptions 1 holds and* $\epsilon_n \propto \log n/n$ *then*

$$d_H(B_n, B^*) = o_P(1),$$

where d_H *is the Hausdorff distance between sets*

$$d_H\left(B_n, B^*\right) = \max\left[\sup_{\beta_n \in B_n} \inf_{\beta \in B^*} |\beta_n - \beta|, \sup_{\beta \in B^*} \inf_{\beta_n \in B_n} |\beta_n - \beta|\right].$$

Under Assumption 2 the result holds for $\epsilon_n = 0$.

Moreover, under Assumption 1 the model-predicted probabilities are consistent, for any $\beta_n \in B_n$, *and each* j *and* k,

$$P^*_{jk} = \sum_{m=1}^{J} \pi_{km}(\beta_n) \mathcal{L}\left(Y^j \mid X^k, \alpha_{km}(\beta_n), \beta_n\right) \to_p \mathcal{P}^*_{jk}, \tag{28}$$

where $\{\pi_{km}(\beta_n), \alpha_{km}(\beta_n),\ \forall k, m\}$ *is a solution to the minimum distance problem (26) for any* $\epsilon_n \to 0$, *where we assume that* $\mathcal{P}^*$ *is unique.*

7.2 Marginal Effects

We next consider the problem of estimation of marginal effects, which is of our prime interest. An immediate issue that arises is that we can not directly use the solution to the minimum distance problem to estimate the marginal effects. Indeed, the constraints of the linear programming programs for these effects in (25) many not hold for any $\beta \in B_n$ when $\mathcal{P}$ is replaced by P due to sampling variation or under misspecification. In order to resolve the infeasibility issue, we replace the nonparametric estimates P_{jk} by the probabilities predicted by the model P^*_{jk} as defined in (28), and we re-target our estimands to the marginal effects defined in the best approximating model.

To describe the estimator of the bounds for the marginal effects, it is convenient to introduce some notation. Let

$$\begin{array}{rl} \underline{\mu}^*_k(\beta, \Pi) = & \min_{\alpha_k, \pi_k} D^{-1} \sum_{m=1}^{J} [F\left(\tilde{x}'\beta + \alpha_{km}\right) - F\left(\bar{x}'\beta + \alpha_{km}\right)]\pi_{km} \\ \text{s.t.} & \Pi^*_{jk} = \sum_{m=1}^{J} \mathcal{L}\left(Y^j \mid X^k, \alpha_{km}, \beta\right) \pi_{km}\ \forall j, \\ & \alpha_k = (\alpha_{k1}, \ldots, \alpha_{kJ}) \in \mathbb{C}, \\ & \pi_k = (\pi_{k1}, \ldots, \pi_{kJ}) \in \mathbb{S}, \end{array} \tag{29}$$

and

$$\begin{array}{rl} \overline{\mu}^*_k(\beta, \Pi) = & \max_{\alpha_k, \pi_k} D^{-1} \sum_{j=1}^{J} [F\left(\tilde{x}'\beta + \alpha_{km}\right) - F\left(\bar{x}'\beta + \alpha_{km}\right)]\pi_{km} \\ \text{s.t.} & \Pi^*_{jk} = \sum_{m=1}^{J} \mathcal{L}\left(Y^j \mid X^k, \alpha_{km}, \beta\right) \pi_{km}\ \forall j, \\ & \alpha_k = (\alpha_{k1}, \ldots, \alpha_{kJ}) \in \mathbb{C}, \\ & \pi_k = (\pi_{k1}, \ldots, \pi_{kJ}) \in \mathbb{S}, \end{array} \tag{30}$$

where $\Pi^* = (\Pi^*_{jk}, j = 1, ..., J, k = 1, ..., K)$ denotes the the projection of Π onto Ξ, i.e., $\Pi^* = \Pi^*(\Pi)$ as defined in (27). Thus, the upper and lower bounds on the true marginal effects of the best approximating model take the form:

$$\underline{\mu}^*_k = \min_{\beta \in B^*} \underline{\mu}^*_k(\beta, \mathcal{P}), \quad \overline{\mu}^*_k = \max_{\beta \in B^*} \overline{\mu}^*_k(\beta, \mathcal{P}).$$

Under correct specification, these correspond to the lower and upper bounds on the marginal effects in (14) and (15). We estimate the bounds by

$$\hat{\underline{\mu}}^*_k = \min_{\beta \in B_n} \underline{\mu}_k(\beta, P), \quad \hat{\overline{\mu}}^*_k = \max_{\beta \in B_n} \overline{\mu}^*_k(\beta, P).$$

THEOREM 13: *If Assumptions 1 is satisfied and $\epsilon_n \propto \log n/n$ then*

$$\hat{\underline{\mu}}^*_k = \underline{\mu}^*_k + o_p(1), \quad \hat{\overline{\mu}}^*_k = \overline{\mu}^*_k + o_p(1).$$

Under Assumption 2 the result holds for $\epsilon_n = 0$.

8 Inference

8.1 Modified Projection Method

The following method projects a confidence region for conditional choice probabilities onto a simultaneous confidence region for all possible marginal effects and other structural parameters. If a single marginal effect is of interest, then this approach is conservative; if all (or many) marginal effects are of interest, then this approach is sharp (or close to sharp). In the next section, we will present an approach that appears to be sharp, at least in large samples, when a particular single marginal effect is of interest.

It is convenient to describe the approach in two stages.

Stage 1. The nonparametric space Ξ_N of conditional choice probabilities is the product of K simplex sets $\mathbb{S}$ of dimension J, that is, $\Xi_N = \mathbb{S}^K$. Thus we can begin by constructing a confidence region for the true choice probabilities $\mathcal{P}$ by collecting all probabilities $\Pi \in \Xi_N$ that pass a goodness-of-fit test:

$$CR_{1-\alpha}(\mathcal{P}) = \left\{ \Pi \in \Xi_N : W(\Pi, P) \leq c_{1-\alpha}(\chi^2_{K(J-1)}) \right\},$$

where $c_{1-\alpha}(\chi^2_{K(J-1)})$ is the $(1-\alpha)$-quantile of the $\chi^2_{K(J-1)}$ distribution and W is the goodness-of-fit statistic:

$$W(\Pi, P) = n \sum_{j,k} P_k \frac{(P_{jk} - \Pi_{jk})^2}{\Pi_{jk}}.$$

Stage 2. To construct confidence regions for marginal effects and any other structural parameters we project each $\Pi \in CR_{1-\alpha}(\mathcal{P})$ onto Ξ, the space of conditional choice probabilities that are compatible with the model. We obtain this projection $\Pi^*(\Pi)$ by solving the minimum distance problem:

$$\Pi^*(\Pi) = \arg\min_{\tilde{\Pi}\in\Xi} W(\tilde{\Pi}, \Pi), \quad W(\tilde{\Pi}, \Pi) = n \sum_{j,k} P_k \frac{(\Pi_{jk} - \tilde{\Pi}_{jk})^2}{\Pi_{jk}}. \tag{31}$$

The confidence regions are then constructed from the projections of all the choice probabilities in $CR_{1-\alpha}(\mathcal{P})$. For the identified set of the model parameter, for example, for each $\Pi \in CR_{1-\alpha}(\mathcal{P})$ we solve

$$B^*(\Pi) = \left\{ \beta \in \mathbb{B} : \exists Q \in \mathbb{Q} \text{ s.t. } \int \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right) dQ_k(\alpha) = \Pi^*_{jk}, \forall (j,k), \Pi^* = \Pi^*(\Pi) \right\}. \tag{32}$$

Denote the resulting confidence region as

$$CR_{1-\alpha}(B^*) = \{B^*(\Pi) : \Pi \in CR_{1-\alpha}(\mathcal{P})\}.$$

We may interpret this set as a confidence region for the set B^* collecting all values β^* that are compatible with the best approximating model $\mathcal{P}^*$. Under correct specification, B^* is just the identified set B.

If we are interested in bounds on marginal effects, for each $\Pi \in CR_{1-\alpha}(\mathcal{P})$ we get

$$\underline{\mu}^*_k(\Pi) = \min_{\beta\in B^*(\Pi)} \underline{\mu}^*_k(\beta, \Pi), \quad \overline{\mu}^*_k(\Pi) = \max_{\beta\in B^*(\Pi)} \overline{\mu}^*_k\ (\beta, \Pi),\ k = 1, ..., K.$$

Denote the resulting confidence regions as

$$CR_{1-\alpha}[\underline{\mu}^*_k, \overline{\mu}^*_k] = \{[\underline{\mu}^*_k(\Pi), \overline{\mu}^*_k(\Pi)] : \Pi \in CR_{1-\alpha}(\mathcal{P})\}.$$

These sets are confidence regions for the sets $[\underline{\mu}^*_k, \overline{\mu}^*_k]$, where $\underline{\mu}^*_k$ and $\overline{\mu}^*_k$ are the lower and upper bounds on the marginal effects induced by any best approximating model in $(B^*, \mathcal{P}^*)$. Under correct specification, these will include the upper and lower bounds on the marginal effect $[\underline{\mu}_k, \overline{\mu}_k]$ induced by any true model in $(B, \mathcal{P})$.

In a canonical projection method we would implement the second stage by simply intersecting $CR_{1-\alpha}(\mathcal{P})$ with Ξ, but this may give an empty intersection either in finite samples or under misspecification. We avoid this problem by using the projection step instead of the intersection, and also by re-targeting our confidence regions onto the best approximating model. In order to state the result about the validity of ourmodified projection method in large samples, let Δ be the set of vectors with all components bounded away from zero by some $\epsilon > 0$.

THEOREM 14: *Suppose Assumption 1 holds, then for (any sequence of true parameter values)* $\mathcal{P}_0 = (\mathcal{P}', \mathcal{P}^{X\prime})' \in \Delta$

$$\lim_{n\to\infty} \Pr_{\mathcal{P}_0} \left\{ \begin{array}{lcl} \mathcal{P} & \in & CR_{1-\alpha}(\mathcal{P}) \\ B^* & \in & CR_{1-\alpha}(B^*) \\ [\underline{\mu}_k^*, \overline{\mu}_k^*] & \in & CR_{1-\alpha}[\underline{\mu}_k^*, \overline{\mu}_k^*], \forall k \end{array} \right\} = 1 - \alpha.$$

8.2 Perturbed Bootstrap

In this section we present an approach that appears to be sharper than the projection method, at least in large samples, when a particular single marginal effect is of interest. The estimators for parameters and marginal effects are obtained by nonlinear programming subject to data-dependent constraints that are modified to respect the constraints of the model. The distributions of these highly complex estimators are not tractable, and are also non-regular in the sense that the limit versions of these distributions do not vary with perturbations of the DGP in a continuous fashion. This implies that the usual bootstrap is not consistent. To overcome all of these difficulties we will rely on a variation of the bootstrap, which we call the perturbed bootstrap.

The usual bootstrap computes the critical value – the α-quantile of the distribution of a test statistic – given a consistently estimated data generating process (DGP). If this critical value is not a continuous function of the DGP, the usual bootstrap fails to consistently estimate the critical value. We instead consider the perturbed bootstrap, where we compute a set of critical values generated by suitable perturbations of the estimated DGP and then take the most conservative critical value in the set. If the perturbations cover at least one DGP that gives a more conservative critical value than the true DGP does, then this approach yields a valid inference procedure.

The approach outlined above is most closely related to the Monte-Carlo inference approach of Dufour (2006); see also Romano and Wolf (2000) for a finite-sample inference procedure for the mean that has a similar spirit. In the set-identified context, this approach was first applied in the MIT thesis work of Rytchkov (2007); see also Chernozhukov (2007).

Recall that the complete description of the DGP is provided by the parameter vector $(\Pi', \Pi^{X\prime})'$, where $\Pi = (\Pi_{jk}, j = 1, ..., J, k = 1, ..., K)'$, $\Pi^X = (\Pi_k, k = 1, ..., K)'$, $\Pi_{jk} = \Pr(Y = Y^j | X = X^k)$, and $\Pi_k = \Pr(X = X^k)$. The true value of the parameter vector is $(\mathcal{P}', \mathcal{P}^{X\prime})$ and the nonparametric empirical estimate is $(P', P^{X\prime})'$. As before, we condition on the observed distribution of X and thus set $\Pi^X = P^X$ and $\mathcal{P}^X = P^X$.

We consider the problem of performing inference on a real parameter θ^*. For example, θ^*

can be an upper (or lower) bound on the marginal effect μ_k such as

$$\theta^*(\Pi) = \max_{\beta \in B^*(\Pi), Q \in \mathbb{Q}} D^{-1} \int [F(\tilde{x}'\beta + \alpha) - F(\bar{x}'\beta + \alpha)] Q_k (d\alpha) \text{ s.t. } \mathcal{L}_{jk}(\beta, Q_k) = \Pi^*_{jk}, \ \forall j,$$

where $\Pi^* = (\Pi^*_{jk}, j = 1, ..., J, k = 1, ..., K)$ denotes the projection of Π onto the model space, as defined in (31), and $B^*(\Pi)$ is the corresponding projection for the identified set of the parameter defined as in (32). Alternatively, θ^* can be an upper (or lower) bound on a scalar functional c'^* of the parameter β^*. Then we define

$$\theta^*(\Pi) = \max_{\beta \in B^*(\Pi)} c'\beta.$$

As before, we project Π onto the model space in order to address the problem of infeasibility of constraints defining the parameters of interest under misspecification or sampling error. Under misspecification, we interpret our inference as targeting the parameters of interest in the best approximating model.

In order to perform inference on the true value $\theta^* = \theta^*(\mathcal{P})$ of the parameter, we use the statistic

$$S_n = \hat{\theta} - \theta^*,$$

where $\hat{\theta} = \theta^*(P)$. Let $G_n(s, \Pi)$ denote the distribution function of $S_n(\Pi) = \hat{\theta} - \theta^*(\Pi)$, when the data follow the DGP Π. The goal is to estimate the distribution of the statistic S_n under the true DGP $\Pi = \mathcal{P}$, that is, to estimate $G_n(s, \mathcal{P})$.

The method proceeds by constructing a confidence region $CR_{1-\gamma}(\mathcal{P})$ that contains the true DGP $\mathcal{P}$ with probability $1-\gamma$, close to one. For efficiency purposes, we also want the confidence region to be an efficient estimator of $\mathcal{P}$, in the sense that as $n \to \infty$, $d_H(CR_{1-\gamma}(\mathcal{P}), \mathcal{P}) = O_p(n^{1/2})$,where d_H is the Hausdorff distance between sets. Specifically, in our case we use

$$CR_{1-\gamma}(\mathcal{P}) = \{\Pi \in \Xi_N : W(\Pi, P) \leq c_{1-\gamma}(\chi^2_{K(J-1)})\},$$

where $c_{1-\gamma}(\chi^2_{K(J-1)})$ is the $(1-\gamma)$-quantile of the $\chi^2_{K(J-1)}$ distribution and W is the goodness-of-fit statistic:

$$W(\Pi, P) = n \sum_{j,k} P_k \frac{(P_{jk} - \Pi_{jk})^2}{\Pi_{jk}}.$$

Then we define the estimates of lower and upper bounds on the quantiles of $G_n(s, \mathcal{P})$ as

$$\underline{G}_n^{-1}(\alpha, \mathcal{P}) / \overline{G}_n^{-1}(\alpha, \mathcal{P}) = \inf / \sup_{\Pi \in CR_{1-\gamma}(\mathcal{P})} G_n^{-1}(\alpha, \Pi), \tag{33}$$

where $G_n^{-1}(\alpha, \Pi) = \inf\{s : G_n(s, \Pi) \geq \alpha\}$ is the α-quantile of the distribution function $G_n(s, \Pi)$. Then we construct a $(1 - \alpha - \gamma) \cdot 100\%$ confidence region for the parameter of interest as

$$CR_{1-\alpha-\gamma}(\theta^*) = \left[\underline{\theta}, \overline{\theta}\right]$$

where, for $\alpha = \alpha_1 + \alpha_2$,

$$\underline{\theta} = \hat{\theta} - \overline{G}^{-1}(1-\alpha_1, \mathcal{P}), \quad \overline{\theta} = \hat{\theta} - \underline{G}^{-1}(\alpha_2, \mathcal{P}).$$

This formulation allows for both one-sided intervals (either $\alpha_1 = 0$ or $\alpha_2 = 0$) or two-sided intervals ($\alpha_1 = \alpha_2 = \alpha/2$).

The following theorem shows that this method delivers (uniformly) valid inference on the parameter of interest.

THEOREM 15. *Suppose Assumption 1 holds, then for (any sequence of true parameter values)* $\mathcal{P}_0 = (\mathcal{P}', \mathcal{P}^{X\prime})' \in \Delta$

$$\lim_{n\to\infty} \Pr_{\mathcal{P}_0}(\theta^* \in \left[\underline{\theta}, \overline{\theta}\right]) \geq 1 - \alpha - \gamma.$$

In practice, we use the following computational approximation to the procedure described above:

1. Draw a potential DGP $\Pi_r = (\Pi'_{r1}, ..., \Pi'_{rK})$, where $\Pi_{rk} \sim \mathcal{M}(nP_k, (P_{1k}, ..., P_{Jk}))/(nP_k)$ and $\mathcal{M}$ denotes the multinomial distribution.

2. Keep Π_r if it passes the chi-square goodness of fit test at the γ level, using $K(J-1)$ degrees of freedom, and proceed to the next step. Otherwise reject, and repeat step 1.

3. Estimate the distribution $G_n(s, \Pi_r)$ of $S_n(\Pi_r)$ by simulation under the DGP Π_r.

4. Repeat steps 1 to 3 for $r = 1, ..., R$, obtaining $G_n(s, \Pi_r)$, $r = 1, ..., R$.

5. Let $\hat{\underline{G}}^{-1}(\alpha, \mathcal{P})/\hat{\overline{G}}^{-1}(\alpha, \mathcal{P}) = \min / \max\{G_n^{-1}(\alpha, \Pi_1), ..., G_n^{-1}(\alpha, \Pi_R)\}$, and construct a $1 - \alpha - \gamma$ confidence region for the parameter of interest as $CR_{1-\alpha-\gamma}(\theta^*) = \left[\underline{\theta}, \overline{\theta}\right]$, where $\underline{\theta} = \hat{\theta} - \hat{\overline{G}}^{-1}(1-\alpha_1, \mathcal{P})$, $\overline{\theta} = \hat{\theta} - \hat{\underline{G}}^{-1}(\alpha_2, \mathcal{P})$, and $\alpha_1 + \alpha_2 = \alpha$.

The computational approximation algorithm is necessarily successful, if it generates at least one draw of DGP Π_r that gives more conservative estimates of the tail quantiles than the true DGP does, namely $[G^{-1}(\alpha_2, \mathcal{P}), G^{-1}(1-\alpha_1, \mathcal{P})] \subseteq [\underline{G}^{-1}(\alpha_2, \Pi_r), \overline{G}^{-1}(1-\alpha_1, \Pi_r)]$.

9 Empirical Example

We now turn to 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, and 1994. 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 more 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 presences have the strongest impact on the mother participation decision. LFP is stable across the years considered, whereas $kids$ is increasing. The proportion of women that change fertility status grows steadily with the number of time periods of the panel, but there are still 49% of the women in the sample for which the effect of fertility is not identified after 3 periods.

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

$$LFP_{it} = \mathbf{1}\left\{\beta \cdot kids_{it} + \alpha_i + \epsilon_{it} \geq 0\right\}, \tag{34}$$

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 mean model and semiparametric logit and probit models described in Sections 2 and 5, 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.[3] The estimates from the general model impose monotonicity of the effects. For the semiparametric estimators, we use the algorithm described in the appendix with penalty $\lambda_n = 1/(n \log n)$ and iterate the quadratic program 3 times with initial weights $w_{jk} = nP_k$. This iteration makes the estimates insensitive to the penalty and 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 and 3 time periods, both of them starting in 1990.

Tables 3 and 4 report estimates of the model parameters and marginal effects for 2 and 3

[3]The analytical corrections use the estimators of the bias based on expected quantities in Fernández-Val (2008). The Jackknife bias correction uses the procedure described in Hahn and Newey (2004).

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 (95% N) and nonparametric bootstrap with 200 repetitions (95% B). For the logit and probit models, the confidence regions are obtained by the modified projection method (95% MP), where the confidence interval for $\mathcal{P}$ in the first stage is approximated by 50,000 DGPs drawn from the empirical multinomial distributions that pass the goodness of fit test; and the perturbed bootstrap method (95% PB) with $R = 100$, $\gamma = .01$, $\alpha_1 = \alpha_2 = .02$, and 200 simulations from each DGP to approximate the distribution of the statistic. We also include confidence intervals obtained by a canonical projection method (95% MP) that intersects the nonparametric confidence interval for $\mathcal{P}$ with the space of probabilities compatible with the semiparametric model Ξ:

$$CR_{1-\alpha}(\mathcal{P}) = \left\{ \Pi \in \Xi : W(\Pi, P) \leq c_{1-\alpha}(\chi^2_{K(J-1)}) \right\}.$$

For the fixed effects estimators, the confidence regions are based on the asymptotic normal approximation. The semiparametric estimates are shown for $\epsilon_n = 0$, i.e., for the solution that gives the minimum value in the quadratic problem.

Overall, we find that the estimates and confidence regions based on the general conditional mean 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 canonical projection method clearly arises in this application. Thus, this procedure gives empty confidence regions for the panel with 3 periods. The modified projection and perturbed bootstrap methods produce similar (non-empty) confidence regions for the model parameters and marginal effects.

10 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 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.

11 Appendix

11.1 Proofs

PROOF OF THEOREM 1: By eq. (3),

$$\begin{aligned}\sum_t (X_t^k - r^k) E[Y_{it} \mid X_i = X^k] &= Tr^k(1-r^k)\int m(1,\alpha)Q_k^*(d\alpha) \\ &\quad + T(1-r^k)(-r^k)\int m(0,\alpha)Q_k^*(d\alpha) = T\sigma_k^2\mu_k. \end{aligned} \tag{35}$$

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

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

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

PROOF OF THEOREM 2: 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 $X_{\tilde{t}}^k = \tilde{x}$ we have $E[Y_{i\tilde{t}} \mid X_i = X^k] = \int m(\tilde{x},\alpha)Q_k^*(d\alpha)$. Therefore, for $\tilde{r}^k = \#\{t : X_t^k = \tilde{x}\}/T$ and $\bar{r}^k = \#\{t : X_t^k = \bar{x}\}/T$, by the law of large numbers,

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

$$\frac{1}{n}\sum_{i=1}^n 1(\tilde{r}_i > 0)1(\bar{r}_i > 0) \xrightarrow{p} E[1(\tilde{r}_i > 0)1(\bar{r}_i > 0)] = \sum_{k\in\mathcal{K}^*}\mathcal{P}_k.$$

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

PROOF OF LEMMA 3: As before let $Q_k^*(\alpha)$ denote the conditional CDF of α given $X_i = X^k$. Note that

$$\bar{m}_t^k = \frac{E[Y_{it} \mid X_i = X^k]}{D} = \frac{\int m(X_t^k, \alpha) Q_k^*(d\alpha)}{D}.$$

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 $X_{\tilde{t}}^k = \tilde{x}$ and $X_{\bar{t}}^k = \bar{x}$

$$\bar{m}_{\tilde{t}}^k - \bar{m}_{\bar{t}}^k = \frac{\int m(\tilde{x}, \alpha) Q_k^*(d\alpha)}{D} - \frac{\int m(\bar{x}, \alpha) Q_k^*(d\alpha)}{D} = \mu_k.$$

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

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

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

$$\bar{m}_{\tilde{t}}^k - B_u = \frac{\int m(\tilde{x}, \alpha) Q_k^*(d\alpha)}{D} - B_u \leq \mu_k \leq \frac{\int m(\tilde{x}, \alpha) Q_k^*(d\alpha)}{D} - B_\ell = \bar{m}_{\tilde{t}}^k - B_\ell.$$

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 $X_{\tilde{t}}^{k^*} = \tilde{x}$ and $X_{\bar{t}}^{k^*} = \bar{x}$, 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 4: See text.

PROOF OF THEOREM 5: 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 $Z_{iT} \xrightarrow{as} b(\alpha_i) = \min\{\Pr(X_{it} = \tilde{x} \mid \alpha_i), \Pr(X_{it} = \bar{x} \mid \alpha_i)\} > 0$. Therefore $\Pr(A_{iT} \mid \alpha_i) \geq \Pr(Z_{iT} > 0 \mid \alpha_i) \longrightarrow 1$ for almost all α_i. It then follows by the dominated convergence theorem that

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

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 6: Let $\mathcal{P}_1$ and $\mathcal{P}_K$ be as in equation (6). By the Markov assumption,

$$\begin{aligned}
\mathcal{P}_1 &= \Pr(X_{i1} = \cdots = X_{iT} = 0) = E[\Pr(X_{i1} = \cdots = X_{iT} = 0 \mid \alpha_i)] \\
&= E[\Pi_{t=J+1}^{T} \Pr(X_{it} = 0 \mid X_{i,t-1} = \cdots = X_{i,t-J} = 0, \alpha_i) \Pr(X_{iJ} = \cdots = X_{i,t-J} = 0 \mid \alpha_i)] \\
&\leq E[(p_i^1)^{T-J}]. \\
\mathcal{P}_K &\leq E[(p_i^K)^{T-J}].
\end{aligned}$$

The first bound then follows as in (6). The second bound then follows from the condition $p_i^k \leq 1 - \varepsilon$ for $k \in \{1, K\}$. Now suppose that there is a set A of possible α_i such that $\Pr(A) > 0$, $Q_i = \Pr(X_{i1} = \cdots = X_{iJ} = 0 \mid \alpha_i) > 0$ and $p_i^1 = 1$ Then

$$\mathcal{P}_1 = E[(p_i^1)^{T-J} Q_i] \geq E[1(\alpha_i \in A)(p_i^1)^{T-J} Q_i] = E[1(\alpha_i \in A) Q_i] > 0.$$

Therefore, for all T the probability $\mathcal{P}_1$ is bounded away from zero, and hence $\mu_\ell \nrightarrow \mu_0$ or $\mu_u \nrightarrow \mu_0$.Q.E.D.

PROOF OF THEOREM 7: Note that every $X_i \in \mathcal{X}^t(x)$ has $X_{it} = x$. Also, the X_{is} for $s > t$ are completely unrestricted by $X_i \in \mathcal{X}^t(x)$. Therefore, it follows by the key implication that

$$E[Y_{it} \mid X_i \in \mathcal{X}^t(x)] = \int m(x, \alpha) Q^*(d\alpha \mid X_i \in \mathcal{X}^t(x)).$$

Then by iterated expectations,

$$\begin{aligned}
\int m(x, \alpha) Q^*(d\alpha) = \bar{\mathcal{P}}(x) \int m(x, \alpha) Q^*(d\alpha \mid X_i \in \bar{\mathcal{X}}(x)) \\
+ \sum_{t=1}^{T} \Pr(X_i \in \mathcal{X}^t(x)) \int m(x, \alpha) Q^*(d\alpha \mid X_i \in \mathcal{X}^t(x)) \\
= \bar{\mathcal{P}}(x) \int m(x, \alpha) Q^*(d\alpha \mid X_i \in \bar{\mathcal{X}}(x)) + E[\sum_{t=1}^{T} 1(X_i \in \mathcal{X}^t(x)) Y_{it}].
\end{aligned}$$

Using the bound and dividing by D then gives

$$\begin{aligned}
E[\sum_{t=1}^{T} 1(X_i \in \mathcal{X}^t(x)) Y_{it}]/D + \bar{\mathcal{P}}(x) B_\ell \leq \int m(x, \alpha) Q^*(d\alpha)/D \\
\leq E[\sum_{t=1}^{T} 1(X_i \in \mathcal{X}^t(x)) Y_{it}]/D + \bar{\mathcal{P}}(x) B_u.
\end{aligned}$$

Differencing this bound for $x = \tilde{x}$ and $x = \bar{x}$ gives the result. 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 $\mathcal{Q}_{k\beta} = \{Q_k : \int \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right) Q_k\left(d\alpha\right) = \mathcal{P}_{jk},\ \ j = 1, ..., J\}$. 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 $\mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right) = \prod_{t=1}^{T} F(X_t^k\beta+\alpha)^{Y_t^j}\{1-F(X_t^k\beta+\alpha)\}^{1-Y_t^j}$, 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 only has $T + 1$ different elements given by $(1, F(\alpha), ..., F(\alpha)^T)$. The feasible set simplifies to:

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

where the moments M_{kt} are identified by the data. Here $\int F(\alpha)Q_k\left(d\alpha\right) = M_{k1}$ 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:

$$\overline{\mu}_k - \underline{\mu}_k = \max_{G_k \in \mathcal{G}_{k\beta}, \beta \in B} D^{-1} \int_0^1 h_\beta(z)G_k(dz) - \min_{G_k \in \mathcal{G}_{k\beta}, \beta \in B} D^{-1} \int_0^1 h_\beta(z)G_k(dz), \tag{36}$$

where $\mathcal{G}_{k\beta} = \{G_k : \int_0^1 z^t G_k(dz) = M_{kt},\ \ t = 0, ..., T\}$, $h_\beta(z) = F(\beta + F^{-1}(z))$, 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

$$h_\beta(z) = P_\beta(z, T) + R_\beta(z, T), \tag{37}$$

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

$$\|R_\beta(z,T)\|_\infty \leq \frac{(T-r)!}{T!}\left(\frac{\pi}{4}\right)^r \bar{h}^r_\beta = O\left(T^{-r}\right), \tag{38}$$

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

$$\overline{\mu}_k - \underline{\mu}_k = \max_{G_k \in \mathcal{G}_{k\beta}} \int_0^1 R_{\beta^*}(z,T)G_k(dx) - \min_{G_k \in \mathcal{G}_{k\beta}} \int_0^1 R_{\beta^*}(z,T)G_k(dx) \leq 2\bar{h}^r_{\beta^*} = O\left(T^{-r}\right). \tag{39}$$

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

$$h_\beta(z) = \frac{ze^\beta}{1-(1-e^\beta)z}, \tag{40}$$

with derivatives

$$h^r_\beta(z) = r!\frac{e^\beta(1-e^\beta)^{r-1}}{[1-(1-e^\beta)z]^r}. \tag{41}$$

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: Note that for $T = 2$ and X binary, we have that $K = 4$. Let $X^1 = (0,0)$, $X^2 = (0,1)$, $X^3 = (1,0)$, and $X^4 = (1,1)$. By Lemma 3, μ_I is identified by

$$\mu_I = \mathcal{P}^*_2[\Pr\{Y=(0,1) \mid X^2\} - \Pr\{Y=(1,0) \mid X^2\}] + \mathcal{P}^*_3[\Pr\{Y=(1,0) \mid X^3\} - \Pr\{Y=(0,1) \mid X^3\}].$$

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

$$\tilde{\mu}_I = \sum_{k=2}^3 \mathcal{P}^*_k[\Pr\{Y=(0,1) \mid X^k\} + \Pr\{Y=(1,0) \mid X^k\}][F(\tilde{\beta}/2) - F(-\tilde{\beta}/2)].$$

The condition for consistency $\tilde{\mu}_I = \mu_I$ can be written as

$$F(\tilde{\beta}/2) = \frac{\mathcal{P}_2 \Pr\{Y=(0,1) \mid X^2\} + \mathcal{P}_3 \Pr\{Y=(1,0) \mid X^3\}\}}{\sum_{k=2}^3 \mathcal{P}_k[\Pr\{Y=(0,1) \mid X^k\} + \Pr\{Y=(1,0) \mid X^k\}]},$$

but this is precisely the first order condition of the program (16). This result follows, after some algebra and using the symmetry property of F, by solving the profile problem

$$\tilde{\beta} = \arg\max_\beta \sum_{k=1}^K \mathcal{P}_k[\Pr\{Y=(0,1) \mid X^k\} \log F(\Delta X^k\beta/2) + \Pr\{Y=(1,0) \mid X^k\} \log F(-\Delta X^k\beta/2)],$$

where $\Delta X^k = X_2^k - X_1^k$. Q.E.D.

PROOF OF LEMMA 10: First, by $\beta \in B$, we have that $T(\beta; \mathcal{P}) = 0$ and therefore any $Q_{k\beta} \in \arg\max_{Q_k} \sum_{j=1}^{J} \omega_{jk}(\mathcal{P}) \left(\mathcal{P}_{jk} - \mathcal{L}_{jk}(\beta, Q_k)\right)^2$ satisfies $\mathcal{L}_{jk}(\beta, Q_{k\beta}) = \mathcal{P}_{jk}$ $\forall j$, for each k.

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

$$\mathcal{L}_k(\alpha, \beta) \equiv \left(\mathcal{L}\left(Y^1 \mid X^k, \alpha, \beta\right), ..., \mathcal{L}\left(Y^J \mid X^k, \alpha, \beta\right)\right)'.$$

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 $\mathbb{C}$, and has at most dimension $J-1$ since the sum of the elements of $\mathcal{L}_k(\beta, \alpha)$ is one $\forall \alpha$. Now, let $\mathcal{M}_k(\beta)$ denote the convex hull of $\Gamma_k(\beta)$. For any $\beta \in B$ we have that there is at least one $Q_{k\beta}$ such that $\mathcal{L}_{jk}(\beta, Q_{k\beta}) = \mathcal{P}_{jk}$ $\forall j$, i.e.,

$$(\mathcal{P}_{1k}, ..., \mathcal{P}_{Jk}) \in \mathcal{M}_k(\beta).$$

By Carathéodory Theorem any point in $\mathcal{M}_k(\beta)$ can be written as a convex combination of at most J vectors located in $\Gamma_k(\beta)$. Then, we can write

$$(\mathcal{P}_{1k}, ..., \mathcal{P}_{Jk}) = \sum_{m=1}^{J} \pi_{km} \mathcal{L}_k(\alpha_{km}, \beta),$$

where $(\pi_{k1}, ..., \pi_{kJ})$ is on the unit simplex $\mathbb{S}$ of dimension J. Thus, the discrete distribution with J support points at $(\alpha_{k1}, ..., \alpha_{kJ})$ and probabilities $(\pi_{k1}, ..., \pi_{kJ})$ solves the population problem for $Q_{k\beta}$. The result also follows from Lindsay (1995, Theorem 18, p. 112, and Theorem 21, p. 116) (though Lindsay does not provide proofs for his theorems). Q.E.D.

PROOF OF LEMMA 11: For $\beta \in B$, let $\mathcal{Q}_{k\beta} = \{Q_k : \mathcal{L}_{jk}(\beta, Q_k) = \mathcal{P}_{jk}, \quad j = 1, ..., J\}$. 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 $(\pi_{k1}, ..., \pi_{kM})$ among $(\alpha_1, ..., \alpha_M)$ in order to solve

$$\max_{(\pi_{k1}, ..., \pi_{kM})} \sum_{m=1}^{M} [F\left(\tilde{x}'\beta + \alpha_m\right) - F\left(\bar{x}'\beta + \alpha_m\right)] \pi_{km}$$

subject to the constraints:

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

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 $\pi_{km} = 0$, see, e.g., Theorem 2.3 and Definition 2.9 (ii) in Bertsimas and Tsitsiklis (1997). Hence a basic solution to this linear programming problem will have at least $M - J$ zeroes, that is at most J strictly positive π_{km}'s.[4] Thus, we have shown that given the original $\bar{Q}^M_{k\beta}$ with $M \gg J$ points of support there exists a distribution $\bar{Q}^L_{k\beta} \in \mathcal{Q}_{k\beta}$ with just J points of support such that

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

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

The final claim is that there exists a distribution $\bar{Q}^L_{k\beta} \in \mathcal{Q}_{k\beta}$ with J points of support $(\alpha_{k1}, ..., \alpha_{kJ})$ such that

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

Suppose otherwise, then it must be that

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

for some $\epsilon > 0$ and for *all* $\bar{Q}^L_{k\beta}$ 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.

[4]Note that $rank(A) \leq J - 1$, since $\sum_{j=1}^{J} \mathcal{L}\left(Y^j \mid X^k, \alpha, \beta\right) = 1$. 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 $X^k_1 \neq X^k_2$, $\beta \neq 0$, and F is any continuous distribution different from the logistic.

LEMMA A1: *Let* $T(\beta, Q; \Pi) = \sum_{j,k} \omega_{jk}(\Pi) \left(\Pi_{jk} - \mathcal{L}_{jk}(\beta, Q_k)\right)^2$. *If Assumption 1 is satisfied then, for* $\mathbb{Q}$ *equal to the collection of distributions with support contained in a compact set* $\mathbb{C}$,

$$\sup_{\beta \in \mathbb{B}, Q \in \mathbb{Q}} |T(\beta, Q; P) - T(\beta, Q; \mathcal{P})| = o_{P^*}(1).$$

Proof: Note that we can write

$$\begin{aligned} T(\beta, Q; P) - T(\beta, Q; \mathcal{P}) &= \sum_{j,k} \omega_{jk}(P)(P_{jk} - \mathcal{P}_{jk})^2 + 2 \sum_{j,k} \omega_{jk}(P)(P_{jk} - \mathcal{P}_{jk}) \left(\mathcal{P}_{jk} - \mathcal{L}_{jk}(\beta, Q_k)\right) \\ &\quad + \sum_{j,k} (\omega_{jk}(P) - \omega_{jk}(\mathcal{P})) \left(\mathcal{P}_{jk} - \mathcal{L}_{jk}(\beta, Q_k)\right)^2. \end{aligned}$$

The result then follows from $P_{jk} - \mathcal{P}_{jk} = o_P(1)$ and $\omega_{jk}(P) - \omega_{jk}(\mathcal{P}) = o_P(1)$ by the continuous mapping theorem. Q.E.D.

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

LEMMA A2: *Let* $T(\beta; \Pi) = \inf_{Q \in \mathbb{Q}} T(\beta, Q; \Pi)$. *If Assumption 1 is satisfied then*

$$\sup_{\beta \in \mathbb{B}} |T(\beta; P) - T(\beta; \mathcal{P})| = o_{P^*}(1).$$

Proof: Let $\hat{Q}_\beta \in \arg\inf_{Q \in \mathbb{Q}} T(\beta, Q; P)$ and $Q_\beta \in \arg\inf_{Q \in \mathbb{Q}} T(\beta, Q; \mathcal{P})$. By definition of $\hat{Q}_\beta$ and Q_β, we have uniformly in β and for all n,

$$T(\beta, \hat{Q}_\beta; P) - T(\beta, \hat{Q}_\beta; \mathcal{P}) \leq T(\beta, \hat{Q}_\beta; P) - T(\beta, Q_\beta; \mathcal{P}) \leq T(\beta, Q_\beta; P) - T(\beta, Q_\beta; \mathcal{P}).$$

Hence

$$\left|T(\beta, \hat{Q}_\beta; P) - T(\beta, Q_\beta; \mathcal{P})\right| \leq \max\left[\left|T(\beta, \hat{Q}_\beta; P) - T(\beta, \hat{Q}_\beta; \mathcal{P})\right|, |T(\beta, Q_\beta; P) - T(\beta, Q_\beta; \mathcal{P})|\right] = o_{P^*}(1),$$

uniformly in β by Lemma A1. Q.E.D.

LEMMA A3: *If Assumption 1 is satisfied then* $T(\beta; \mathcal{P})$ *is continuous in* β.

Proof: By Lemma 10, the problem

$$\inf_{Q \in \mathbb{Q}} T(\beta, Q; \mathcal{P})$$

can be rewritten as

$$\min_{\substack{(\alpha_{1k}, \ldots, \alpha_{Jk}) \in \mathbb{C}, \forall k \\ (\pi_{1k}, \ldots, \pi_{Jk}) \in \mathbb{S}, \forall k}} \sum_{j,k} \omega_{jk}(\mathcal{P}) \left[\mathcal{P}_{jk} - \sum_{m=1}^{J} \pi_{km} \mathcal{L}\left(Y^j \mid X^k, \alpha_{km}, \beta\right)\right]^2,$$

where J and K are finite, and $\mathbb{S}$ denotes the unit simplex in $\mathbb{R}^J$. Here, $(\alpha_{1k}, \ldots, \alpha_{jk})$ and $(\pi_{1k}, \ldots, \pi_{Jk})$ characterize discrete distributions with no more than J points of support. Because

the objective function is continuous in $(\beta, \alpha_{11}, \ldots, \alpha_{JK}, \pi_{11}, \ldots, \pi_{JK})$, and because $\mathbb{C}^K \times \mathbb{S}^K$ 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.

LEMMA A4: *If Assumption 1 is satisfied then*

$$\sup_{\beta \in B} |T(\beta; P) - T(\beta; \mathcal{P})| = O_{P^*}\left(n^{-1}\right).$$

Proof: Let $Q_\beta \in \arg\min_{Q \in \mathbb{Q}} T(\beta, Q; \mathcal{P})$. By Lemma 10, we have that $\mathcal{P}_{jk} = \mathcal{L}_{jk}(\beta, Q_{k\beta})$ and $T(\beta; \mathcal{P}) = 0 \ \forall \beta \in B$. Then, we have

$$\sup_{\beta \in B} |T(\beta; P) - T(\beta; \mathcal{P})| = \sup_{\beta \in B} T(\beta; P) \leq \sup_{\beta \in B} T(\beta, Q_\beta; P) = \sum_{j,k} \omega_{jk}(P) \left(P_{jk} - \mathcal{P}_{jk}\right)^2 = O_{P^*}(n^{-1}),$$

where the last equality follows from $P_{jk} - \mathcal{P}_{jk} = O_P(n^{-1/2})$, $\omega_{jk}(P) = \omega_{jk}(\mathcal{P}) + o_P(1)$ by the continuous mapping theorem, and J and K being finite. Q.E.D.

PROOF OF THEOREM 12. The consistency result under Assumption 1 and $\epsilon_n \propto \log n/n$ follows from Theorem 3.1 in Chernozhukov, Hong, and Tamer (2007) with $a_n = n$. Indeed, the Condition C.1 in Chernozhukov, Hong, and Tamer (2007) follows by Assumption 1 ($\mathbb{B}$ compact), Lemma A3 ($T(\beta; \mathcal{P})$ continuous), Lemma A2 (uniform convergence of $T(\beta; P)$ to $T(\beta; \mathcal{P})$ in $\mathbb{B}$), and Lemma A4 (uniform convergence of $T(\beta; P)$ to $T(\beta; \mathcal{P})$ in B at a rate n).

The consistency result under Assumptions 1 and 2 and $\epsilon_n = 0$ follows from Theorem 3.2 in Chernozhukov, Hong, and Tamer (2007) with $a_n = n$. It is not difficult to show that Assumption 3.2 implies condition C.3 in Chernozhukov, Hong, and Tamer (2007), which along with other conditions verified above, implies the consistency result.

The second result follows by redefining the estimation problem as

$$P^* \in \epsilon_n - \arg\min_{\Pi \in \Xi} W(\Pi, P), \quad W(\Pi, P) = \sum_{j,k} \omega_{jk}(P) \left(P_{jk} - \Pi_{jk}\right)^2,$$

where $P^* = (P^*_{jk}, j = 1, ..., J, k = 1, ..., K)$ and Ξ is the space of conditional choice probabilities that are compatible with the model. Under Assumption 1, Ξ is compact, the function $\Pi \mapsto W(\Pi, P)$ is continuous for each P in the neighborhood of $\mathcal{P}$, and therefore $W(\Pi, P) - W(\Pi; \mathcal{P}) = o_p(1)$ uniformly in $\Pi \in \Xi$, as $P = \mathcal{P} + o_p(1)$. Moreover, $\Pi \mapsto W(\Pi, \mathcal{P})$ is uniquely minimized at $\Pi = \mathcal{P}^*$ by assumption. Therefore, by the consistency theorem for approximate argmin estimators, it follows that the ϵ_n-argmin P^* is consistent for $\mathcal{P}^*$. Q.E.D.

PROOF OF THEOREM 13. We consider the upper bounds only, since the proof for lower bounds is analogous. We have that (i) the projection

$$\Pi^* = \Pi^*(\Pi) \in \arg\min_{\tilde{\Pi} \in \Xi} \sum_{j,k} w_{jk}(\Pi)(\Pi_{jk} - \tilde{\Pi}_{jk})^2$$

is continuous at $\mathcal{P}$ by the theorem of the maximum, (ii) the parameter space for β and Π is compact, (iii) the function defining the constraints

$$(\Pi, \beta, \alpha_{k1}, ..., \alpha_{kJ}, \pi_{k1}, ..., \pi_{kJ}) \mapsto \Pi^*_{jk} - \sum_{m=1}^{J} \mathcal{L}(Y^j \mid X^k, \alpha_{km}, \beta)\pi_{km}$$

is continuous by Assumption 1 and the continuity of the projection, and (iv) the criterion function

$$(\Pi, \beta, \alpha_{k1}, ..., \alpha_{kJ}, \pi_{k1}, ..., \pi_{kJ}) \mapsto \sum_{m=1}^{J} [F\left(\tilde{x}'\beta + \alpha_{km}\right) - F\left(\bar{x}'\beta + \alpha_{km}\right)]\pi_{km}$$

is continuous by the assumed continuity of F. Then, using the theorem of the maximum, we conclude that the maximal mapping

$$(\beta, \Pi) \mapsto \overline{\mu}^*_k(\beta, \Pi)$$

is continuous. By Theorem 12 and the extended continuous mapping theorem we have that

$$d_H(B_n, B^*) \to_p 0, \quad P \to_p \mathcal{P}, \quad P^* \to_p \mathcal{P}^*,$$

implies that

$$d_H(\overline{\mu}^*_k(B_n, P), \overline{\mu}^*_k(B^*, \mathcal{P})) \to_p 0,$$

where $\overline{\mu}^*_k(A, \Pi) = \{\overline{\mu}^*_k(\beta, \Pi) : \beta \in A\}$. The conclusion of the theorem then immediately follows. Q.E.D.

Proof of Theorem 14: By the uniform central limit theorem, $W(\mathcal{P}, P)$ converges in law to $\chi^2_{J(K-1)}$ under any sequence of true DGPs with $\mathcal{P}_0$ in Δ. It follows that

$$\lim_{n\to\infty} \Pr_{\mathcal{P}_0}\{\mathcal{P} \in CR_{1-\alpha}(\mathcal{P})\} = 1 - \alpha.$$

Further, the event $\mathcal{P} \in CR_{1-\alpha}(\mathcal{P})$ implies event $\mathcal{P}^* \in \{\Pi^*(\Pi) : \Pi \in CR_{1-\alpha}(\mathcal{P})\}$ by construction, which in turn implies the events $B^* \in CR_{1-\alpha}(B^*)$ and $[\underline{\mu}^*_k, \overline{\mu}^*_k] \in CR_{1-\alpha}[\underline{\mu}^*_k, \overline{\mu}^*_k], \forall k$. Q.E.D.

Proof of Theorem 15. We have that for $S_n(\mathcal{P}) = \hat{\theta} - \theta^* = \hat{\theta} - \theta^*(\mathcal{P})$

$$\begin{aligned}
&\Pr_{\mathcal{P}_0}\{\theta^* \notin \left[\underline{\theta}, \overline{\theta}\right]\} = \Pr_{\mathcal{P}_0}\{S_n(\mathcal{P}) \notin [\underline{G}^{-1}(\alpha_2, \mathcal{P}), \overline{G}^{-1}(1-\alpha_1, \mathcal{P})]\} \\
&\leq \Pr_{\mathcal{P}_0}[\{S_n(\mathcal{P}) \notin [\underline{G}^{-1}(\alpha_2, \mathcal{P}), \overline{G}^{-1}(1-\alpha_1, \mathcal{P})]\} \cap \{\mathcal{P} \in \mathrm{CR}_{1-\gamma}(\mathcal{P})\}] \\
&\quad + \Pr_{\mathcal{P}_0}\{\mathcal{P} \notin \mathrm{CR}_{1-\gamma}(\mathcal{P})\} \\
&\leq \Pr_{\mathcal{P}_0}[\{S_n(\mathcal{P}) \notin [G^{-1}(\alpha_2, \mathcal{P}), G^{-1}(1-\alpha_1, \mathcal{P})]\} \cap \{\mathcal{P} \in \mathrm{CR}_{1-\gamma}(\mathcal{P})\}] \\
&\quad + \Pr_{\mathcal{P}_0}\{\mathcal{P} \notin \mathrm{CR}_{1-\gamma}(\mathcal{P})\} \\
&\leq \Pr_{\mathcal{P}_0}\{S_n(\mathcal{P}) \notin [G^{-1}(\alpha_2, \mathcal{P}), G^{-1}(1-\alpha_1, \mathcal{P})]\} + \Pr_{\mathcal{P}_0}\{\mathcal{P} \notin \mathrm{CR}_{1-\gamma}(\mathcal{P})\} \\
&\leq \alpha + \Pr_{\mathcal{P}_0}\{\mathcal{P} \notin \mathrm{CR}_{1-\gamma}(\mathcal{P})\}.
\end{aligned}$$

Thus if $\limsup_n \Pr_{\mathcal{P}_0}\{\mathcal{P} \notin \mathrm{CR}_{1-\gamma}(\mathcal{P})\} \leq \gamma$, we obtain that $\lim_n \Pr_{\mathcal{P}_0}\{\theta_0 \notin \left[\underline{\theta}, \overline{\theta}\right]\} \leq \alpha + \gamma$, which is the desired conclusion.

It now remains to show that $\limsup_{n\to\infty} \Pr_{\mathcal{P}_0}\{\mathcal{P} \notin \mathrm{CR}_{1-\gamma}(\mathcal{P})\} \leq \gamma$. We have that

$$\Pr_{\mathcal{P}_0}\{\mathcal{P} \notin \mathrm{CR}_{1-\gamma}(\mathcal{P})\} = \Pr_{\mathcal{P}_0}\{W(\mathcal{P}, P) > c_{1-\gamma}(\chi^2_{K(J-1)})\}.$$

By the uniform central limit theorem, $W(\mathcal{P}, P)$ converges in law to $\chi^2_{K(J-1)}$ under any sequence $\mathcal{P}_0$ in Δ. Therefore,

$$\lim_{n\to\infty} \Pr_{\mathcal{P}_0}\{W(\mathcal{P}, P) > c_{1-\gamma}(\chi^2_{K(J-1)})\} = \Pr\{\chi^2_{K(J-1)} > c_{1-\gamma}(\chi^2_{K(J-1)})\} = \gamma.$$

Q.E.D.

11.2 Generic Uniqueness of Projections of Probabilities onto the Model Space

The following lemma is motivated by the analysis of Newey (1986) on generic uniqueness of quasi-maximum likelihood population parameter values.

LEMMA A5. *Let $\mathcal{G}$ be a set of vectors $\Pi = (\Pi_{jk}, j = 1, ..., k, j = 1, ..., J) > 0$ that satisfy the system of linear constraints $\sum_{j=1}^J \Pi_{jk} = 1, \quad k = 1, ..., K$. Let $\mathrm{proj}(\Pi) = \arg\min_{\Pi\in\Xi} d(\Pi, \Pi')$, where $d(\Pi, \Pi') = \left(\sum_{k=1}^K \sum_{j=1}^J \omega_{jk}(\Pi_{jk} - \Pi'_{jk})^2\right)^{1/2}$, be the projection of Π on the set Ξ, where $\omega_{jk} > 0$ for all (j, k) are weights normalized so that d is a proper distance, and $\Xi = \{\Xi(\beta), \beta \in \mathbb{B}\}$ where $\mathbb{B}$ is compact and $\Xi(\beta) = \{\Pi \in \Xi_N : (\Pi_{1k}, ...\Pi_{Jk})' \in \Gamma_k(\beta), \forall k\}$, where Γ_k is defined as in Section 7, with link function F being twice continuously differentiable. The set $\mathcal{G}_0 = \{\Pi \in \mathcal{G} :$ proj(Π) is unique$\}$ is an open dense subset of $\mathcal{G}$.*

Proof: We first note that Ξ is compact, $\Pi' \mapsto d(\Pi, \Pi')$ is continuous, so that the minimum is attainable, and the projection exists. The rest of the proof has two steps: verification of openness of $\mathcal{G}_0$ and verification of denseness of $\mathcal{G}_0$ relative to $\mathcal{G}$.

To verify openness, we take $\Pi_0 \in \mathcal{G}_0$ and find an open neighborhood $\mathcal{N}$ of Π_0 in $\mathcal{G}$ such that $\mathcal{N} \subset \mathcal{G}_0$. We consider two cases. First, if proj(Π_0) is in the interior of Ξ, then there exists an open neighborhood $\mathcal{N}'$ of Π_0 in Ξ. For each Π in $\mathcal{N}$, we necessarily have that proj$(\Pi) = \Pi$, so we can take $\mathcal{N} = \mathcal{N}'$. Second, if Π_0^* is on the boundary of Ξ, the verification follows by an argument similar to that given by Newey (1986), p.7.

To verify denseness, we take $\Pi_0 \in \mathcal{G} \setminus \mathcal{G}_0$, so that proj$(\Pi_0)$ is not unique. For this to happen it must be that $\Pi_0 \notin \Xi$. Take any element Π_0^* of proj(Π_0). Then we can construct a sequence Π_n approaching Π_0 such that proj$(\Pi_n) = \Pi_0^*$, so that $\Pi_n \in \mathcal{G}_0$. Indeed, simply take

$$\Pi_n = \frac{1}{n}\Pi_0^* + \frac{n-1}{n}\Pi_0.$$

Clearly, $\Pi_n \in \mathcal{G}$ and it approaches Π_0. Also, note that by definition Π_0^* is a point of intersection of Ξ with the contour set or ellipse $C_0 = \{\Pi' \in \mathcal{G} : d(\Pi_0, \Pi') = t\}$ for $t = \min_{\tilde{\Pi}=\Xi} d(\Pi_0, \tilde{\Pi})$. Also, note that the contour set or sphere $C_n = \{\Pi' \in \mathcal{G} : d(\Pi_n, \Pi') = t'\}$, where $t' = \min_{\tilde{\Pi}=\Xi} d(\Pi_n, \tilde{\Pi})$ is a strict subset of the sphere C_0, since by convexity of the distance

$$t' \leq d(\Pi_n, \Pi_0^*) \leq \frac{1}{n} d(\Pi_0^*, \Pi_0^*) + \frac{n-1}{n} d(\Pi_0, \Pi_0^*) = \frac{n-1}{n} t \leq t,$$

with only one common point $C_n \cap C_0 = \Pi_0^* \in \Xi$. This establishes that $\text{proj}(\Pi_n) = \Pi_0^*$. Q.E.D.

11.3 Computation

The quadratic problem (26) can be solved using computational techniques developed for finite mixture models such as the EM algorithm or vertex direction methods, see, e.g., Laird (1978), Böhning (1995), Lindsay (1995, Chap. 6) and Aitkin (1999). These iterative algorithms, however, are sensitive to initial values and can be very slow to converge in this problem where we estimate several mixtures over a grid of values for β. Moreover, a slow algorithm is specially inconvenient for the resampling based inference that we develop in the next section. The main computational difficulty in the mixture problems is to find the location of the support points; see, e.g., Aitkin (1999). Since the mixtures are nuisance parameters in our problem, we propose solving the following penalized quadratic problem:

$$\begin{aligned} T_\lambda(\beta; P) &= \min_{\pi_{km}} \sum_{j,k} \left[\omega_{jk} \left(P_{jk} - \sum_{m=1}^{M} \pi_{km} \mathcal{L}\left(Y^j \mid X^k, \alpha_m, \beta\right) \right)^2 + \lambda_n \sum_{m=1}^{M} \pi_{km}^2 \right], \quad (42) \\ &\text{s.t.} \quad \sum_{m=1}^{L} \pi_{km} = 1, \ \pi_{km} \geq 0, \ \forall j, k. \end{aligned}$$

where M is large and λ is small. For the weights, we set $w_{jk} = nP_k / \sum_{m=1}^{L} \tilde{\pi}_{km} \mathcal{L}(Y^j \mid X^k, \alpha_m, \tilde{\beta})$, where $(\tilde{\beta}, \{\tilde{\pi}_{km}, \ \forall (k,m)\})$ is an initial estimate. This is a convex quadratic programming problem for which there are reliable algorithms to find the solution in polynomial time; see, e.g., the `quadprog` package in `R` (Weingessel, 2007). The penalty λ_n acts choosing a distribution among the set of discrete distributions with support contained in a large grid $\{\alpha_1, ..., \alpha_M\}$. In general there is an infinite number of solutions for Q_k, one of them is a discrete distribution with no more than $J << M$ 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 the parameter space $\mathbb{C}$. By making the partition fine enough we guarantee to cover a solution to the problem, without having to find explicitly the location of the support points. The error of the finite grid approximation approaches zero as $M \to \infty$ if $\mathbb{C}$ is compact and the objective function has boundable variation with respect to α_m; see, e.g., Lindsay (1995;

Chap. 6). The penalty favors distributions with large supports. This regularization therefore addresses the computational difficulties created by the non-identifiability of Q_k^*.

The final estimates of the identified sets for the parameters and marginal effects are computed by solving the linear programming problems (24) and (25) for all the parameter values β which satisfy the condition $T_\lambda(\beta; P) \leq \min_\beta T_\lambda(\beta; P) + \epsilon_n$, and replacing the $\mathcal{P}_{jk}$'s by the probabilities predicted by the model P_{jk}^*'s for this parameter value β, defined as in (28).

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] Beresteanu, A., and Molinari, F. (2008), "Asymptotic properties for a class of partially identified models," *Econometrica* 76(4), 763–814.

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

[7] 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.

[8] Böhning, D. (1995), "A Review of Reliable Maximum Likelihood Algorithms for Semiparametric Mixture Models," *Journal of Statistical Planning and Inference* 47, 5–28.

[9] 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.

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

[11] 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.

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

[13] Chamberlain, G. (1982), "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5–46.

[14] Chamberlain, G. (1984), "Panel Data," in Z. Griliches and M. Intriligator eds *Handbook of Econometrics*. Amsterdam: North-Holland.

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

[16] 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.

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

[18] Chernozhukov, V., J.Hahn, and W.K.Newey (2004), "Bound Analysis in Panel Models with Correlated Random Effects," *unpublished manuscript*.

[19] Chernozhukov, V., H. Hong, and E. Tamer (2007), "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica* 75(5), pp. 1243–1284.

[20] 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.

[21] Feller, W. (1943), "On a General Class of Contagious Distributions," *Annals of Statistics,* 14, 389-400.

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

[23] Graham, B.W. and J.L. Powell (2008), "Semiparametric Identification and Estimation of Correlated Random Coefficient Models for Panel Data," *unpublished manuscript*.

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

[25] 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.

[26] Hahn, J., and W. Newey (2004), "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica* 72, 1295-1319.

[27] Hahn, J., and G. Kuersteiner (2007), "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects, "*unpublished manuscript.*

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

[29] Heckman, J. J., and T. E. MaCurdy (1980), "A Life Cycle Model of Female Labor Supply," *Review of Economic Studies* 47, 47-74.

[30] 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.

[31] Honoré, B.E., and E. Tamer (2006), "Bounds on Parameters in Dynamic Discrete Choice Models", *Econometrica* 74(3), 611-629.

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

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

[34] Laird, N. (1978), "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution,"*Journal of the American Statistical Association* 73, 805–811.

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

[36] Manski, C.F., and E. Tamer (2002), "Inference on Regressions with Interval Data on a Regressor or Outcome,"*Econometrica* 70, 519 - 546.

[37] McLachlan, G., and D. Peel (2000), *Finite mixture models.* Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

[38] Newey, W.K.(1986), "Generic Uniqueness of the Population Quasi Maximum Likelihood Estimators," *unpublished manuscript.*

[39] Romano, J. P., and M. Wolf, (2000), "Finite sample nonparametric inference and large sample efficiency," *Annals of Statistics*, 28(3), 756–778.

[40] Rytchkov, O.(2007), *Essays on Predictability of Stock Returns.* Doctoral Dissertation. MIT.

[41] Stokey, N.L., and R.E. Lucas (1989), *Recursive Methods in Economic Dynamics*, Harvard University Press: Cambridge.

[42] Weingessel, A. (2007), *quadprog: Functions to solve Quadratic Programming Problems.* R package version 1.4-11. S original by Berwin A. Turlach. `http://www.r-project.org`.

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

[44] 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.

[45] 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 (0.60)	34.63	-91.20 (0.45)	-91.20	-31.07 (0.77)	-31.07
4	12.77 (0.61)	9.91	-61.52 (0.47)	-59.77	20.52 (0.75)	25.32
8	5.76 (0.62)	0.74	-33.16 (0.49)	-20.40	19.90 (0.74)	30.38

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
kids1990	0.38	
kids1992	0.35	0.31
kids1994	0.45	0.51

Notes: LFP - 1 if woman is in the labor force, 0 otherwise; kid - 1 if woman has any child of age less than 3, 0 otherwise. 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	95% B*	Est.	95% CP	95% MP^+	95% PB^	FE	FE-BC	CMLE	FE
						Logit						
β*					-.36	(-.75, .02)	(-.85, .02)	(-.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)	(-.20, .00)	(-.22, .01)	-.05			
μ(0,1)	.14	-.12	(-.20, -.04)	(-.18, -.06)	-.06	(-.11, .00)	(-.15, .00)	(-.16, .01)	-.07			
μ(1,0)	.17	-.03	(-.10, .05)	(-.08, .03)	-.07	(-.13, .00)	(-.15, .00)	(-.15, .01)	-.07			
μ(1,1)	.21	[-.38, 0]	(-.42, 0)	(-.44, 0)	[-.07, -.05]	(-.15, .00)	(-.18, .00)	(-.20, .01)	-.05			
μ_0		[-.49, -.02]	(-.53, .00)	(-.52, -.01)	[-.06, -.05]	(-.15, .00)	(-.17, .00)	(-.19, .01)	-.06	-.04		-.07
									(-.08, -.04)	(-.06, -.02)		(-.11, -.03)
						Probit						
β*					[-.411, -.409]	(-.85, .03)	(-.88, .04)	(-1.06, .10)	-.88	-.51		
									(-1.24, -.52)	(-.86, -.16)		
μ(0,0)	.48	[-.81, 0]	(-.83, 0)	(-.84, 0)	[-.08, -.04]	(-.20, .00)	(-.21, .01)	(-.24, .02)	-.05			
μ(0,1)	.14	-.12	(-.20, -.04)	(-.18, -.06)	-.07	(-.12, .00)	(-.16, .01)	(-.17, .01)	-.07			
μ(1,0)	.17	-.03	(-.10, .05)	(-.08, .03)	-.07	(-.13, .01)	(-.14, .01)	(-.15, .02)	-.07			
μ(1,1)	.21	[-.38, 0]	(-.42, 0)	(-.44, 0)	[-.07, -.05]	(-.16, .01)	(-.16, .01)	(-.19, .02)	-.06			
μ_0		[-.49, -.02]	(-.53, .00)	(-.52, -.01)	[-.07, -.05]	(-.17, .00)	(-.18, .01)	(-.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. N denotes nornal approximation; B denotes nonparametric bootstrap; CP denotes canonical projection; MP denotes modified projection; PB denotes perturbed bootstrap; FE denotes fixed effects maximum likelihood estimator (FEMLE); FE-BC denotes bias corrected FEMLE; CMLE denotes conditional FEMLE; Linear FE denotes the linear within groups estimator. *200 boostraps repetitions. $^+$Based on 50,000 DGPs. ^Based on 100 DGP's and 200 simulations for each DGP.

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

		General CEF Model			Semiparametric Model								Linear
	$P(X^k)$	Est.	95% N	95% B*	Est.	95% CP	95% MP^+	95% PB^	FE	FE-BC	Jack.	CMLE	FE
						Logit							
β*					-.42	-	(-.76, -.07)	(-.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]	-	(-.14, -.01)	(-.14, -.02)	-.06				
μ(0,0,1)	.08	-.12	(-.21, -.04)	(-.20, -.06)	-.07	-	(-.13, -01)	(-.14, -.01)	-.07				
μ(0,1,0)	.06	-.1	(-.20, .01)	(-.17, .01)	-.08	-	(-.16, -.01)	(-.17, -.02)	-.09				
μ(1,0,0)	.14	-.06	(-.14, .01)	(-.10, .01)	-.08	-	(-.15, -.01)	(-.15, -.02)	-.09				
μ(0,1,1)	.08	-.18	(-.26, -.09)	(-.23, -.09)	-.07	-	(-.14, -.01)	(-.16, -.02)	-.08				
μ(1,0,1)	.03	.02	(-.16, .20)	(-.02, .22)	-.08	-	(-.17, -.01)	(-.19, -.02)	-.09				
μ(1,1,0)	.12	-.04	(-.12, .04)	(-.15, -.01)	-.07	-	(-.13, -.02)	(-.15, -.02)	-.08				
μ(1,1,1)	.09	[-.41, 0]	(-.46, 0)	(-.42, 0)	[-.08, -.07]	-	(-.14, -.01)	(-.16, -.02)	-.08				
μ_0		[-.40, -.04]	(-.46, .00)	(-.41, -.02)	[-.07, -.07]	-	(-.13, -.01)	(-.13, -.02)	-.08	-.07	-.09		-.08
									(-.09, -.06)	(-.09, -.05)	(-.11, -.07)		(-.11, -.06)
						Probit							
β*					[-.462, -.460]	-	(-.74, -.17)	(-.73, -.16)	-.78	-.55	-.38		
									(-.99, -.57)	(-.75, -.35)	(-.58, -.18)		
μ(0,0,0)	.4	[-.81, 0]	(-.83, 0)	(-.81, 0)	[-.08, -.06]	-	(-.15, -.02)	(-.15, -.02)	-.06				
μ(0,0,1)	.08	-.12	(-.21, -.04)	(-.20, -.06)	-.08	-	(-.15, -.03)	(-.16, -.02)	-.07				
μ(0,1,0)	.06	-.1	(-.20, .01)	(-.17, .01)	-.08	-	(-.16, -.02)	(-.16, -.03)	-.08				
μ(1,0,0)	.14	-.06	(-.14, .01)	(-.10, .01)	-.09	-	(-.17, -.03)	(-.15, -.02)	-.09				
μ(0,1,1)	.08	-.18	(-.26, -.09)	(-.23, -.09)	-.08	-	(-.14, -.03)	(-.17, -.02)	-.08				
μ(1,0,1)	.03	.02	(-.16, .20)	(-.02, .22)	-.08	-	(-.16, -.02)	(-.22, -.02)	-.09				
μ(1,1,0)	.12	-.04	(-.12, .04)	(-.15, -.01)	-.07	-	(-.13, -.03)	(-.15, -.02)	-.09				
μ(1,1,1)	.09	[-.41, 0]	(-.46, 0)	(-.42, 0)	[-.09, -.07]	-	(-.15, -.03)	(-.17, -.02)	-.08				
μ_0		[-.40, -.04]	(-.46, .00)	(-.41, -.02)	[-.08, -.07]	-	(-.14, -.03)	(-.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. N denotes nornal approximation; B denotes nonparametric bootstrap; CP denotes canonical projection; MP denotes modified projection; PB denotes perturbed bootstrap; FE denotes fixed effects maximum likelihood estimator (FEMLE); FE-BC denotes bias corrected FEMLE; Jack. denotes panel Jackknife bias corrected FEMLE; CMLE denotes conditional FEMLE; Linear FE denotes the linear within groups estimator. *200 boostraps repetitions. $^+$Based on 50,000 DGPs. ^Based on 100 DGP's and 200 simulations for each DGP.

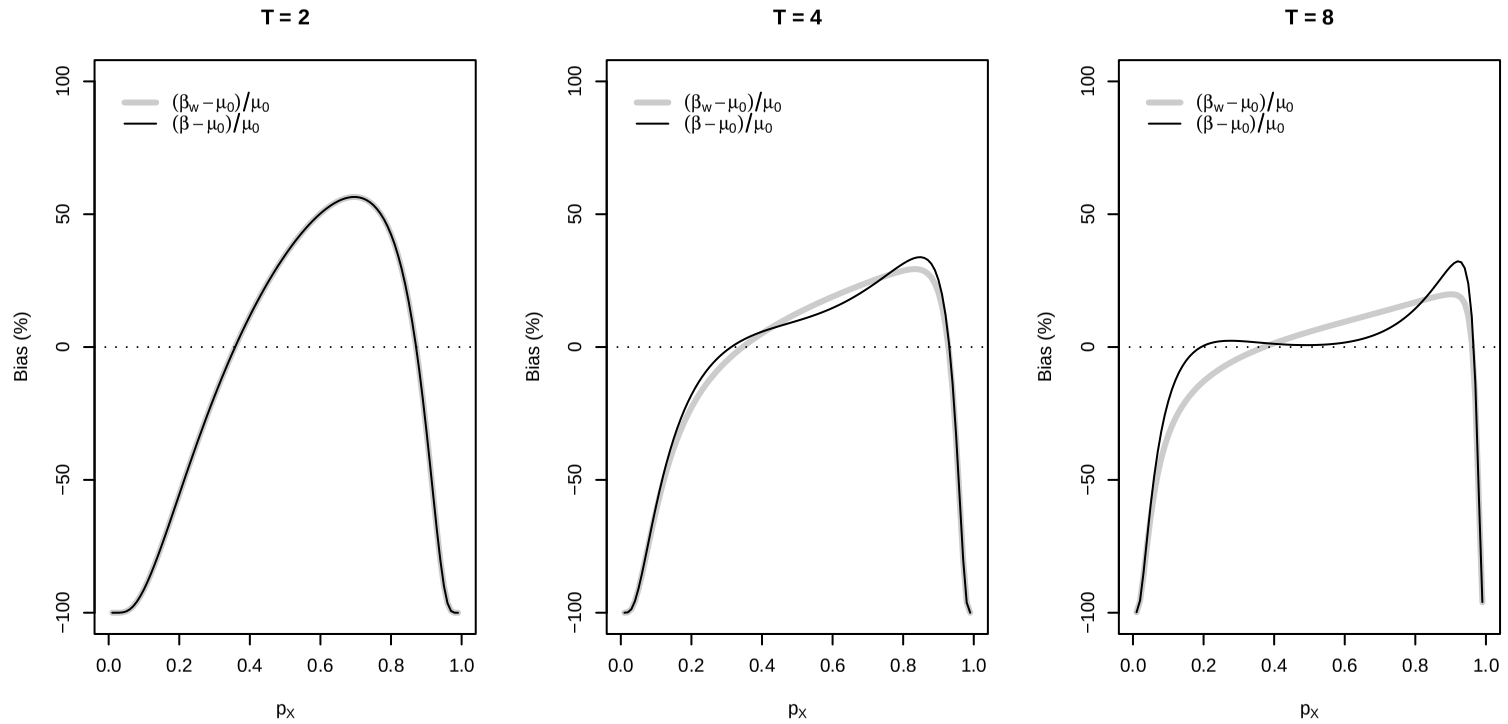

Figure 1: Biases of linear probability model estimators in percentage of marginal effect. Probit model with a single binary regressor with parameter equal to one and individual location effect. Individual effect is the standardized individual mean of the regressor.

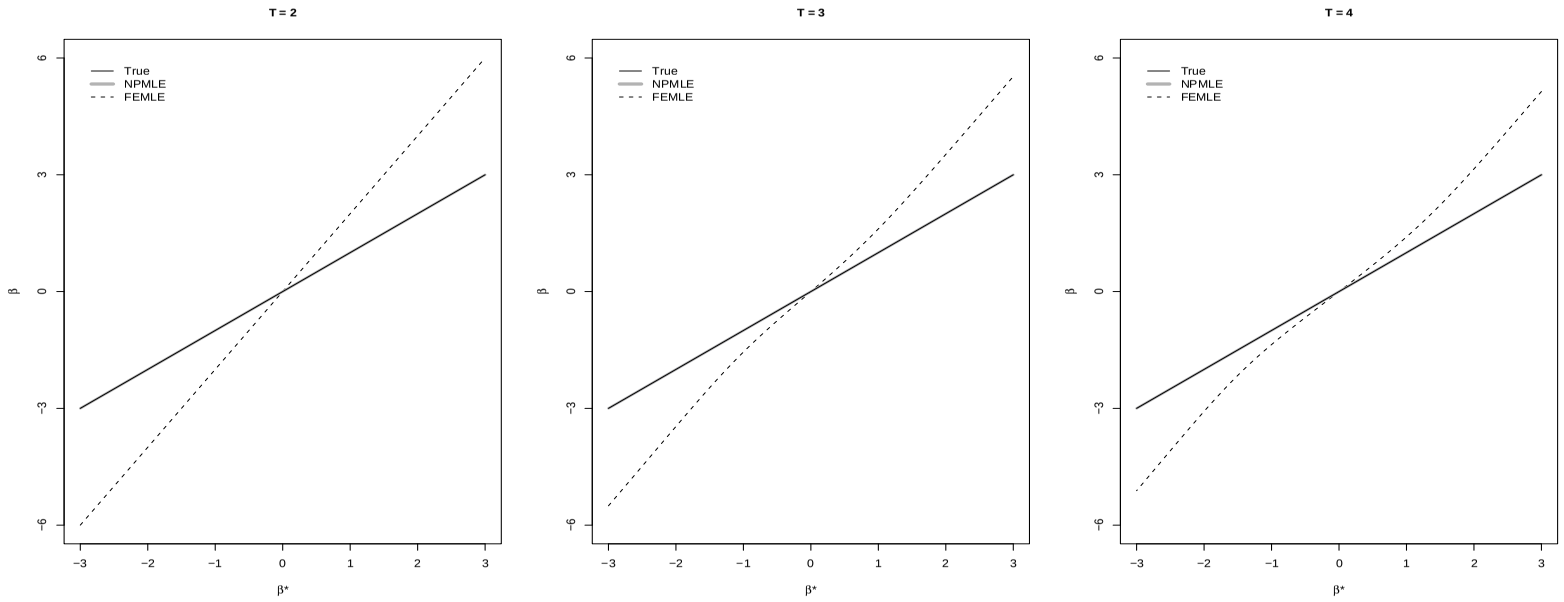

Figure 2: Logit model: Nonparametric MLE identification sets for model parameter β_0 and probability limits of fixed effects estimators.

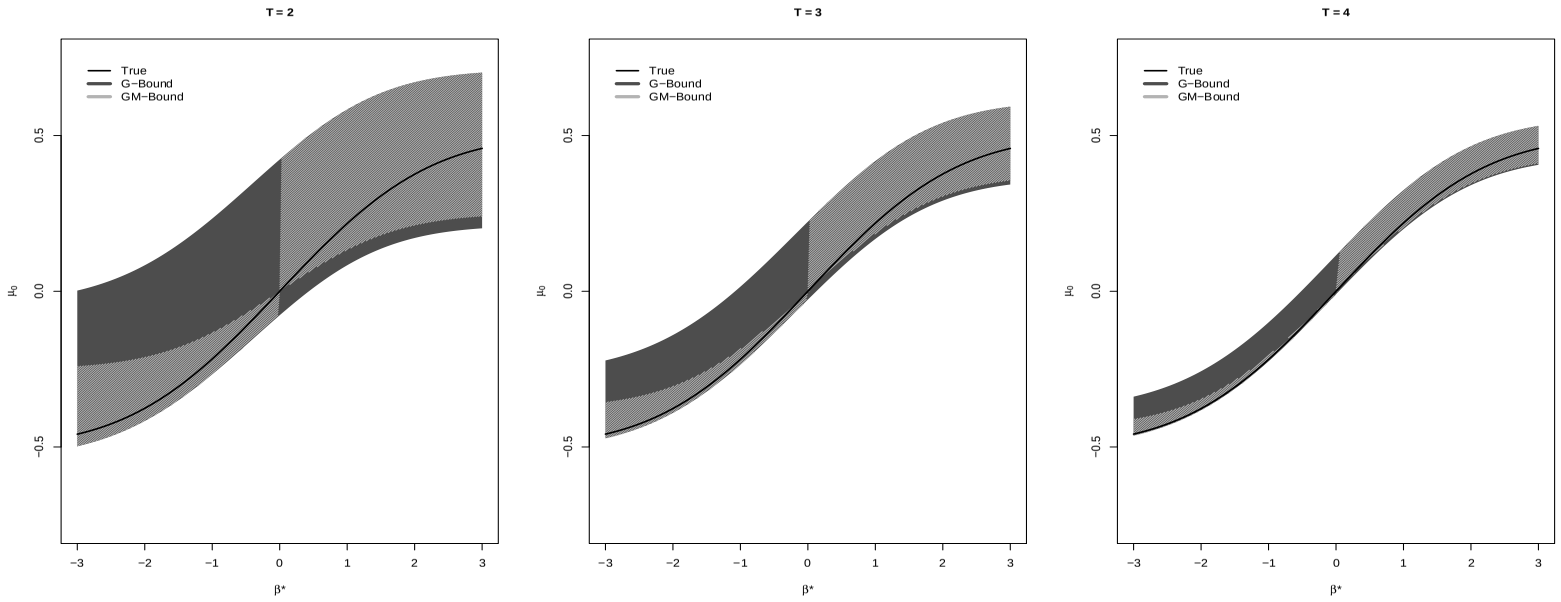

Figure 3: Logit model: Identification sets for average marginal effects μ_0 based on general model. G-bounds are obtained using equation (6) and GM-bounds impose monotonicity of the marginal effects.

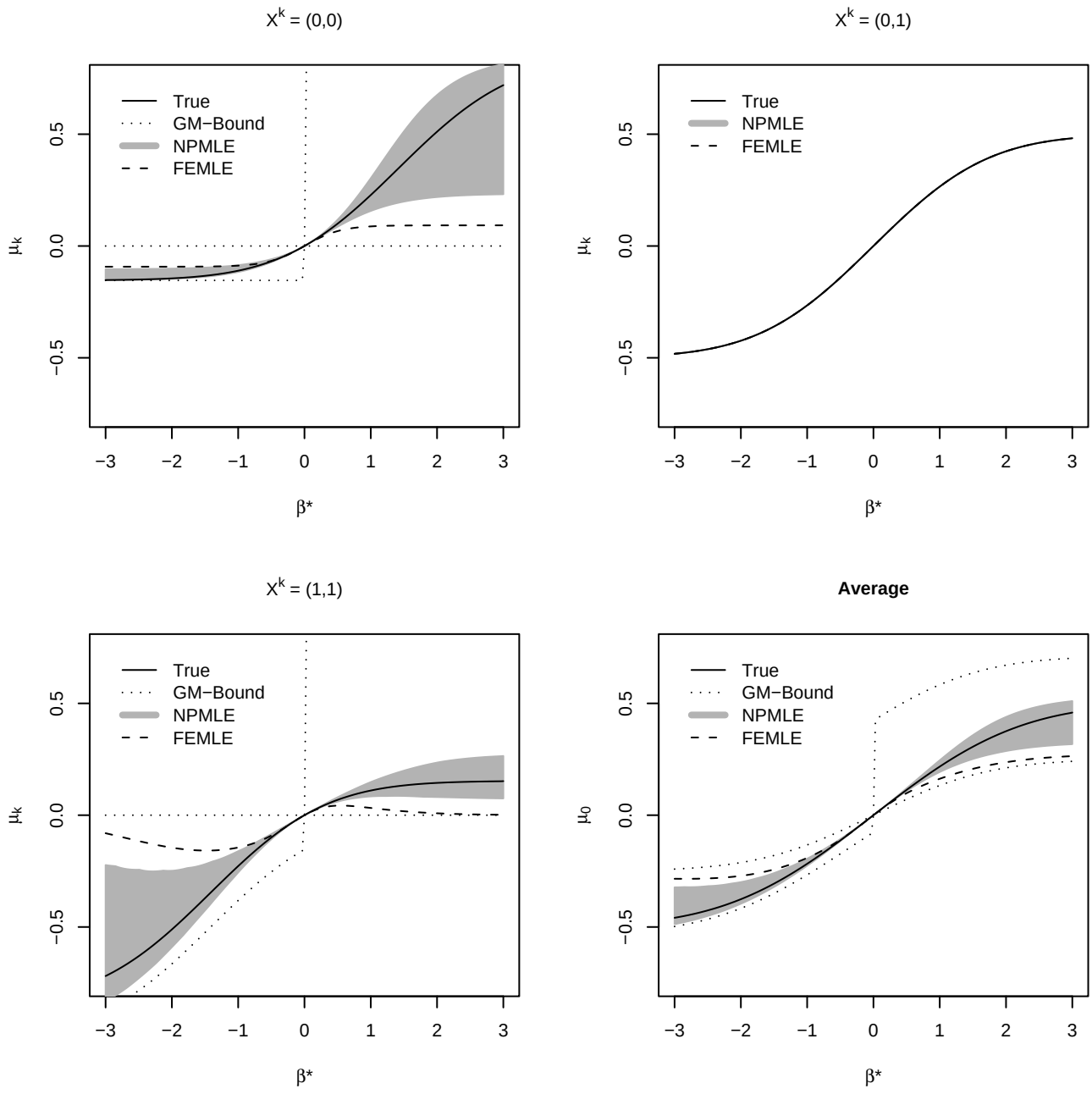

Figure 4: Logit model (T = 2). Identification sets for marginal effects and probability limits of fixed effects estimators.

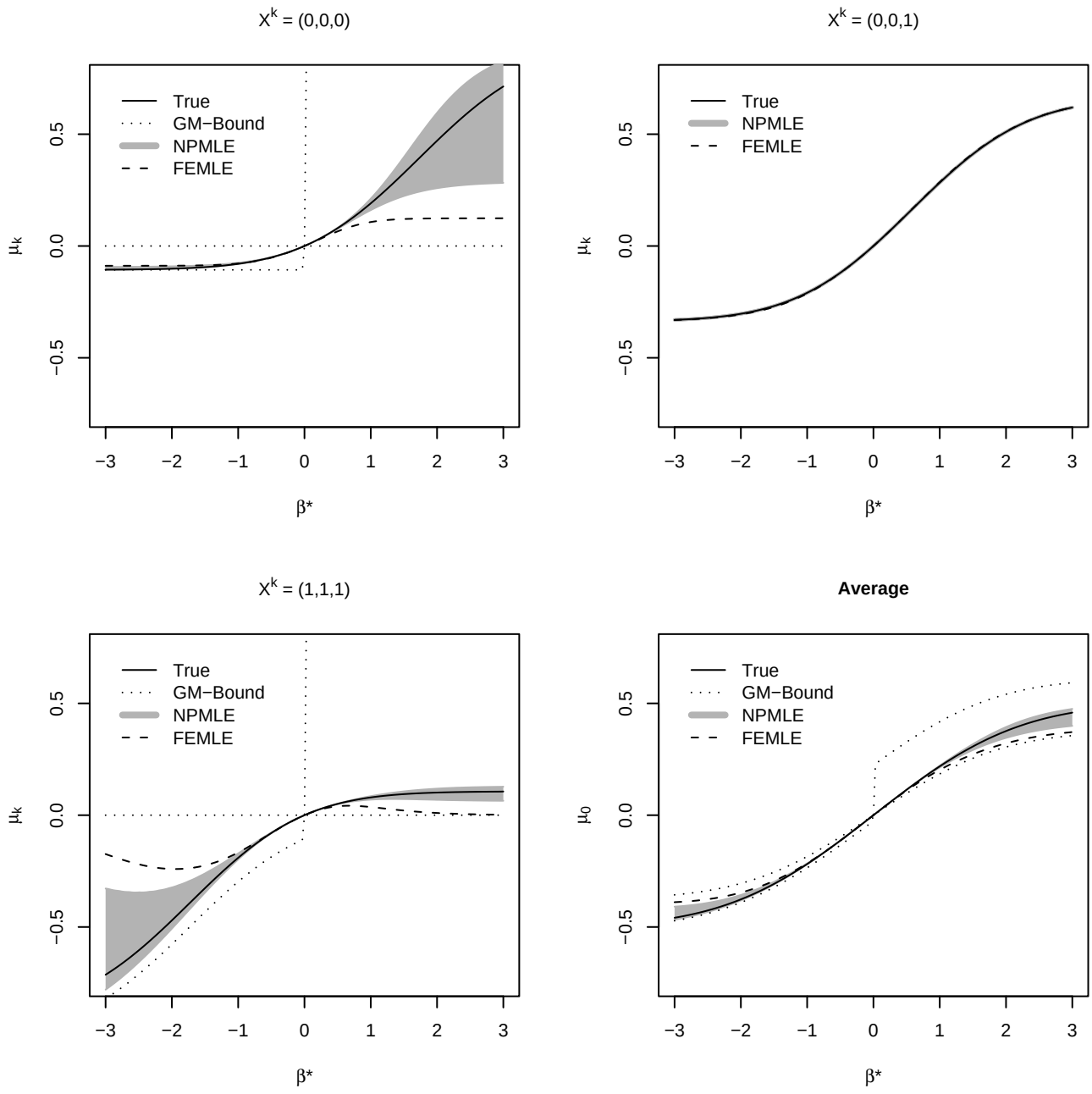

Figure 5: Logit model (T = 3). Identification sets for marginal effects and probability limits of fixed effects estimators.

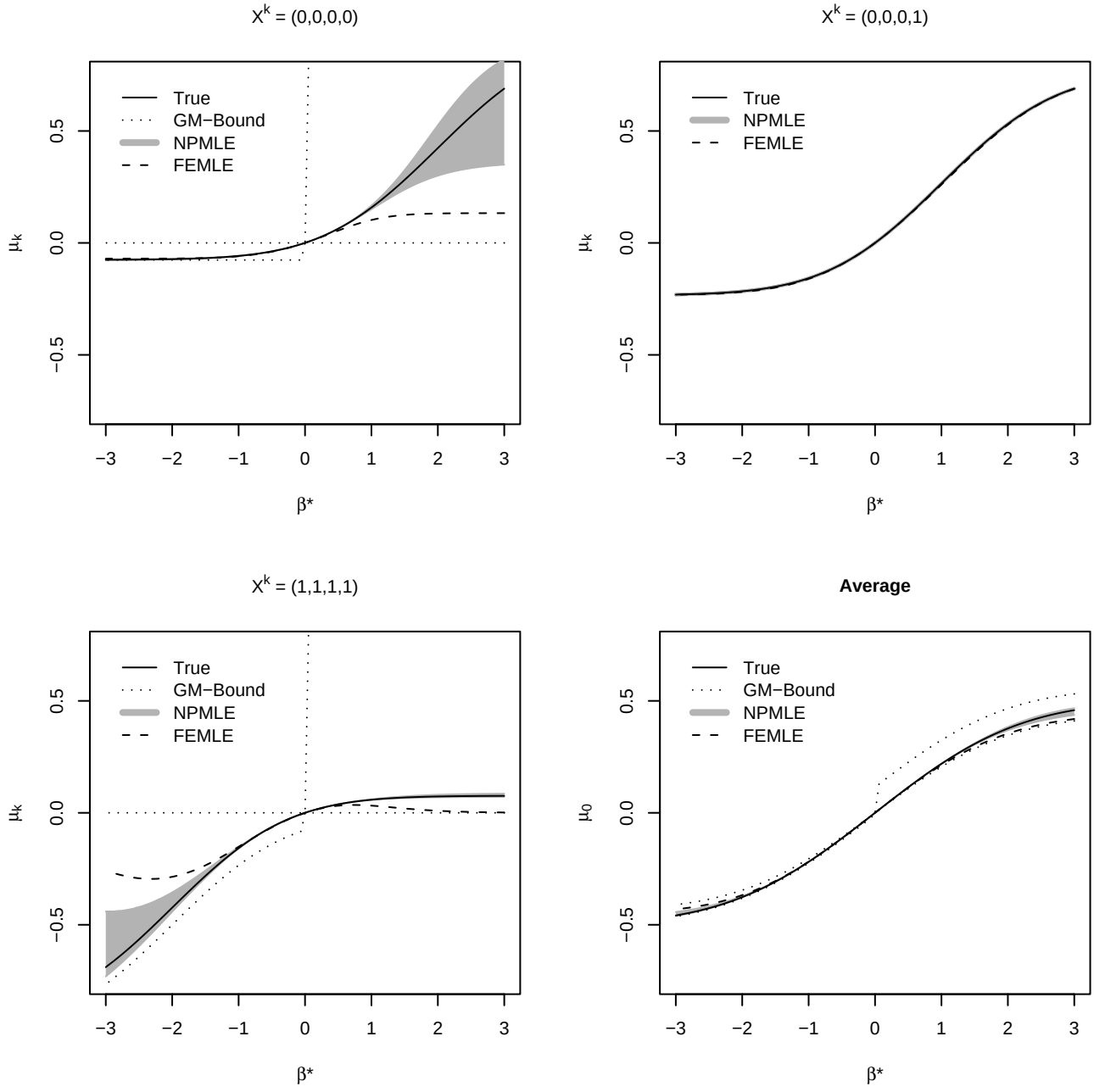

Figure 6: Logit model (T = 4). Identification sets for marginal effects and probability limits of fixed effects estimators.

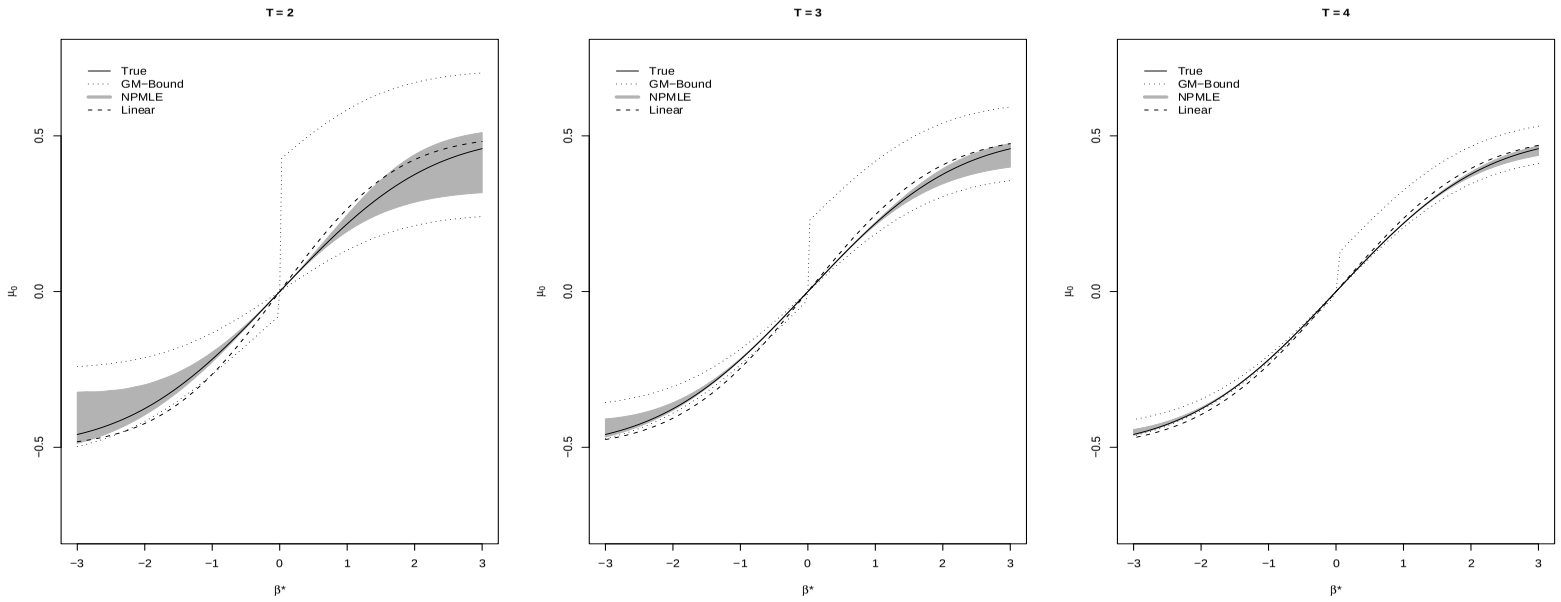

Figure 7: Logit model: Identification sets for average marginal effects and probability limits of linear model estimators.

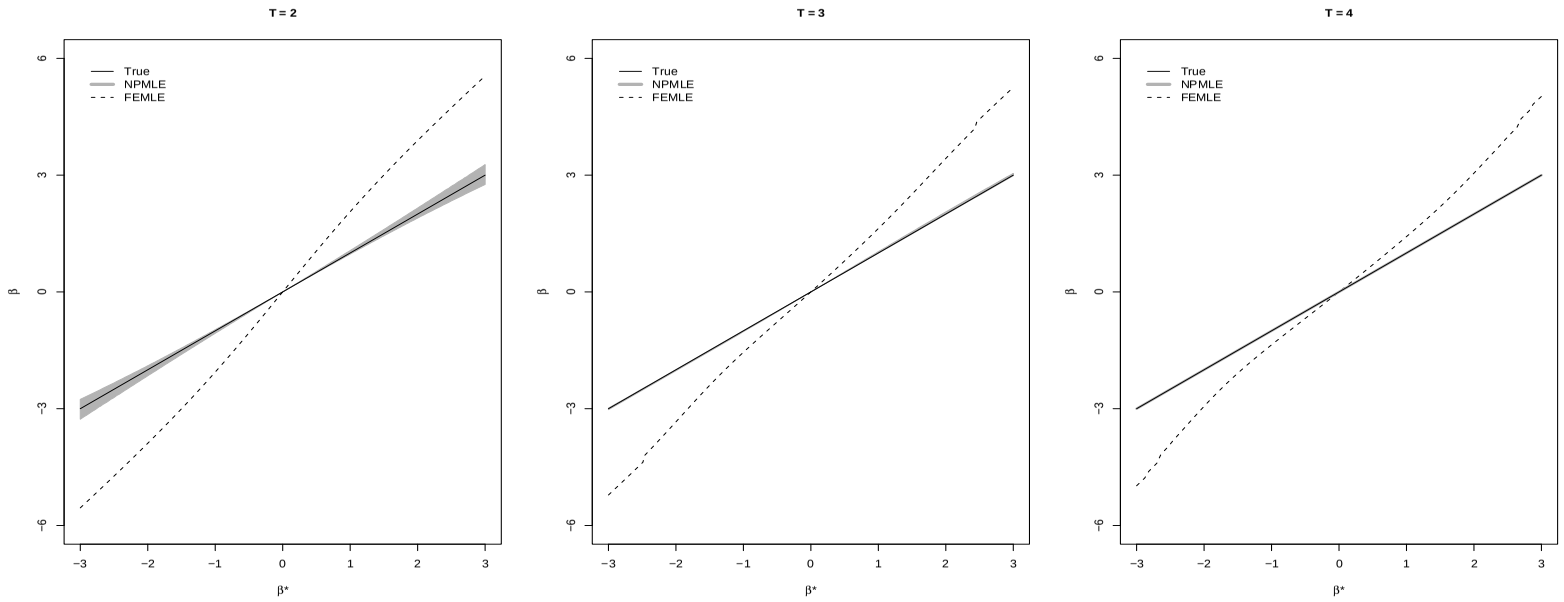

Figure 8: Probit model: Nonparametric MLE identification sets for model parameter β_0 and probability limits of fixed effects estimators.

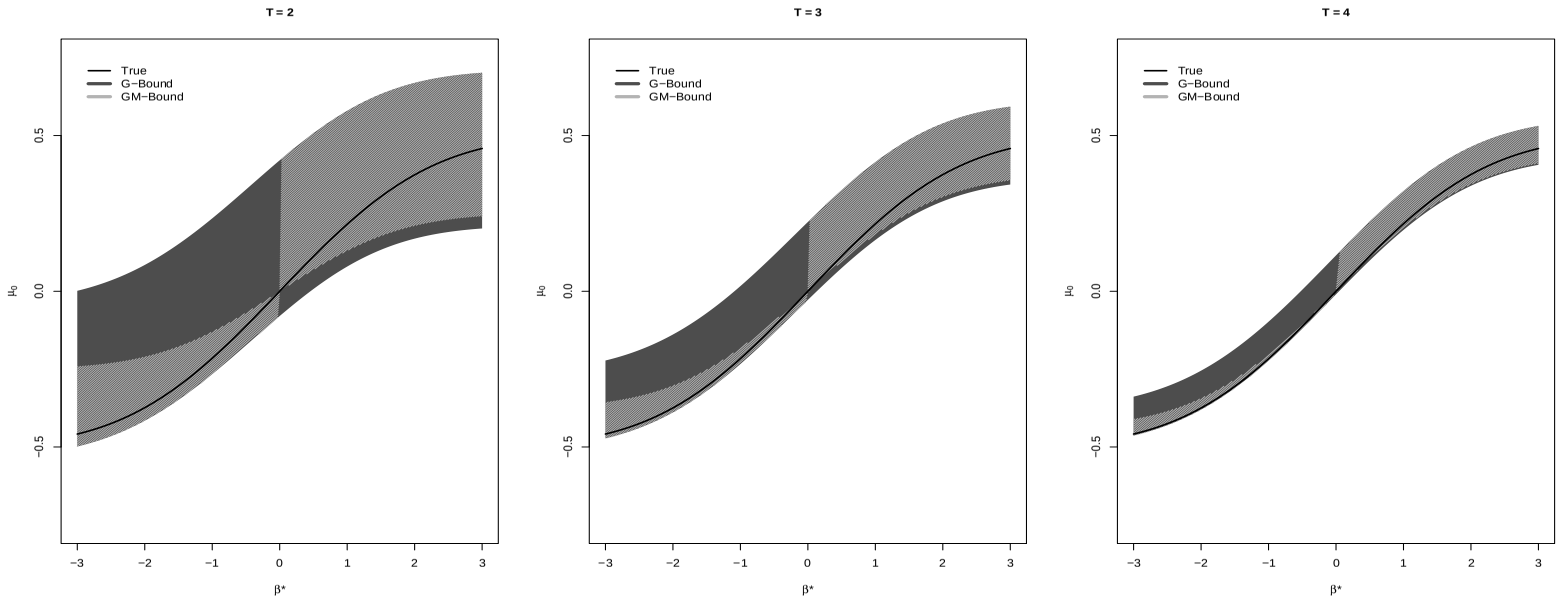

Figure 9: Probit model: Identification sets for average marginal effects μ_0 based on general model. G-bounds are obtained using equation (6) and GM-bounds impose monotonicity of the marginal effects.

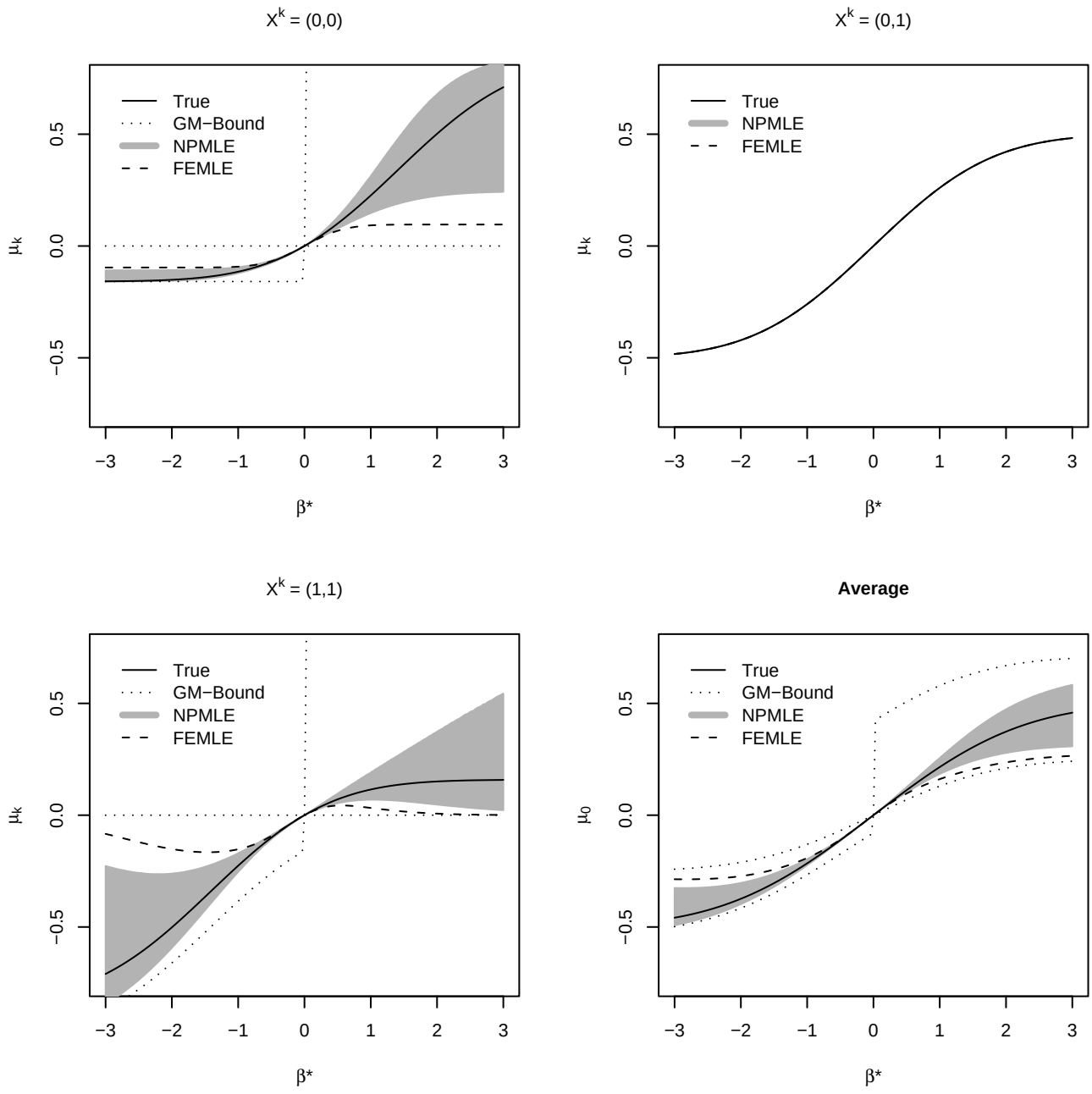

Figure 10: Probit model (T = 2). Identification sets for marginal effects and probability limits of fixed effects estimators.

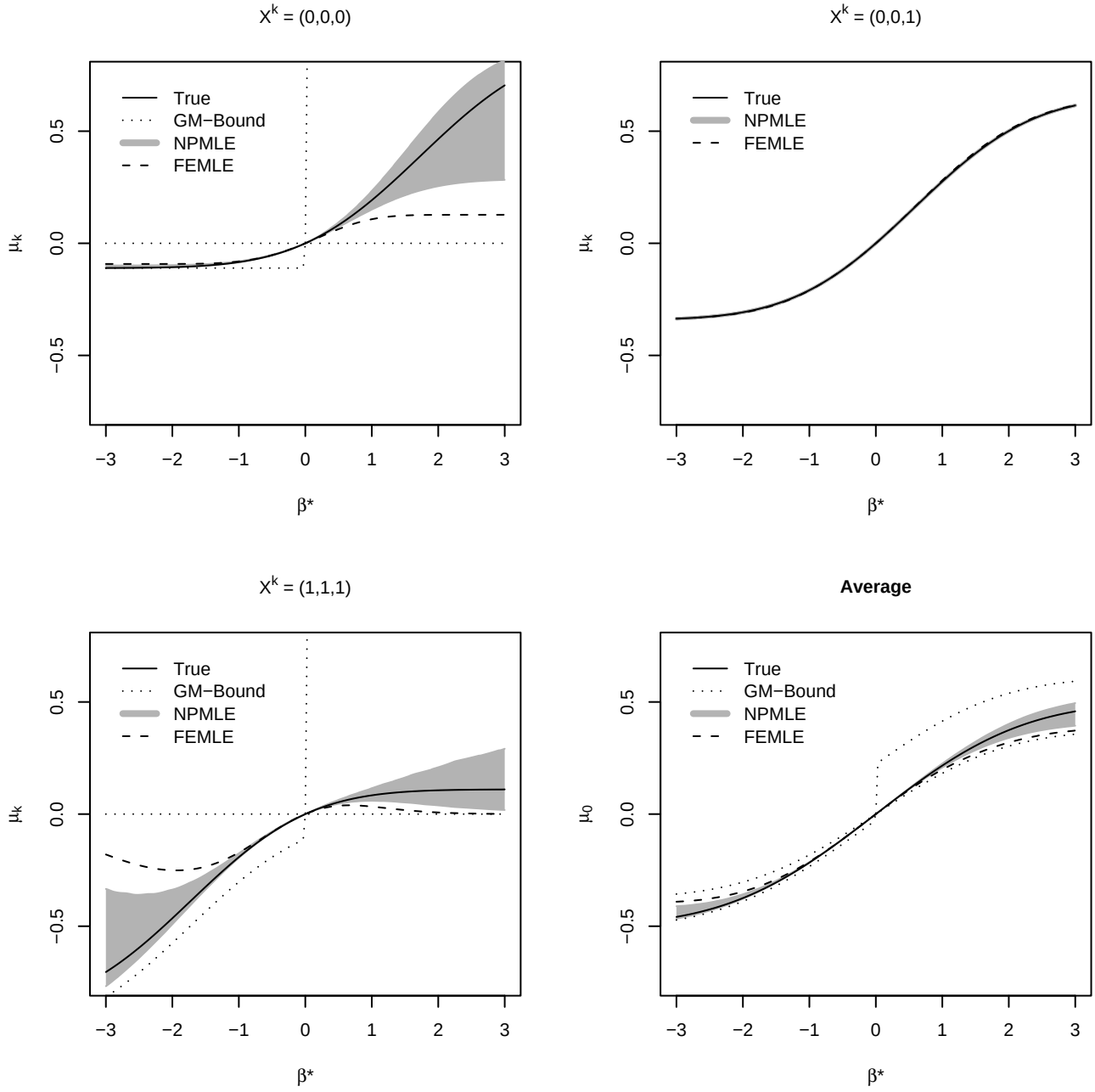

Figure 11: Probit model (T = 3). Identification sets for marginal effects and probability limits of fixed effects estimators.

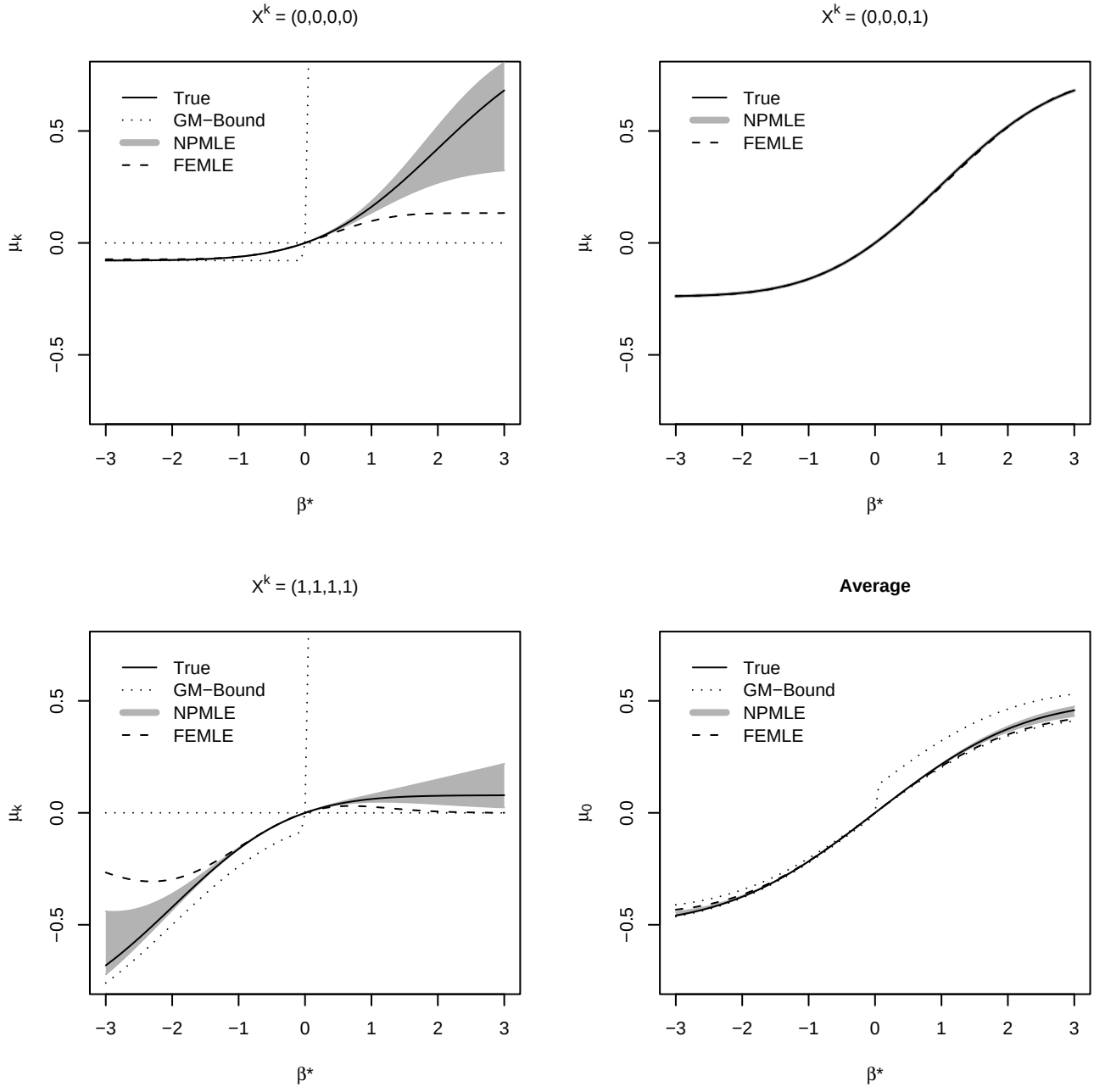

Figure 12: Probit model (T = 4). Identification sets for marginal effects and probability limits of fixed effects estimators.

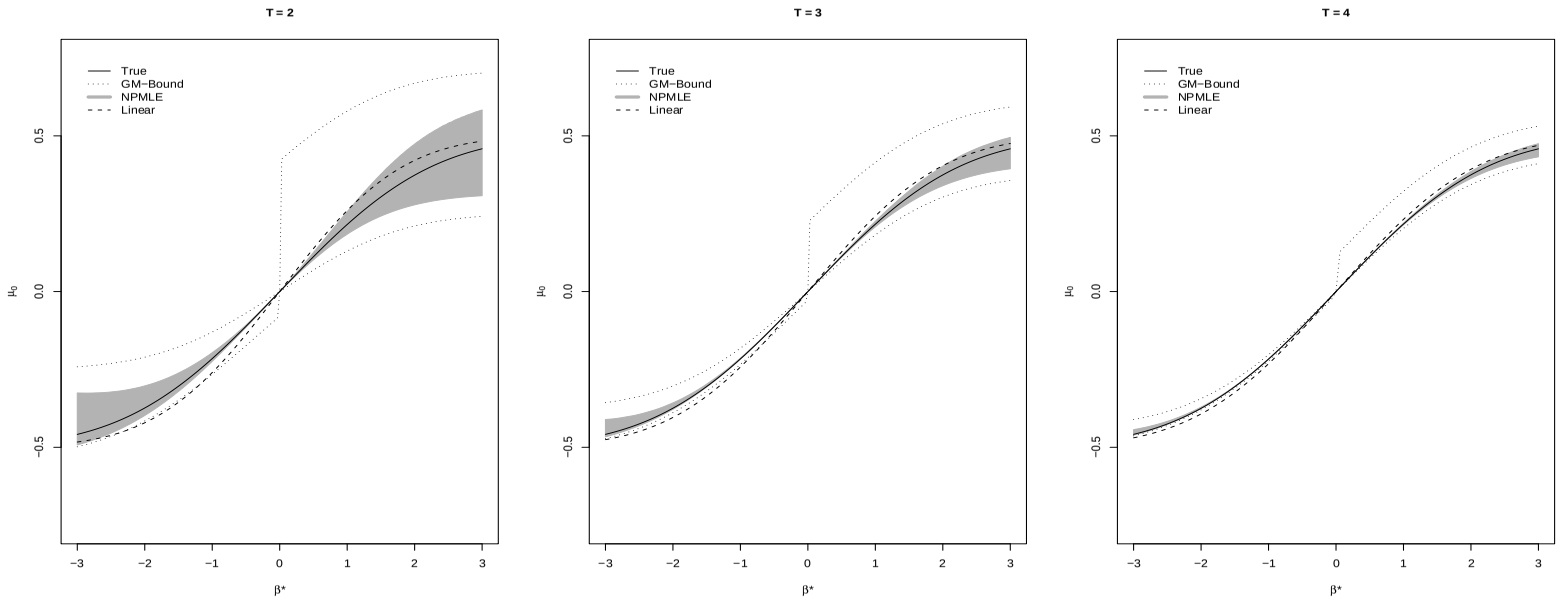

Figure 13: Probit model: Identification sets for average marginal effects and probability limits of linear model estimators.