Average and Quantile Effects in Nonseparable Panel Models [1]

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

November 6, 2010

[1]We thank J. Angrist, G. Chamberlain, B. Graham, J. Hausman, and many seminar participants for comments. Brad Larsen provided capable research assistance. Parts of this paper were given at the 2007 CEMMAP Microeconometrics: Measurement Matters Conference, the Shanghai Lecture of the 2010 World Congress of the Econometric Society, and conferences inbetween. We gratefully acknowledge research support from the NSF.

Abstract

This paper gives identification and estimation results for average and quantile effects in nonseparable panel models. Nonseparable models are important for modeling in a variety of economic settings, including discrete choice. We find that linear fixed effects estimators are not consistent for the average effect, due in part to that effect not being identified. Nonparametric bounds for quantile and average effects are derived for discrete regressors that are strictly exogenous or predetermined. We allow for location and scale time effects and show how monotonicity can be used to shrink the bounds. We derive rates at which the bounds tighten as the number T of time series observations grows. We also consider semiparametric discrete choice models and find that the bounds for average effects tighten considerably. In numerical calculations we find that the bounds may be very tight for small numbers of observations, suggesting their use in practice. We propose two novel inference methods for parameters defined as solutions to linear and nonlinear programs such as average effects in multinomial choice models. We show that these methods produce uniformly valid confidence regions in large samples. We give empirical illustrations.

1 Introduction

Interesting empirical questions are often formulated in term of the ceteris paribus effect of x on y, when observed x is an individual choice variable partly determined by preferences or technology. Panel data holds out the hope of controlling for individual preferences or technology by using multiple observations for a single economic agent. This hope is particularly difficult to realize with discrete or other nonseparable models and/or multidimensional individual effects. These models are, by nature, not additively separable in unobserved individual effects, making them challenging to identify and estimate. There are some simple solutions, such as the conditional MLE for the slope parameter of a logit model with an individual location effect. However these are rare and dependent on specific models or distributions.

A fundamental idea for using panel data to identify the ceteris paribus effect of x on y is to use changes in x over time to estimate the effect. In order for changes over time in x to correspond to ceteris paribus effects, the distribution of variables other than x must not vary over time. This restriction is like "time being randomly assigned." In this paper we consider identification via such time homogeneity conditions. They are also the basis of many previous panel results, including Chamberlain (1982), Manski (1987), and Honore (1992). Here we consider the identifying power of time homogeneity for general nonseparable models and for semiparametric discrete choice models. We also allow for multidimensional heterogeneity, as motivated by the economic and empirical examples of Browning and Carro (2007). Because time homogeneity can include homoskedasticity over time, that often does not hold in applications, we also allow for some time effects.

Models with discrete regressors have many applications and are the subject of most of this paper. With discrete regressors, time homogeneity only leads to partial identification of nonparametric quantile and average treatment effects, as shown by Chamberlain (1982) for the average effect in an important example, and here more generally. Recently Graham and Powell (2008) and Hoderlein and White (2009) have used time homogeneity to obtain some identification results in nonseparable models with continuous regressors.

We show that average and quantile effects on those individuals for whom x changes are identified, as Chamberlain (1982) did for the average effect in a linear random coefficient example. We give simple estimators for the identified effects. We also find that linear fixed effects estimates a time variance weighted average effect rather than the average effect. We also show that overall quantile and average effects are not identified but that the data can be informative about them. To show how much information time homogeneity provides with discrete regressors we derive sharp bounds for quantile and average treatment effects in a static nonparametric model. We also derive bounds for these effects in a dynamic model that allows for lagged dependent variables.

The dynamic bounds provide a partial solution to the problem of estimating state dependence in the presence of unobserved heterogeneity. In an Appendix we allow for location and scale time effects in a static model and show how monotonicity can be used to shrink the bounds. We derive rates at which the nonparametric bounds tighten as the number T of time series observations grows, obtaining exponential rates in some interesting cases. We show in examples that the nonparametric bounds can be informative but also may be quite wide.

The width of the nonparametric bounds motivates models that impose more restrictions, leading to tighter bounds. To that end we consider semiparametric discrete choice models, where the conditional distribution of the individual effects is unknown. We find that in a semiparametric binary choice model with a location individual effect, the bounds are much tighter than in the nonparametric model. In numerical calculations we find that the bounds may be very tight even for small numbers of time periods, suggesting they should be informative in practice. For the logit case of this model we also derive an exponential convergence rate for the average partial effect bounds as the number of time periods T grows. The semiparametric models we consider are also flexible enough to accommodate multiple sources of individual heterogeneity, although we leave to future research numerical calculations, empirical examples, and the derivation of rates as T grows for such cases.

We show that semiparametric discrete choice models have finite dimensional parameterizations. This reduces bounds calculation and estimation to a finite dimensional problem, albeit a large dimensional, highly nonlinear, and computationally difficult one. To make computation more feasible we use grids of fixed values for individual effects, so that average choice probabilities are finite dimensional linear combinations. We combine this with minimum squared distance fitting of data cell probabilities to obtain a quadratic programming approach for estimating the individual effect distributions. This approach is computationally convenient and overcomes problems with previously proposed methods, as further discussed below. We also allow the grid to grow in order to approximate the true support points. It turns out that because the model is finite dimensional there is no need to limit the number of grid points. Mathematically, a richer fixed grid simply corresponds to a bigger submodel of the finite dimensional model.

The semiparametric bounds build on Honoré and Tamer (2006) and Chernozhukov, Hahn, and Newey (2004). Both papers gave results for bounds in semiparametric nonlinear panel data models. Honore and Tamer (2006) proposed linear programming, minimum distance, and maximum likelihood methods for dynamic models. Chernozhukov, Hahn, and Newey (2004) proposed sieve likelihood estimation of bounds for static models. These approaches are not very useful for estimation. Plugging in sample frequencies in place of cell probabilities in the linear programming algorithm produces empty identification regions because the frequencies need not satisfy constraints imposed by the model. Also, the minimum distance objective

function is computationally difficult, as is sieve maximum likelihood, given the dimensionality of the individual effect distributions. Honore and Tamer (2006) also assumed a fixed known grid for true individual effects, while we consider an approximation to an unknown grid.

The inferential problem for the semiparametric models is also rather challenging. The models impose data-dependent constraints that are often infeasible in finite samples or under misspecification, which produces empty confidence regions. We overcome these difficulties by projecting these data-dependent constraints onto the model space using the quadratic programming approach mentioned above, thus producing an always feasible data-dependent constraint set. We then suggest 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 a best approximating model under incorrect specification. We also 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.

We give two empirical illustrations. One is to estimation of the effect of unions on earnings quantiles. There we obtain static and dynamic estimates, finding that a decline in the union effect as the quantile increases can be attributed to individual heterogeneity. The other illustration is to estimation of the effects of fertility on women's labor force participation. There we compare nonparametric and semiparametric estimates.

Another useful assumption for panel data is existence of a control variable, where conditioning on that variable makes x and the individual effects independent, also referred to as correlated random effects. This condition has been used by Chamberlain (1980, 1984), Altonji and Matzkin (2005), and Bester and Hansen (2008). This is a powerful assumption that leads to relatively simple estimators of interesting identified effects, but it does restrict dependence between individual effects and regressors. We try to avoid such restrictions and focus instead on time homogeneity.

Bias corrected fixed effects estimation of semiparametric models has been proposed by Hahn and Kuersteiner (2002), Alvarez and Arellano (2003), Woutersen (2002), Hahn and Newey (2004), and Fernández-Val (2009). These estimators have good theoretical properties for large T and work well in many examples. In an Appendix we show that with small T, nonlinear fixed effects consistently estimates the identified average effect for those whose x changes. However, these methods are dependent on large T, while the bounds analysis we give applies to any T.

Section 2 describes the models and effects we consider. Section 3 discusses estimation of identified effects. Section 4 and 5 derive bounds for the static and dynamic nonparametric models respectively. Section 6 considers identification and rates as T grows. Section 7 describes and gives results for semiparametric discrete choice models. Section 8 gives results and numerical

examples on calculation of population bounds. Section 9 discusses estimation and Section 10 inference for semiparametric models. Section 11 gives the empirical examples. The Appendix contains results that allow for time effects, impose monotonicity, and other results, as well as proofs.

2 The Models and Effects

The data consist of n observations on $Y_i = (Y_{i1}, ..., Y_{iT})'$ and $X_i = [X_{i1}, ..., X_{iT}]'$, for a dependent variable Y_{it} and a vector of regressors X_{it}. Throughout we assume that the observations (Y_i, X_i), $(i = 1, ..., n)$, are independent and identically distributed. The nonparametric models we consider satisfy

ASSUMPTION 1: *There is a function $g_0(x, \alpha, \varepsilon)$ and vectors α_i and $\varepsilon_{it}, (t = 1, ..., T)$ of random variables such that*

$$Y_{it} = g_0(X_{it}, \alpha_i, \varepsilon_{it}), (i = 1, ..., n; t = 1, ..., T).$$

The vector α_i consists of time invariant individual effects that often represent individual heterogeneity. The vector ε_{it} represents period specific disturbances. Altonji and Matzkin (2005) considered models satisfying Assumption 1. The invariance of g_0 over time in this Assumption does not actually impose any time homogeneity. If there are no restrictions on ε_{it} then t could be one of the components of ε_{it}, allowing the function to vary over time in a completely general way. The next condition together with Assumption 1 imposes time homogeneity on the conditional distribution of ε_{it}.

ASSUMPTION 2: $\varepsilon_{it}|X_i, \alpha_i \stackrel{d}{=} \varepsilon_{i1}|X_i, \alpha_i$, *for all t.*

This is a static, or "strictly exogenous" time homogeneity condition, where all leads and lags of the regressor are included in the conditioning variable X_i. It requires that the conditional distribution of ε_{it} given X_i and α_i does not depend on t, but does allow for dependence of ε_{it} over time. An equivalent condition is $\tilde{\varepsilon}_{it}|X_i \stackrel{d}{=} \tilde{\varepsilon}_{i1}|X_i$ for $\tilde{\varepsilon}_{it} = (\alpha_i, \varepsilon_{it})$. Thus, the time invariant α_i has no distinct role in this model. The condition is just that whatever the unobserved disturbances are, their conditional distribution given X_i does not depend on t.

This seems a basic condition that helps panel data provide information about the effect of x on y. It is like the time period being "randomly assigned," with the distribution of factors other than x not varying over time, so that changes in x over time can help identify the effect of x on y. It also turns out to be a natural strengthening of linear model conditions. To see this consider a linear model with $Y_{it} = X_{it}'\beta_0 + \tilde{\varepsilon}_{it}$ and $E^*(\tilde{\varepsilon}_{it}|X_i) = E^*(\tilde{\varepsilon}_{i1}|X_i)$ for all t, where $E^*(\tilde{\varepsilon}_{it}|X_i)$ is the

linear projection of $\tilde{\varepsilon}_{it}$ on X_i (assuming all second moments exist). The invariance of $E^*(\tilde{\varepsilon}_{it}|X_i)$ to t is a linear version of Assumption 2 with time homogeneity of the linear projection replacing time homogeneity of the conditional distribution. Then since $\varepsilon_{it} = \tilde{\varepsilon}_{it} - E^*(\tilde{\varepsilon}_{it}|X_i)$ is orthogonal to all rows of X_i,

$$Y_{it} = X_{it}'\beta_0 + \alpha_i + \varepsilon_{it}; E[X_{is}\varepsilon_{it}] = 0, \forall s, t; \alpha_i = E^*[\tilde{\varepsilon}_{i1}|X_i].$$

This is a standard linear model with an additive individual effect α_i and an idiosyncratic disturbance ε_{it} that is orthogonal to all leads and lags of the regressors. Thus, since a linear version of Assumption 2 leads to a standard panel data linear model, Assumption 2, which applies to a distribution rather than linear projection, seems appropriate for a nonlinear model.

Although they seem appropriate for a nonlinear model, the time homogeneity conditions are strong. In particular they do not allow for heteroskedasticity over time, which is often thought to be imporant in applications. We partially address this problem in Appendix B by allowing for location and scale time effects.

A dynamic version of the model can be obtained by only including current and lagged X_{is} in the conditioning set for each t, as in the following condition:

ASSUMPTION 3: $\varepsilon_{it}|X_{it}, ..., X_{i1}, \alpha_i \overset{d}{=} \varepsilon_{i1}|X_{i1}, \alpha_i$, *for all* t.

This is a "predetermined" version of time homogeneity, where only the conditional distribution given current and lagged regressors must be time invariant. It does imply that the conditional distribution of ε_{it} given current and lagged regressors only depends on X_{i1}, and so embodies a conditional independence restriction. Here conditioning on α_i has an important role, acting as a kind of "control variable" by making ε_{it} be independent of all X_{is} for $1 < s \leq t$. This model is dynamic in the sense that ε_{it} and X_{is} can be dependent for $s > t$. For instance, X_{it} could be $Y_{i,t-1}$, in which case $Y_{it} = g_0(Y_{it-1}, \alpha_i, \varepsilon_{it})$ is an explicit nonseparable dynamic model with ε_{it} being time shocks that are independent of $Y_{it-1}, ..., Y_{i1}$. An important example is one where where $Y_{it} \in \{0, 1\}$ is binary, representing state dependence, with α_i representing unobserved heterogeneity.

We will focus in the nonparametric model on two effects of x on y, the average structural function (ASF) of Blundell and Powell (2003) and the quantile structural function (QSF) of Imbens and Newey (2009). The ASF is

$$\mu(x) = E[g_0(x, \alpha_i, \varepsilon_{it})] = \int g_0(x, \alpha, \varepsilon) F(d\alpha, d\varepsilon).$$

This object is useful for quantifying the effect of x on the mean of the outcome Y_{it}. In the treatment effects literature the average treatment effect (ATE) of changing x from $\bar{x}$ to $\tilde{x}$ is

$$\mu(\tilde{x}) - \mu(\bar{x}).$$

The QSF $q(\lambda, x)$ is the λ^{th} quantile of $g_0(x, \alpha_i, \varepsilon_{it})$. Under conditions specified below the QSF will equal the inverse of the cumulative distribution function (CDF),

$$q(\lambda, x) = G^{-1}(\lambda, x), G(y, x) = E[1(g_0(x, \alpha_i, \varepsilon_{it}) \leq y)].$$

In the treatment effects literature the λ^{th} quantile treatment effect of changing x from $\bar{x}$ to $\tilde{x}$ is

$$q(\lambda, \tilde{x}) - q(\lambda, \bar{x}),$$

as in Lehmann (1974).

The ATE is the average effect of x on y integrating over both α_i and ε_{it}. Chamberlain (1982), Hahn (2001), Wooldridge (2005), and Chernozhukov et. al (2007) have also considered a conditional mean model $E[Y_{it}|X_i, \alpha_i] = m_0(X_{it}, \alpha_i)$ and the average partial effect $\int [m_0(\tilde{x}, \alpha) - m_0(\bar{x}, \alpha)]F(d\alpha)$. It turns out that the nonseparable models given here imply conditional mean models where the average partial effect is the ATE, as shown in the following result.

THEOREM 1: *Suppose that Assumption 1 is satisfied,* $E[|Y_{it}|] < \infty$, *and* $E[|g_0(x, \alpha_i, \varepsilon_{it})|] < \infty$ *for all* x. *If Assumption 2 is satisfied then for* $\tilde{\alpha} = X$ *and* $m_0(x, \tilde{\alpha}) = \int g_0(x, \alpha, \varepsilon)F(d\alpha, d\varepsilon|\tilde{\alpha})$,

$$E[Y_{it}|X_i, \tilde{\alpha}_i] = m_0(X_{it}, \tilde{\alpha}_i), \mu(x) = \int m_0(x, \tilde{\alpha})F(d\tilde{\alpha}).$$

If Assumption 3 is satisfied then for $\tilde{\alpha} = (\alpha, X_1)$ *and* $m_0(x, \tilde{\alpha}) = \int g_0(x, \alpha, \varepsilon)F(d\varepsilon|\tilde{\alpha})$,

$$E[Y_{it}|X_{it}, ..., X_{i1}, \tilde{\alpha}_i] = m_0(X_{it}, \tilde{\alpha}_i), \mu(x) = \int m_0(x, \tilde{\alpha})F(d\tilde{\alpha}).$$

Proofs of all of the results are given in an Appendix. From the expression for $\mu(x)$ given in this result it follows that the ATE equals the average partial effect. Consequently, bounds for the average partial effect, such as those in Chernozhukov et. al. (2007), will imply bounds for the ATE. Indeed the ATE bounds given here are identical to the average partial effect bounds from Chernozhukov et. al. (2007).

To help explain these and other results, it is useful to consider examples. Binary choice is a very important model for panel data, having many applications, and so we choose that as our main example. The most common model has been one with a scalar individual effect that is an additive shift to a linear combination of X_{it}, where

$$Y_{it} = 1(X_{it}'\beta_0 + \alpha_i \geq \varepsilon_{it}),$$

for scalar ε_{it}. In this example $g_0(x, \alpha, \varepsilon) = 1(x'\beta_0 + \alpha \geq \varepsilon)$ and the ASF is

$$\mu(x) = \int 1(x'\beta_0 + \alpha \geq \varepsilon)F(d\varepsilon, d\alpha).$$

This is an unusual object, but Theorem 1 helps relate it to the more familiar average partial effect. Consider a special case of Theorem 1 where ε_{it} is independent of (X_i, α_i) with CDF $H(\varepsilon)$ for each t. Then

$$\mu(x) = \int 1(x'\beta_0 + \alpha \geq \varepsilon)F(d\varepsilon)F(d\alpha) = \int H(x'\beta_0 + \alpha)F(d\alpha).$$

Here $H(x'\beta_0 + \alpha)$ is the choice probability given x and α, so the ATE will be the partial effect $H(\tilde{x}'\beta_0 + \alpha) - H(\bar{x}'\beta_0 + \alpha)$ averaged over α.

In this paper we will focus on discrete regressors. We formalize that focus by imposing the following condition from here on:

ASSUMPTION 4: *The support of X_i is finite, and is equal to $\{X^1, ..., X^K\}$.*

With discrete X_{it} the model can also be written as a multiple regression with random coefficients, though we find it convenient to use the notation given here. One interesting example of a discrete regressor is a binary X_{it}, where $X_{it} \in \{0, 1\}$. In this example X_i will be a vector of zeros and ones.

3 Estimation of Identified Effects

To explain identification in the static model, i.e. under Assumption 2, it is helpful to consider the conditional ASF given by

$$\mu(x|X_i) = E[g_0(x, \alpha_i, \varepsilon_{i1})|X_i].$$

It will turn out that this conditional ASF will be identified for any X_i where there is a t with $X_{it} = x$. It will also be the case that if there is no time period with $X_{it} = x$ then $\mu(x|X_i)$ will not be identified, leading to $\mu(x)$ not being identified.

To describe these results, define $d_{it}(x) = 1(X_{it} = x)$. Note that

$$d_{it}(x)Y_{it} = d_{it}(x)g_0(X_{it}, \alpha_i, \varepsilon_{it}) = d_{it}(x)g_0(x, \alpha_i, \varepsilon_{it}).$$

Then since $d_{it}(x)$ is a function of X_i,

$$\begin{aligned} d_{it}(x)E[Y_{it}|X_i] &= E[d_{it}(x)g_0(x, \alpha_i, \varepsilon_{it})|X_i] \\ &= d_{it}(x)E[g_0(x, \alpha_i, \varepsilon_{it})|X_i] = d_{it}(x)\mu(x|X_i), \end{aligned} \tag{1}$$

where the last equality follows by Assumption 2. Therefore, if for some $t(x)$ we have $X_{it(x)} = x$ then substituting $d_{it(x)}(x) = 1$ in the previous equation we see that

$$\mu(x|X_i) = E[Y_{it(x)}|X_i] \tag{2}$$

This shows that $\mu(x|X_i)$ is identified for any X_i where $X_{it} = x$ for some t. Time homogeneity is essential to this result. The fact that "time is randomly assigned" implies that $E[g_0(x, \alpha_i, \varepsilon_{it})|X_i]$ is invariant to t so that we can choose whichever t has $X_{it} = x$ without affecting the result.

Identification of the conditional ASF leads to identification of a corresponding conditional ATE, defined to be $\Delta(X_i) = \mu(\tilde{x}|X_i) - \mu(\bar{x}|X_i)$. By identification of the conditional ASF $\mu(x|X_i)$ whenever $X_{it} = x$ for some t, the conditional ATE is identified whenever $X_{i\tilde{t}} = \tilde{x}$ for some $\tilde{t}$ and $X_{i\bar{t}} = \bar{x}$ for some $\bar{t}$, with

$$\Delta(X_i) = E[Y_{i\tilde{t}}|X_i] - E[Y_{i\bar{t}}|X_i] = E[Y_{i\tilde{t}} - Y_{i\bar{t}}|X_i].$$

This equation is a precise formulation of the idea that one can use variation of x over time to identify the ceteris paribus effect of x on y. The conditional ATE is obtained by varying t over time periods where $\tilde{x}$ and $\bar{x}$ occur.

This identification result implies identification of the ATE δ conditional on X_i including both $\tilde{x}$ and $\bar{x}$. Define $T_i(x) = \sum_{t=1}^{T} d_{it}(x)$ and $D_i = 1(T_i(\tilde{x}) > 0)1(T_i(\bar{x}) > 0)$, so that $D_i = 1$ if and only if X_i includes both $\tilde{x}$ and $\bar{x}$ for some time periods. For all X_i with $D_i = 1$ define $\tilde{t}_i$ and $\bar{t}_i$ such that $X_{i\tilde{t}_i} = \tilde{x}$ and $X_{i\bar{t}_i} = \bar{x}$.

$$\begin{aligned} \delta &= E[g_0(\tilde{x}; \alpha_i, \varepsilon_{i1}) - g_0(\bar{x}, \alpha_i, \varepsilon_{i1})|D_i = 1] = E[\Delta(X_i)D_i]/E[D_i] \\ &= E[E[Y_{i\tilde{t}_i} - Y_{i\bar{t}_i}|X_i]D_i]/E[D_i] = E[D_i(Y_{i\tilde{t}_i} - Y_{i\bar{t}_i})]/E[D_i]. \end{aligned}$$

This δ is the ATE for those individuals who have both $\tilde{x}$ and $\bar{x}$ among their regressor values.

The identified effect δ may be of interest in many settings. For example, when Y_{it} is log earnings and $X_{it} \in \{0, 1\}$ represents union status, δ would be the effect of union status on earnings for those who changed union status over the time periods we observe. For a given number of time periods T, this is all one could hope to identify nonparametrically. However, we may be interested in other effects too. We might be interested in union effects for those who ever change union status, which we could identify as T gets large. Or we might even be interested in the effect for those who were ever in a union. The data will also be informative about these effects and in the next Section we will provide bounds for them.

The conditional ASF $\mu(x|X_i)$ is not identified if X_{it} does not take on the value x for any time period. Intuitively, $g_0(x, \alpha_i, \varepsilon_{it})$ is never observed in that case so that the data does not provide any information about $\mu(x|X_i)$. Furthermore, as long as the support of X_{it} is the same for each t and the support of X_i is the Cartesian product of the supports of X_{it}, there will be some X_i with positive probability where $X_{it} \neq x$ for all t. Hence $\mu(x|X_i)$ will not be identified for some X_i, and so $\mu(x) = E[\mu(x|X_i)]$ is not identified either. The data still may be informative about $\mu(x)$ if g_0 is bounded, as discussed in the next Section, but in general $\mu(x)$ is not identified, and hence neither is the ATE.

Consider the binary $X_{it} \in \{0,1\}$ example with and $T = 2$. Let $\mathcal{P}^k = \Pr(X_i = X^k)$. Assume the support of X_i is $\{X^1, ..., X^4\}$ with $X^1 = (0,0)'$, $X^2 = (0,1)'$, $X^3 = (1,0)'$, $X^4 = (1,1)'$. Then $\mu(1|X_i = X^1)$ is not identified, because $X_{it} = 0$ for each t when $X_i = X^1$. Similarly, $\mu(0|X_i = X^4)$ is also not identified. Hence none of $\mu(1)$, $\mu(0)$, and the ATE $\mu(1) - \mu(0)$ are identified. In this example the identified conditional ATE δ is the ATE conditional on $X_i \in \{X^2, X^3\}$. Here, where X_{it} is binary, this is the ATE conditional on X_{it} changing over time. This result also follows from Theorem 1 and Chamberlain (1982), who showed nonidentification of the average partial effect in a linear random coefficient model with a discrete regressor.

These identification insights can be used to analyze the properties of linear fixed effects (FE), that has sometimes been used to try to estimate the ATE from panel data. Firstly, since there is no consistent estimator for an unidentified parameter, FE will not be consistent for the ATE. It also turns out that FE is not consistent for the identified conditional ATE δ either. Instead, FE converges to a weighted average of $\Delta(X_i)$.

To simplify the exposition we derive the limit of FE for binary $X_{it} \in \{0,1\}$. FE is $\hat{\delta}_w$ from least squares on $Y_{it} = X_{it}\delta + \gamma_i + v_{it}$,where each γ_i is estimated. For $\bar{X}_i = \sum_{t=1}^T X_{it}/T$,

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

Let $r_i = \#\{t : X_{it} = 1\}/T$ and $\sigma_i^2 = r_i^k(1 - r_i^k)$.

THEOREM 2: *If Assumptions 1, 2, and 4 are satisfied, (X_i, Y_i) has finite second moments, and $E[\sigma_i^2] > 0$, then*

$$\hat{\delta}_w \xrightarrow{p} \delta_w = \frac{E[\sigma_i^2 \Delta(X_i)]}{E[\sigma_i^2]}. \tag{3}$$

The value of $\Delta(X_i)$ has no impact on δ_w where $\sigma_i^2 = 0$, i.e. when $X_i = (0, ..., 0)'$ or $X_i = (1, ..., 1)'$, consistent with these $\Delta(X_i)$ being unidentified. Also, δ_w is a weighted average of conditional ATE's, where the weights σ_i^2 are the variances across time of the X_i vectors. If $T \geq 4$ then these weights vary over the positive σ_i^2 and so the limit δ_w of FE is not the identified conditional ATE δ.

Theorem 2 is different than Yitzhaki (1996) and Angrist (1998), who gave weighted average interpretations of least squares in other, non panel settings. Theorem 2 is also different from Hahn (2001), who found that $\hat{\delta}_w$ consistently estimates the marginal effect. Hahn (2001) considered $T = 2$ and excluded $(0,0)'$ and $(1,1)'$ from the support of X_i, so neither feature that causes inconsistency of $\hat{\delta}_w$ is present. As noted by Hahn (2001), those conditions are quite special. Theorem 2 is also different from Wooldridge (2005), who showed that if $b_i = E[g_0(1, \alpha_i, \varepsilon_{it}) - g_0(0, \alpha_i, \varepsilon_{it})|\alpha_i]$ 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. For instance, if $T = 2$, $X_{i2} - \bar{X}_i$ 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 $\Delta(X^2) = \Delta(X^3) = [\Pr(X^1)\Delta(X^1) + \Pr(X^4)\Delta(X^4)]/[\Pr(X^1) + \Pr(X^4))$. 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.

A simple estimator of the identified conditional ATE δ is

$$\hat{\delta} = \sum_{i=1}^{n} D_i \left(\frac{\sum_{t=1}^{T} d_{it}(\tilde{x}) Y_{it}}{T_i(\tilde{x})} - \frac{\sum_{t=1}^{T} d_{it}(\bar{x}) Y_{it}}{T_i(\bar{x})} \right) / \sum_{i=1}^{n} D_i. \tag{4}$$

Here we average each Y_{it} over all time periods with $d_{it}(x) = 1$ rather than just using one time period. This is a simple approach to using information from more than one time period, but may not be efficient. We leave to future work the treatment of efficiency of estimators of δ. This extends Chamberlain's (1982) estimator to multivariate regressions with discrete X_{it} that are not binary. The following result shows that $\hat{\delta}$ is a consistent estimator of δ.

THEOREM 3: *If Assumptions 1, 2 and 4 are satisfied, (X_i, Y_i) has finite second moments, and $E[D_i] > 0$ then $\hat{\delta} \xrightarrow{p} \delta$.*

This estimator may be of interest in practice where it is important to allow for time effects. For this reason we give a brief discussion of time effects here, reserving to the Appendix a detailed analysis of time effects for bounds. We consider time effects that constitute a location and scale shift of g_0, with $Y_{it} = \tau_t + s_t g_0(X_{it}, \alpha_i, \varepsilon_{it})$ and the normalization $\tau_1 = 0$, $s_1 = 1$. These time effects can be estimated by doing instrumental variables with residual $Y_{it} - \tau_t - s_t Y_{i1}$ on all observations with $X_{it} = X_{i1}$ and a constant and dummy variable for a group of X_{i1} values as instruments. The identified effect $\hat{\delta}_1$ for $t = 1$ can be estimated by replacing Y_{it} by $(Y_{it} - \hat{\tau}_t)/\hat{s}_t$ in equation (4). An effect $\hat{\delta}_t$ for each $t > 1$ and an overall effect $\tilde{\delta}$ can be estimated by

$$\hat{\delta}_t = \hat{\tau}_t + \hat{s}_t \hat{\delta}_1, \tilde{\delta} = \frac{1}{T} \sum_{t=1}^{T} \hat{\delta}_t.$$

Because this is a two step estimator it may be easiest to obtain asymptotic confidence intervals using the bootstrap, resampling from the empirical distribution of (Y_i, X_i).

To get some information about how δ and δ_w can differ from the ATE we give a numerical example. Numerical calculations are reported in Table 1 for a probit model where

$$Y_{it} = 1(X_{it} + \alpha_i > \varepsilon_{it}), \varepsilon_{it} \sim N(0,1), X_{it} = 1(\xi + \alpha_i > \eta_{it}), \eta_{it} \sim N(0,1).$$

We consider three different distributions for α_i,

$$\alpha_i \sim U(-1,1), \alpha_i \text{ symmetric, triangular with support } [-2,2], \alpha_i \sim N(0,1).$$

The value of ξ is chosen to calibrate $P(X_{it} = 1)$ and calculations are done by simulating $500,000$ individuals. We report values of $(\delta_w - \Delta)/\Delta$ and $(\delta - \Delta)/\Delta$, where $\Delta = \mu(1) - \mu(0)$. We find that the biases (inconsistencies) can be substantial in percentage terms. We also find that δ_w has the largest percentage inconsistency when there is little variation in the regressor. Also the bias of δ_w is similar for all T, which is surprising given how weights change with T. In contrast, δ gets close to the ATE as T grows, consistent with the ATE being identified as $T \longrightarrow \infty$, as discussed below.

In a similar way we can also estimate identified quantile effects. Let

$$\hat{G}(y, x|D_i = 1) = \sum_{i=1}^{n} D_i \left[\frac{\sum_{t=1}^{T} d_{it}(x)\Phi(\frac{y - Y_{it}}{h})}{T_i(x)} \right] / \sum_{i=1}^{n} D_i,$$

where Φ is a strictly monotonic CDF and h is a bandwidth. This is an estimator of the distribution function of $g(x, \alpha_i, \varepsilon_{it})$ conditioned both $\tilde{x}$ and $\bar{x}$ appearing in X_i. The indicator function $1(Y_{it} < y)$ has been smoothed, as suggested by Yu and Jones (1998) for estimating a conditional CDF. An estimator of the λ^{th} quantile treatment effect from changing x from $\bar{x}$ to $\tilde{x}$ over all individuals for which both occur is then

$$\hat{G}^{-1}(\lambda, \tilde{x}|D_i = 1) - \hat{G}^{-1}(\lambda, \bar{x}|D_i = 1).$$

4 Bounds in the Static Model

When $g_0(x, \alpha, \varepsilon)$ is bounded the data will be informative about the ATE. In some applications such bounds are implied by the data structure. For example, in the binary choice model, where $Y_{it} \in \{0, 1\}$, lower and upper bounds are 0 and 1 respectively. To describe the bounds, let $d_{it}(x) = 1(X_{it} = x)$ and $T_i(x) = \sum_{t=1}^{T} d_{it}(x)$ as in Section 3. Let $\bar{\mathcal{P}}(x) = \Pr(T_i(x) = 0)$ be the probability that x does not appear in any time period for X_i.

THEOREM 4: *If Assumptions 1, 2, and 4 are satisfied and $B_\ell \leq g_0(x, \alpha_i, \varepsilon_{it}) \leq B_u$ for constants B_ℓ and B_u and all x, then $\mu_\ell(x) \leq \mu(x) \leq \mu_u(x)$ for*

$$\mu_\ell(x) = E[1(T_i(x) > 0)\frac{\sum_{t=1}^{T} d_{it}(x)Y_{it}}{T_i(x)}] + \bar{\mathcal{P}}(x)B_\ell, \mu_u(x) = \mu_\ell(x) + \bar{\mathcal{P}}(x)(B_u - B_\ell),$$

and these bounds are sharp.

Corresponding ATE bounds are

$$\mu_\ell(\tilde{x}) - \mu_u(\bar{x}) \leq \mu(\tilde{x}) - \mu(\bar{x}) \leq \mu_u(\tilde{x}) - \mu_\ell(\bar{x}).$$

The width of these ATE bounds is $[\bar{\mathcal{P}}(\tilde{x}) + \bar{\mathcal{P}}(\bar{x})](B_u - B_\ell)$. Tighter bounds may be obtained by imposing restrictions, such as monotonicity of treatment effects, as shown in an Appendix.

An example may help clarify these bounds. Consider the binary $X_{it} \in \{0,1\}$ example of Section 3 where $T = 2$ and the support of X_i is $\{X^1, ..., X^4\}$ with $X^1 = (0,0)'$, $X^2 = (0,1)'$, $X^3 = (1,0)'$, $X^4 = (1,1)'$. Then

$$\begin{aligned}\mu_\ell(1) &= \mathcal{P}^2 E[Y_{i2}|X^2] + \mathcal{P}^3 E[Y_{i1}|X^3] + \mathcal{P}^4 E[\frac{Y_{i1}+Y_{i2}}{2}|X^4] + \mathcal{P}^1 B_\ell, \\ \mu_u(1) &= \mu_\ell(1) + \mathcal{P}^1(B_u - B_\ell),\end{aligned}$$

Then the width of the bounds is $\mathcal{P}^1(B_u - B_\ell)$. For general T and binary X_{it}, the width of the bounds for $\mu(1)$ is $\Pr(X_i = (0,...,0)')(B_u - B_\ell)$.

The sharpness conclusion of Theorem 4 depends on being able to let $g_0(x, \alpha_i, \varepsilon_{it})$ take any value between B_ℓ and B_u. That is not possible for binary choice, where the outcome is restricted to zero or one. Nevertheless the bounds can still be shown to be sharp.

The QSF bounds are obtained by replacing Y_{it} by $1(Y_{it} \leq y)$ in the ASF bounds and inverting as a function of y. As in equation (1),

$$d_{it}(x) E[1(Y_{it} \leq y)|X_i] = d_{it}(x) E[1(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y)|X_i].$$

For any monotonic function $0 \leq G(y) \leq 1$ that is everywhere increasing and strictly increasing on the interior of its range and any $0 < \lambda < 1$ let

$$Q(\lambda, G(\cdot)) = \begin{cases} -\infty, \ \lambda \leq \inf_y G(y) \\ G^{-1}(\lambda), \ \inf_y G(y) < \lambda < \sup_y G(y) \\ +\infty, \ \lambda \geq \sup_y G(y) \end{cases}$$

THEOREM 5: *If Assumptions 1, 2, and 4 are satisfied and*

$$G_\ell(y, x) = E[1(T_i(x) > 0) T_i(x)^{-1} \sum_{t=1}^{T} d_{it}(x) 1(Y_{it} \leq y)]$$

is continuous and strictly increasing in y on the interior of its range then $q_\ell(\lambda, x) \leq q(\lambda, x) \leq q_u(\lambda, x)$ for

$$q_\ell(\lambda, x) = Q(\lambda, G_\ell(\cdot, x) + \bar{\mathcal{P}}(x)), q_u(\lambda, x) = Q(\lambda, G_\ell(\cdot, x)).$$

If $G_\ell(y, x)$ is continuous and strictly increasing in y then these bounds are sharp.

Bounds for quantile treatment effects can then be formed in the usual way as

$$q_\ell(\lambda, \tilde{x}) - q_u(\lambda, \bar{x}) \leq q(\lambda, \tilde{x}) - q(\lambda, \bar{x}) \leq q_u(\lambda, \tilde{x}) - q_\ell(\lambda, \bar{x}).$$

The width of these bounds depends on the shape of distribution of Y. They are infinitely wide for the λ^{th} quantile when $\lambda \leq \max\{\bar{\mathcal{P}}(\tilde{x}), \bar{\mathcal{P}}(\bar{x})\}$ or $\lambda \geq \min\{1 - \bar{\mathcal{P}}(\tilde{x}), 1 - \bar{\mathcal{P}}(\bar{x})\}$, and will be tighter when $\bar{\mathcal{P}}(\tilde{x})$ and $\bar{\mathcal{P}}(\bar{x})$ are small.

Estimation of the ATE bounds is straightforward. Let $\bar{P}(x) = n^{-1} \sum 1(T_i(x) = 0)$. Estimates of the ASF bounds are given by

$$\hat{\mu}_\ell(x) = n^{-1} \sum_{i=1}^{n} 1(T_i(x) > 0) \frac{\sum_{t=1}^{T} d_{it}(x) Y_{it}}{T_i(x)} + \bar{P}(x) B_\ell, \hat{\mu}_u(x) = \hat{\mu}_\ell(x) + \bar{P}(x)(B_u - B_\ell).$$

Note that no estimate of the support of X_i is required for estimation of the bounds. These estimates will be consistent and jointly asymptotically normal under our i.i.d. sampling framework for (Y_i, X_i). Also, they are averages over i of explicit, simple functions of the data and so it is straightforward to estimate the joint asymptotic variance of the upper and lower bounds for the ASF and for the ATE. Confidence intervals for the identified set can then be formed using results of Chernozhukov, Hong, and Tamer (2007) or Beresteanu and Molinari (2008) on bounds estimators that are jointly asymptotically normal.

Estimation of the quantile bounds proceeds similar, replacing Y_{it} with a smoother version of the indicator $1(Y_{it} \leq y)$ and then inverting as a function of y to get quantile bounds. Let $\Phi(u)$ be some CDF, h be a bandwidth. Then an estimator of the lower bounds for $G(y, x)$ is

$$\hat{G}_\ell(y, x) = n^{-1} \sum_{i=1}^{n} 1(T_i(x) > 0) \frac{\sum_{t=1}^{T} d_{it}(x) \Phi(\frac{y - Y_{it}}{h})}{T_i(x)}, \hat{G}_u(y, x) = \hat{G}_\ell(y, x) + \bar{P}(x).$$

By construction $\hat{G}_\ell(y, x)$ is strictly monotonic increasing as long as $T_i(x) > 0$ for some i and so can be inverted to give

$$\hat{q}_\ell(\lambda, x) = Q(\lambda, \hat{G}_\ell(\cdot, x) + \bar{P}(x)), \hat{q}_u(\lambda, x) = Q(\lambda, \hat{G}_\ell(\cdot, x)).$$

These estimators will be joint asymptotically normal for different values of x, but their asymptotic variance is complicated because of the function inversion, so it may be easiest to estimate asymptotic variance by the bootstrap. Inference about identified intervals can then be carried out as for the ATE estimators.

Similarly to the treatment effects literature, we may be interested in the average structural function, or the average treatment effect, conditional on certain X_i values. For example, if $X_{it} \in \{0, 1\}$ represents treatment then we might be interested on the effect of treatment conditional on ever treated, i.e. conditional on $X_i \neq (0, ..., 0)'$. Tighter bounds for such effect can be formed and in some cases the effects may be identified. These effects can be estimated by including the indicator function for the event of interest inside each sum over i and dividing the sum by the sample probability of that event.

5 Bounds in the Dynamic Model

The bounds for the static model are based on decomposing the average treatment effect into components conditional on the entire vector X_i. This can be thought of as a partition based

on X_i. In the dynamic model we can no longer use a partition based on the entire vector X_i because of endogeneity of the lead variables. For the dynamic model we instead use a partition that conditions only on lagged X_{it}. Specifically, we partition the support of X_i into sets where the first occurrence of x is at time t and the set where x never occurs. This partition is given by

$$\mathcal{X}_t(x) = \{X : X_t = x,\ X_s \neq x\ \forall s < t\}, t = 1, ..., T; \bar{\mathcal{X}}(x) = \{X : X_t \neq x\ \forall t\}.$$

There is a fundamental result that provides partial identification using this partition. Define $d_{it}^{\mathcal{X}}(x) = 1(X_i \in \mathcal{X}_t(x))$ and note that $d_{it}^{\mathcal{X}}(x)$ only depends on $X_{it}, ..., X_{i1}$. For all t,

$$\begin{aligned} E[d_{it}^{\mathcal{X}}(x)Y_{it}] &= E[d_{it}^{\mathcal{X}}(x)g_0(x, \alpha_i, \varepsilon_{it})] = E[d_{it}^{\mathcal{X}}(x)E[g_0(x, \alpha_i, \varepsilon_{it})|X_{it}, ..., X_{i1}, \alpha_i]] \\ &= E[d_{it}^{\mathcal{X}}(x)E[g_0(x, \alpha_i, \varepsilon_{iT})|X_{iT}, ..., X_{i1}, \alpha_i]] = E[d_{it}^{\mathcal{X}}(x)g_0(x, \alpha_i, \varepsilon_{iT})], \end{aligned} \tag{5}$$

where the first equality follows by $X_{it} = x$ when $d_{it}^{\mathcal{X}}(x) = 1$ and the third equality by Assumption 3. Combining this result with the fact that the distribution of $(\alpha_i, \varepsilon_{it})$ does not vary with t (also implied by Assumption 3) leads to the following bounds:

THEOREM 6: *Suppose that Assumptions 1, 3, and 4 are satisfied. If $B_\ell \leq g_0(x, \alpha_i, \varepsilon_{it}) \leq B_u$ for constants B_ℓ and B_u and all x, then $\mu_\ell(x) \leq \mu(x) \leq \mu_u(x)$ for*

$$\mu_\ell(x) = E[\sum_{t=1}^{T} d_{it}^{\mathcal{X}}(x)Y_{it}] + \bar{\mathcal{P}}(x)B_\ell, \mu_u(x) = \mu_\ell(x) + \bar{\mathcal{P}}(x)(B_u - B_\ell).$$

If $\tilde{G}_\ell(y, x) = E[\sum_{t=1}^{T} d_{it}^{\mathcal{X}}(x)1(Y_{it} \leq y)]$ is continuous and strictly increasing in y on the interior of its range then $q_\ell(\lambda, x) \leq q(\lambda, x) \leq q_u(\lambda, x)$ for

$$q_\ell(\lambda, x) = Q(\lambda, \tilde{G}_\ell(\cdot, x) + \bar{\mathcal{P}}(x)), q_u(\lambda, x) = Q(\lambda, \tilde{G}_\ell(\cdot, x)).$$

An important example is the binary $Y_{it} \in \{0, 1\}$ case where $X_{it} = Y_{i,t-1}$. In this case $B_\ell = 0$, $B_u = 1$. Here $\bar{\mathcal{P}}(0) = \Pr(X_i = (1, ..., 1)')$ and $\bar{\mathcal{P}}(1) = \Pr(X_i = (0, ..., 0)')$. The bounds for $\mu(0)$ and $\mu(1)$ will be

$$\begin{aligned} E[\sum_{t=1}^{T} d_{it}^{\mathcal{X}}(0)Y_{it}] &= \mu_\ell(0) \leq \mu(0) \leq \mu_u(0) = \mu_\ell(0) + \bar{\mathcal{P}}(0), \\ E[\sum_{t=1}^{T} d_{it}^{\mathcal{X}}(1)Y_{it}] &= \mu_\ell(1) \leq \mu(1) \leq \mu_u(1) = \mu_\ell(1) + \bar{\mathcal{P}}(1). \end{aligned}$$

Then for $\delta^{\mathcal{X}} = \sum_{t=1}^{T} E[\{d_{it}^{\mathcal{X}}(1) - d_{it}^{\mathcal{X}}(0)\}Y_{it}]$ we have

$$\delta^{\mathcal{X}} - \bar{\mathcal{P}}(0) \leq \mu(1) - \mu(0) \leq \delta^{\mathcal{X}} + \bar{\mathcal{P}}(1).$$

The width of the bounds is $\bar{\mathcal{P}}(0) + \bar{\mathcal{P}}(1)$, that will tend to be large in short panels but more informative in long ones. This is a bounds solution to the problem of identifying state dependence in the presence of unobserved heterogeneity (Feller, 1943, and Heckman, 1981), since

$$\mu(1) - \mu(0) = \int [\Pr(Y_{it} = 1|Y_{i,t-1} = 1, \alpha) - \Pr(Y_{it} = 1|Y_{i,t-1} = 0, \alpha)] F(d\alpha)$$

is the effect of lagged Y_{it}, holding α_i fixed, averaged over α_i. Note that the joint distribution of $\Pr(Y_{it} = 1|Y_{i,t-1} = 1, \alpha)$ and $\Pr(Y_{it} = 1|Y_{i,t-1} = 0, \alpha)$ is entirely unrestricted, allowing general effects of heterogeneity on dynamics. This set up is like Browning and Carro (2007), though we give bounds and they suggest estimators.

In the dynamic bounds $d_{it}^{\mathcal{X}}(x) = 1$ for at most one time period. This means that although the bounds for the dynamic model also apply to the static model there may be advantages to using the static bounds when they apply. One advantage is that the bounds for the static model use more than one time period, which should help reduce sampling variability in estimators.

Figure 1 shows the width of the bounds in a numerical example based on a dynamic probit model where

$$Y_{it} = 1(\beta_0 Y_{i,t-1} + \alpha_i \geq \varepsilon_{it}), \varepsilon_{it} \sim N(0,1), \alpha_i \sim N(0,1), \Pr(Y_{i0} = 1) = .5.$$

We consider different DGPs indexed by $\beta_0 \in [-2, 2]$ and compute the width of the bounds for $T \in \{2, 4, 8, 16, 32, 64\}$. The width is asymmetric with respect to $\beta_0 = 0$ because $\Pr(X_i = (1, ..., 1)')$ grows with β_0, whereas $\Pr(X_i = (0, ..., 0)')$ does not depend on β_0. We find that the bounds can be substantially wide for high values of β_0 even for large T. It may be possible to tighten these bounds using a semiparametric model as we do for bounds in the static case in Section 7.

Estimators for the bounds can be constructed analogously to the static model. The bounds for the ASF can be estimated by

$$\hat{\mu}_\ell(x) = n^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} d_{it}^{\mathcal{X}}(x) Y_{it} + \bar{P}(x) B_\ell, \hat{\mu}_u(x) = \hat{\mu}_\ell(x) + \bar{P}(x)(B_u - B_\ell).$$

The bounds for the QSF can be estimated analogously. For brevity we omit details.

6 Identification and Rates as $T \longrightarrow \infty$

It is interesting to consider whether the average and quantile treatment effect become identified as T grows. Of course, this will only be possible if $\tilde{x}$ and $\bar{x}$ both eventually occur for every individual, or more generally each x in the support of X_{it} eventually shows up in X_i. Mathematically, this corresponds to $\bar{\mathcal{P}}(x) \longrightarrow 0$ as $T \longrightarrow \infty$. The following result gives sufficient conditions for this to occur.

THEOREM 7: *Suppose that Assumptions 1, 3, and 4 are satisfied,* $\overrightarrow{X}_i = (X_{i1}, X_{i2}, ...)$ *is stationary, the support of each* X_{it} *conditional on* α_i *is the marginal support of* X_{it}*, and* $\overrightarrow{X}_i$ *is ergodic conditional on* α_i*. If* $E[|g_0(x, \alpha_i, \varepsilon_{i1})|] < \infty$ *for each* x *in the support of* X_{it} *then* $\delta \longrightarrow \mu(\tilde{x}) - \mu(\bar{x})$ as $T \longrightarrow \infty$*. If* $B_\ell \leq g_0(x, \alpha_i, \varepsilon_{it}) \leq B_u$ *for constants* B_ℓ *and* B_u *and all* x*, then* $\mu_\ell(x) \longrightarrow \mu(x)$ *and* $\mu_u(x) \longrightarrow \mu(x)$ as $T \longrightarrow \infty$*. If* $0 < \lambda < 1$ *and* $G(y, x)$ *is continuous and strictly monotonic in* y *on* $\{y : 0 < G(y, x) < 1\}$ *then* $q_\ell(\lambda, x) \longrightarrow q(\lambda, x)$ *and* $q_u(\lambda, x) \longrightarrow q(\lambda, x)$ *as* $T \longrightarrow \infty$*.*

Clearly identification as $T \longrightarrow \infty$ can only occur if every individual can have every X_{it}. This is the meaning of the condition that the support of X_{it} conditional α_i is the marginal support of X_{it}. If this is not true, then some individuals, as represented by α_i, will never reach some x value, and so we cannot nonparametrically identify any average treatment involving that x value. To explain mathematically, consider a simple example where X_{it} is i.i.d. conditional on α_i. In that case

$$\bar{\mathcal{P}}(x) = E[\Pr(X_{it} \neq x|\alpha_i)^T].$$

This will not go to zero if and only if $\Pr(X_{it} \neq x|\alpha_i) = 1$ with positive probability, that is $\Pr(X_{it} = x|\alpha_i) = 0$ with positive probability. The marginal support being equal to the conditional support rules this out.

In the union example presumably not everyone joins a union at some point so the treatment effect of unions averaged over all people is not identified as $T \longrightarrow \infty$. Even the average treatment effect for those ever in a union may not be identified, because there are individuals who are always in a union. In general the best we can hope for as $T \longrightarrow \infty$ is to identify the average effect for those individuals who will eventually have x equal to $\tilde{x}$ and $\bar{x}$.

The rate at which the width of the bounds shrink is a complicated question. We can give a simple result if the conditional probability for $X_{it} = x$ is bounded away from zero.

THEOREM 8: *Suppose that Assumptions 1, 3, and 4 are satisfied,* $\overrightarrow{X}_i$ *is stationary and Markov of order* J *conditional on* α_i*, and for some* $\varepsilon > 0$*,*

$$\Pr(X_{it} = x|X_{i,t-1}, ..., X_{i,t-J}, \alpha_i) \geq \varepsilon.$$

Then if $B_\ell \leq g_0(x, \alpha_i, \varepsilon_{it}) \leq B_u$*,*

$$\mu_u(x) - \mu_\ell(x) \leq (B_u - B_\ell)(1 - \varepsilon)^{T-J}.$$

Also, if $0 < \lambda < 1$ *and* $G(y, x)$ *is continuously differentiable on a neighborhood of* $y = q(\lambda, x)$ *with a derivative bounded below by* $D_x > 0$*, then for a large enough* T

$$q_u(\lambda, x) - q_\ell(\lambda, x) \leq 2D_x^{-1}(1 - \varepsilon)^{T-J}.$$

This result shows that the rate of convergence of the bounds will be exponential when the conditional probability that $X_{it} = x$ is bounded away from zero. A simple example where X_{it} is i.i.d. conditional on α_i can be used to illustrate what other kinds of results might occur. As discussed above, $\bar{\mathcal{P}}(x) = E[\Pr(X_{it} \neq x|\alpha_i)^T]$, so the rate of shrinkage depends on the thickness of the tails of the distribution of $\Pr(X_{it} \neq x|\alpha_i)$. If too much weight is put on conditional probabilities near one then the convergence may be slow. For example, suppose $X_{it} = 1(\alpha_i \geq \eta_{it})$, $\alpha_i \sim N(0,1)$, $\eta_{it} \sim N(0,1)$. Then

$$\bar{\mathcal{P}}(0) = E[\Phi(\alpha_i)^T] = \int \Phi(\alpha)^T \phi(\alpha) d\alpha = \frac{\Phi(\alpha)^{T+1}}{T+1}\Big|_{-\infty}^{+\infty} = \frac{1}{T+1},$$

which shrinks slower than exponentially. On the other hand, if α_i has any distribution with a compact support, Theorem 8 implies that the bounds shrink exponentially fast in T.

Table 1 provides related examples showing how fast the identified effect δ gets close to the ATE as T grows. The rate at which the percentage error between δ and the ATE shrinks is bounded above by the rate at which the bounds shrink. In the Gaussian case α_i in Table 1 the error in δ shrinks slowly as suggested by the rate example above. Also, in the uniform α_i the error in δ shrinks noticeably more quickly, as Theorem 8 suggests it should.

7 Semiparametric Multinomial Choice Models

The nonparametric bounds are informative but may be quite wide for small T. Using the information implied by a parametric model for the conditional distribution of Y_i given (X_i, α_i) the bounds may be tighter. Such a specification corresponds to a semiparametric model with a nonparametric part that includes the conditional distribution of α_i given X_i. We focus here on multinomial choice models, paying particular attention to the binary choice model where α_i is a scalar location effect.

We take a multinomial panel data model to be one satisfying

ASSUMPTION 5: *$Y_i = (Y_{i1}, ..., Y_{iT})'$ has finite support $\{Y^1, ..., Y^J\}$ and $\Pr(Y_i = Y^j|X_i = X^k) = \int \mathcal{L}_j^k(\alpha, \beta^*) F_k^*(d\alpha)$ where $\mathcal{L}_j^k(\alpha, \beta)$ is a known function and β is a parameter vector with true value β^*.*

This is a semiparametric model, with the parameters β being the parametric part and the conditional distributions $F_k(\alpha), (k = 1, ..., K)$ being a nonparametric part. An important example is a binary choice model where $Y_{it} \in \{0, 1\}$, $\Pr(Y_{it} = 1|X_i, \alpha_i, \beta) = H(X_{it}'\beta + \alpha_i)$ for a CDF H, and $Y_{i1}, ..., Y_{iT}$ are mutually independent conditional on X_i and α_i. In this case each possible realization Y^j of Y_i will be a $T \times 1$ vector of zeros and ones, as in the following condition:

ASSUMPTION 6: *$Y_{it} \in \{0,1\}$ and $\mathcal{L}_j^k(\alpha,\beta) = \prod_{t=1}^T H(X_t^{k\prime}\beta+\alpha)^{Y_t^j}[1-H(X_t^{k\prime}\beta+\alpha)]^{1-Y_t^j}$ where H is a differentiable CDF with derivative h that is positive, bounded, an even function, and monotonically decreasing on $[0,\infty)$.*

One can easily generalize this specification to allow for time effects and to allow for random slopes by changing $\mathcal{L}_j^k(\alpha,\beta)$. Let z_t be a vector of variables that only change with time (e.g. dummy variables for all time periods but one). Consider

$$\mathcal{L}_j^k(\alpha,\beta) = \prod_{t=1}^T H(z_t'\beta_1 + X_{t1}^{k\prime}\beta_2 + X_{t2}^{k\prime}\alpha)^{Y_t^j}[1-H(z_t'\beta_1 + X_{t1}^{k\prime}\beta_2 + X_{t2}^{k\prime}\alpha)]^{1-Y_t^j}.$$

These probabilities specify that the observations over time of Y_{it} are independent conditional on α, allow time effects z_t with coefficients β_1, allow some of the individual regressors to have constant coefficients β_2, and also allow for multidimensional heterogeneity α. Here some of the slope coefficients are included in α and so are allowed to vary over individuals. The general computation and estimation methods described in the following sections apply to this model without further modification. One could even restrict the distribution of the individual effect to not depend on certain regressors by imposing equality restrictions across k on the distribution of α, as further discussed below.

The set up here can also be adapted to the dynamic binary choice setting. Consider a specification where for $k \in \{1,2\}$,

$$\begin{aligned}\mathcal{L}_j^k(\alpha,\beta) &= \prod_{t=2}^T H(\beta_1 Y_{t-1}^j + z_t'\beta_2 + \alpha)^{Y_t^j}[1-H(\beta_1 Y_{t-1}^j + z_t'\beta_2 + \alpha)]^{1-Y_t^j} \\ &\quad \times H(\beta_1(k-1) + z_1'\beta_2 + \alpha)^{Y_1^j}[1-H(\beta_1(k-1) + z_t'\beta_2 + \alpha)]^{1-Y_1^j}.\end{aligned}$$

Here the distribution of α can depend on the initial value Y_{i0} in a completely general way and trend terms z_t are present. This set up is like that of Honore and Tamer (2006). One could apply the estimation methods given below to this model and obtain estimated bounds in dynamic models where the distribution of α is unrestricted.

In addition to bounds for β we also consider bounds for the ATE. In the general multinomial model we will assume that the ATE conditional on $X_i = X^k$ is

$$\Delta^k = \int \Delta(\alpha,\beta^*)F_k^*(d\alpha),$$

where $\Delta(\alpha,\beta)$ denotes a marginal (or average partial) effect conditional on α and from now on we let $\Delta^k = \Delta(X^k)$. For example, in the model of Assumption 6 we could take $\Delta(\alpha,\beta) = H(\tilde{x}'\beta+\alpha) - H(\bar{x}'\beta+\alpha)$.

Neither β nor the ATE need be identified. Instead, there will generally be sets of β and ATE values that are consistent with the distribution of the data. To describe the identified sets let $\mathcal{P} = (\mathcal{P}_1^1, ..., \mathcal{P}_J^1, ..., \mathcal{P}_J^K)'$ denote the vector of population choice probabilities and

$$\mathcal{F}_k(\beta, \mathcal{P}) = \{F_k : \mathcal{P}_j^k = \int \mathcal{L}_j^k(\alpha, \beta) F_k(d\alpha), j = 1, ..., J\},$$

where $\mathcal{F}_k(\beta, \mathcal{P})$ may be empty. The identified set for β is

$$B = \{\beta \text{ s.t. } \mathcal{F}_k(\beta, \mathcal{P}) \neq \varnothing, \forall k = 1, ..., K\}.$$

That is, B is the set where there exist individual effect distributions such that integrals of model probabilities equal population choice probabilities. Sharp upper and lower bounds Δ_u^k and Δ_ℓ^k for Δ^k are given by

$$\Delta_u^k = \sup_{\beta \in B, F_k \in \mathcal{F}_k(\beta, \mathcal{P})} \int \Delta(\alpha, \beta) F_k(d\alpha), \ \Delta_\ell^k = \inf_{\beta \in B, F_k \in \mathcal{F}_k(\beta, \mathcal{P})} \int \Delta(\alpha, \beta) F_k(d\alpha). \tag{6}$$

This characterization of bounds for the ATE extends that of Honore and Tamer (2006) from a finite dimensional F_k, where α is restricted to a known fixed grid, to infinite dimensional F_k where any distribution for α is allowed.

A useful feature of multinomial panel models is that they are finite dimensional, in spite of the presence of distributions. The following lemma shows that one only need consider discrete distributions with J unknown support points in the specification of the likelihood and the bounds for the ATE. Let Υ denote the set of possible values for the individual effect and $\mathbb{B}$ the set of parameters for β.

LEMMA 9: *If Assumptions 4 and 5 are satisfied and $\mathcal{L}_j^k(\alpha, \beta)$ is a measurable function of α for each $\beta \in \mathbb{B}$, then for each β and every CDF F_k on Υ there is a discrete distribution F_k^J with no more than J support points such that $\int \mathcal{L}_j^k(\alpha, \beta) F_k^J(d\alpha) = \int \mathcal{L}_j^k(\alpha, \beta) F_k(d\alpha)$ $(j = 1, ..., J)$. If, in addition, $\Delta(\alpha, \beta)$ is bounded for each β then Δ_u^k and Δ_ℓ^k are not affected by restricting attention to $F_k \in \mathcal{F}_k(\beta)$ that are discrete with no more than J support points.*

Thus, no matter what the dimension of α is, the multinomial panel model is finite dimensional, with number of parameters given by $\dim(\beta) + (2J-1)^K$. Another implication of this result is that the distribution of the individual effect is generally not identified in multinomial models. For example, if the true distribution F_k^* were continuous then Lemma 9 implies that there is a discrete distribution that gives exactly the same likelihood. The proof of this result is similar to Lindsay's (1983) result that the maximum likelihood estimator of a mixture model has a finite support.

The constraints imposed by Assumption 5 lead to the same population formula for the identified components of the ASF as in the nonparametric model. For instance, in the model of

Assumption 6 the ASF conditional on $X_i = X^k$ is $\mu(x|X^k) = \int H(x'\beta^* + \alpha)F_k^*(d\alpha)$. If $X_{\bar{t}}^k = \bar{x}$ for some $\bar{t}$ then

$$\begin{aligned}
\mu(\bar{x}|X^k) &= \int H(\bar{x}'\beta^* + \alpha)F_k^*(d\alpha) = \int \Pr(Y_{i\bar{t}} = 1|X_i = X^k, \alpha)F_k^*(d\alpha) \\
&= \int \sum_{j:Y_{\bar{t}}^j=1} \mathcal{L}_j^k(\alpha, \beta^*) F_k^*(d\alpha) = \sum_{j:Y_{\bar{t}}^j=1} \mathcal{P}_j^k = \Pr(Y_{i\bar{t}} = 1|X_i = X^k) \\
&= E[Y_{i\bar{t}}|X_i = X^k],
\end{aligned}$$

where the fourth equality holds by Assumption 5. This is the same formula for the identified ASF as in the static model; see equation (1).

When for X_i where $X_{it} \neq x$ for all t, the static nonparametric model places no restrictions on $\mu(x|X_i)$ other than being in between the known bounds B_ℓ and B_u. For the binary choice model the only nonparametric restriction is that $0 \leq \mu(x|X_i) \leq 1$ in such a case. The semiparametric bounds are tighter than the nonparametric bounds precisely when Assumption 5 provides information about $\mu(x|X_i)$ for such x. In the binary choice model with additive heterogeneity this occurs because Assumption 6 provides information about the distribution of α conditional on such X_i, as further discussed below.

When slopes vary across individuals the semiparametric bounds may be no tighter than the nonparametric ones. To illustrate consider a binary choice model with a single binary regressor X_{it}, where $Y_{it} = 1(\alpha_{i2}X_{it} + \alpha_{i1} > \varepsilon_{it})$, ε_{it} is independent of $(X_i, \alpha_{i2}, \alpha_{i1})$, and ε_{it} has known CDF H that is strictly increasing on the entire real line. The joint distribution of $H(\alpha_{i1})$ and $H(\alpha_{i2} + \alpha_{i1})$ conditional on $X_i = X^k$ is entirely unrestricted. Therefore when $X^k = (0, ..., 0)'$ the fact that $\mu(0|X^k) = E[H(\alpha_{i1})] = E[Y_{it}|X_i = X^k]$ is identified gives no information about $\mu(1|X^k) = E[H(\alpha_{i2} + \alpha_{i1})|X^i = X^k]$. Thus, $\mu(1|X^k)$ can be anything in the unit interval and the semiparametric bounds will be identical to the nonparametric bounds.

In the model of Assumption 6 the bound $\Delta_u^k - \Delta_\ell^k$ can be very tight for small T, as we show in numerical examples below, and may shrink exponentially fast as T grows. We show this result for the logit model with binary $X_{it} \in \{0, 1\}$ and leave other models to future work.

THEOREM 10: *If Assumptions 4, 5, and 6 are satisfied, $H(v) = e^v/(1 + e^v)$, $X_{it} \in \{0, 1\}$, and $X^k = (0, ..., 0)'$ or $X^k = (1, ..., 1)'$, then there are $C > 0$ and $1 > \varepsilon > 0$ such that*

$$\Delta_u^k - \Delta_\ell^k \leq C(1 - \varepsilon)^T.$$

To explain this exponential rate consider $X^k = (0, ..., 0)'$. Here $\mathcal{L}_j^k(\alpha, \beta) = H(\alpha)^{a_j}[1 - H(\alpha)]^{T-a_j}$ for $a_j = \sum_t Y_t^j$, so that by Assumption 5, $\mathcal{P}_j^k = E[H(\alpha)^{a_j}\{1 - H(\alpha)\}^{T-a_j}|X_i = X^k]$. From this it is straightforward to identify the first T moments of $H(\alpha)$ conditional on $X_i = X^k$.

Since β^* is identified for logit and because H is very smooth, the expectation of $H(\beta^*+\alpha)-H(\alpha)$ can then be approximated exponentially fast using these T moments as T grows, leading to $\Delta_u^k - \Delta_\ell^k$ also shrinking that fast. We expect that kind of result will extend to other smooth H and other discrete regressors X_{it}.

The bounds on the ATE Δ^k conditional on $X_i = X^k$ can be combined to obtain bounds for the overall ATE $\Delta = \int \Delta(\alpha, \beta^*) F^*(d\alpha)$, where F^* is the marginal CDF of α_i. The width of the bounds will now depend on $\Delta_u^k - \Delta_\ell^k$. Imposing the structure implied by the semiparametric model helps shrink the bounds when $\Delta_u^k - \Delta_\ell^k$ is smaller than in the nonparametric case. By Theorem 10, for logit the bounds will shrink to a point exponentially quickly as T grows even if the hypotheses on X_i from Theorems 7 and 8 are not satisfied.

8 Computation of Population Bounds

In this section we discuss computation of population bounds, give examples, and present theoretical results. A challenge for computation and for estimation is the dimensionality of unknown parameters and the nonlinearity of the probabilities in those parameters. Lemma 9 does show that we can take the individual effect CDF F_k to be $2J-1$ dimensional, but this dimension can be large, and the probabilities depend nonlinearly on the support points for the individual effect. We overcome this challenge by using an approximation with fixed but large number of support points for the individual effects. This approximation makes approximate probabilities and the ATE linear in parameters, simplifying computation. Honore and Tamer (2006) used a similar approach, but assumed that the true distribution of individual effects had known support points. We explicitly allow for approximation of unknown support points.

To describe how the approximation can be used to calculate the identified set, let M denote a number of support points for the individual effect and $\Upsilon_M = (\bar{\alpha}_{1M}, ..., \bar{\alpha}_{MM})'$ be a grid of fixed values for the individual effect. Also let $\pi = (\pi^{1\prime}, ..., \pi^{K\prime})'$ denote a $MK \times 1$ vector of possible probabilities, with each π^k an element of the M dimensional unit simplex $\mathcal{S}_M$. Approximate model probabilities are

$$P_j^k(\beta, \pi, M) = \sum_{m=1}^{M} \pi_m^k \mathcal{L}_j^k \left(\bar{\alpha}_{mM}, \beta\right).$$

Consider the function

$$T_\lambda(\beta, \pi, M) = \sum_{j,k} w_j^k \left[\mathcal{P}_j^k - P_j^k(\beta, \pi, M)\right]^2 + \lambda_M \pi'\pi,$$

where w_j^k are positive weights, such as the chi-square ones $\mathcal{P}^k/\mathcal{P}_j^k$, and $\lambda_M > 0$ is a penalty multiplier that controls the impact of the penalty term $\lambda_M \pi'\pi$. This term is present to help regularize the objective function by ensuring a nonsingular Hessian matrix. Let $\tilde{T}_\lambda(\beta, M) =$

$\min_{\pi \in \mathcal{S}_M^K} T_\lambda(\beta, \pi, M)$ and let $\epsilon_M > 0$ be a positive scalar. We approximate the identified set for β by

$$B(M) = \{\beta : \tilde{T}_\lambda(\beta, M) \leq \epsilon_M\}.$$

We calculate the identified set by letting M grow, and λ_M and ϵ_M shrink until there is little change in $B(M)$. Calculation of $\tilde{T}_\lambda(\beta, M)$ is straightforward because it is the minimum of a quadratic function. In practice we have found that $B(M)$ changes little as M increases even when M is quite small. As M grows and ϵ_M shrinks the set $B(M)$ will converge to the identified set under conditions given below.

For the ATE bounds, note

$$D^k(M) = \{\sum_{m=1}^{M} \pi_m^k \Delta(\bar{\alpha}_{mM}, \beta) : T_\lambda(\beta, \pi, M) \leq \epsilon_M\}$$

is the set of possible conditional ATE (given $X = X^k$) that are consistent with $\tilde{T}_\lambda(\beta, M) \leq \epsilon_M$. Approximate lower and upper bounds are

$$\Delta_\ell^k(M) = \min D^k(M), \Delta_u^k(M) = \max D^k(M).$$

As M grows and ϵ_M shrinks these bounds will converge to Δ_ℓ^k and Δ_u^k respectively, under conditions given below.

Computation of these ATE bounds is challenging because it requires searching over a large dimensional set of possible π. In practice we start with a smaller set of probabilities and then try others. Specifically, let $\tilde{\pi}(\beta) \in \arg\min T_\lambda(\beta, \pi, M)$, $\tilde{S}^k(\beta) = \{\pi^k : P_j^k(\beta, \pi, M) = P_j^k(\beta, \tilde{\pi}(\beta), M),\ j = 1, ..., J\}$, and

$$\tilde{\Delta}_\ell^k(M) = \min_{\beta \in B(M), \pi^k \in \tilde{S}^k(\beta)} \sum_{m=1}^{M} \pi_m^k \Delta(\bar{\alpha}_{mM}, \beta),\ \tilde{\Delta}_u^k(M) = \max_{\beta \in B(M), \pi^k \in \tilde{S}^k(\beta)} \sum_{m=1}^{M} \pi_m^k \Delta(\bar{\alpha}_{mM}, \beta).$$

For each β these bounds are easy to calculate by linear programming. We have done so and then checked to see if other values π violate these bounds. We have not found this to be so for values of M that we use to compute β. We conjecture that these bounds also converge to the population bounds as $M \longrightarrow \infty$ although we have not yet been able to prove this (because we have not been able to show that the ATE bounds are continuous in the true probabilities).

We carry out some numerical calculations for the probit model where

$$Y_{it} = 1(\beta^* X_{it} + \alpha_i \geq \varepsilon_{it}), \varepsilon_{it} \sim N(0,1), X_{it} = 1(\alpha_i \geq \eta_{it}), \eta_{it} \sim N(0,1), \alpha_i \sim N(0,1).$$

We consider different DGPs indexed by $\beta^* \in [-2, 2]$ and $T \in \{2, 3\}$. Figures 2 and 3 show nonparametric bounds for ATEs and semiparametric bounds for β^* and ATEs for $T = 2$ and $T = 3$, respectively. The semiparametric bounds are obtained using the computational algorithm

described above with $M = 100$ and $\lambda_M = 1.3 \times 10^{-8}$. The elements of the fixed grid Υ_M are located at the percentiles of the standard normal distribution. We find that β^* is not identified for $T = 2$, extending the result of Chamberlain (2010) to this example without time dummy. This result also holds for $T = 3$, although it is difficult to appreciate in the figure because the identified set B is very small. The nonparametric bounds for the ATEs (NP-bounds) can be very wide, even when we impose monotonicity (NPM-bounds) as described in Appendix B. The semiparametric bounds for the ATEs (SP-bounds) are tighter than the nonparametric bounds and shrink very fast with T. In Appendix B we report similar results for the logit, including nonidentification of the ATEs, except that β^* is identified, as is well known.

To show that the approximate sets converge to the identified set as M grows we impose some conditions. Let $d(\alpha, \tilde{\alpha})$ denote a metric on the set Υ of possible values for α.

ASSUMPTION 7: *(i)* Υ *is a compact metric space with metric* $d(\alpha, \tilde{\alpha})$*; ii)* $\eta(M) = \sup_{\alpha \in \Upsilon} \min_{\tilde{\alpha} \in \Upsilon_M} d(\alpha, \tilde{\alpha}) \longrightarrow 0$ as $M \longrightarrow \infty$*; (iii)* $\mathbb{B}$ *is a compact subset of* $\Re^b$*; (iv) there is* C *such that for all* $(\alpha, \beta), (\tilde{\alpha}, \tilde{\beta}) \in \Upsilon \times \mathbb{B}$, $\left| \mathcal{L}_j^k \left(\tilde{\alpha}, \tilde{\beta} \right) - \mathcal{L}_j^k \left(\alpha, \beta \right) \right| \leq C[d(\tilde{\alpha}, \alpha) + \left\| \tilde{\beta} - \beta \right\|]$*; and iv)* $\Delta(\alpha, \beta)$ *is continuous on* $\Upsilon \times \mathbb{B}$.

Although condition (i) seems restrictive, unbounded individual effects may be allowed if Υ is chosen appropriately. For example, in the binary choice model of Assumption 6 this condition will be satisfied if Υ is taken to be a two-point compactification of the real line and $d(\alpha, \tilde{\alpha})$ is specified appropriately, as shown by the following result.

LEMMA 11: *If Assumptions 4, 5, and 6 are satisfied and* $\mathbb{B}$ *is a compact subset of* $\Re^b$ *then there is a metric* $d(\alpha, \tilde{\alpha})$ *and for each* M *there is* $\Upsilon_M = \{\bar{\alpha}_{1M}, ..., \bar{\alpha}_{MM}\}$ *such that Assumption 7 is satisfied with* $\eta(M) = 1/(M-1)$.

For the convergence results for the identified set we use the Hausdorff set metric,

$$d_H(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\}.$$

THEOREM 12: *If Assumptions 4, 5, and 7 are satisfied,* $\epsilon_M \longrightarrow 0$*, and* $(\eta(M) + \lambda_M)/\epsilon_M \longrightarrow 0$ *then as* $M \longrightarrow \infty$,

$$d_H(B(M), B) \longrightarrow 0, \Delta_\ell^k(M) \longrightarrow \Delta_\ell^k, \Delta_u^k(M) \longrightarrow \Delta_u^k.$$

9 Estimation

The estimation method is like the computational one in using linear in parameters approximations to the probabilities. Let $\hat{P}^k$ and $\hat{P}_j^k$ denote sample cell frequencies, $\hat{w}_j^k$ estimated weights, $\hat{M}$ a choice of M that may depend on the data and sample size, and

$$\hat{T}_\lambda(\beta,\pi) = \sum_{j,k} \hat{w}_j^k \left[\hat{P}_j^k - P_j^k(\beta,\pi,\hat{M})\right]^2 + \lambda_n \pi'\pi.$$

Let $\hat{T}_\lambda(\beta) = \min_{\pi \in \mathcal{S}_M^K} \hat{T}_\lambda(\beta,\pi)$ and $\epsilon_n > 0$ be a positive scalar. We estimate the identified set for β by

$$\hat{B} = \{\beta \in \mathbb{B} : \hat{T}_\lambda(\beta) \leq \epsilon_n\}.$$

where $\mathbb{B}$ is the parameter space and ϵ_n is a cut-off parameter that shrinks to zero with the sample size, as in Manski and Tamer (2002) and Chernozhukov, Hong, and Tamer (2007).

The ATE bounds can be estimated by

$$\hat{\Delta}_\ell^k = \min \hat{D}^k, \hat{\Delta}_u^k = \max \hat{D}^k, \hat{D}^k = \{\sum_{m=1}^M \pi_m^k \Delta(\bar{\alpha}_{mM},\beta) : \hat{T}_\lambda(\beta,\pi) \leq \epsilon_n\}$$

We also try simpler estimates of the bounds corresponding to those described in the computation section. Specifically, for $\hat{\pi}(\beta) \in \arg\min_{\pi \in S_M^K} \hat{T}_\lambda(\beta,\pi)$ let $\hat{S}^k(\beta) = \{\pi^k : P_j^k(\beta,\pi,\hat{M}) = P_j^k(\beta,\hat{\pi}(\beta),\hat{M}),\ j = 1,...,J\}$ and let

$$\check{\Delta}_\ell^k = \min_{\beta \in \hat{B}, \pi^k \in \hat{S}^k(\beta)} \sum_{m=1}^M \pi_m^k \Delta(\bar{\alpha}_{mM},\beta),\ \check{\Delta}_u^k = \max_{\beta \in \hat{B}, \pi^k \in \hat{S}^k(\beta)} \sum_{m=1}^M \pi_m^k \Delta(\bar{\alpha}_{mM},\beta).$$

We use these estimated bounds as starting values and then search over other possible values of π, similar to the computational approach.

This approach to estimation (and computation) can be easily be modifed to handle the case where the distribution of the individual effect is restricted to be the same across some values of k. Such a modification could be implemented by imposing equality of π_m^k across those values of k. An example would be where the distribution of α_i did not depend on some component of X_{it}. That restriction could be imposed setting π_m^k to be equal across k where the other components of X_{it} do not vary. Or in a case with a lagged dependent variable we could restrict the distribution of α to only depend on the initial condition by imposing equality of π_m^k across all k where Y_{i0} takes on a particular value.

The choice of $\hat{M}$ is important for this estimator. In our empirical examples we have proceeded by starting with a small $\hat{M}$, and stopping when the change in the estimated sets is small. We have found that quite small $\hat{M}$ often suffices. The choice of weights $\hat{w}_j^k$ is also important. The optimal choice, corresponding to minimum chi-square would be $\hat{w}_j^k = \mathcal{P}^k/\mathcal{P}_j^k$. Using sample frequencies

in place of population frequencies does not work well due to small cell sizes. One could use a two-step procedure where one first computes the identified set for weights like $\hat{w}_j^k = \hat{P}^k$ and then reestimates the identified set using weights $\hat{w}_j^k = \hat{P}_k / P_j^k(\beta, \hat{\pi}(\beta), \hat{M})$ for some $\beta \in \hat{B}$.

The following is a consistency result.

THEOREM 13: *If Assumptions 4, 5, and 7 are satisfied,* $\hat{w}_j^k \xrightarrow{p} w_j^k > 0$, $\hat{P}_j^k \xrightarrow{p} \mathcal{P}_j^k$, $\epsilon_n \longrightarrow 0$, *and* $\left(n^{-1} + \eta(\hat{M}) + \lambda_n\right)/\epsilon_n \xrightarrow{p} 0$, *then* $d_H(\hat{B}, B) \xrightarrow{p} 0, \hat{\Delta}_\ell^k \xrightarrow{p} \Delta_\ell^k, \hat{\Delta}_u^k \xrightarrow{p} \Delta_u^k$.

It is interesting to note that no upper limit is placed on M in this result or in Theorem 12. That is because the model is finite dimensional so there is no need for such a limit. Mathematically, a richer fixed grid simply corresponds to a bigger submodel of the finite dimensional model.

10 Inference

Under Assumptions 4 and 5 the complete description of the data generating process is provided by the parameter vector $(P_X', P')'$, where $P_X = (P^k, k = 1, ..., K)'$ and $P = (P_j^k, j = 1, ..., J, k = 1, ..., K)'$. The true value of the parameter vector is $\Pi = (\mathcal{P}_X', \mathcal{P}')'$ and the empirical estimate is $\hat{\Pi} = (\hat{P}_X', \hat{P}')'$, where $\hat{P}_X = (\hat{P}^k, k = 1, ..., K)'$ and $\hat{P} = (\hat{P}_j^k, j = 1, ..., J, k = 1, ..., K)'$. For the inference results we condition on the observed distribution of X and thus set $P_X = \mathcal{P}_X = \hat{P}_X$. We make the following assumption about data generating process.

ASSUMPTION 8: $\Pi \in \mathbb{P} = \{(P_X, P) : P^k > 0, P_j^k > 0; j = 1, ..., J, k = 1, ..., K\}$

10.1 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 probabilities $\mathcal{P}_j^k$ belong to the product S_J^K of K unit simplexes of dimension J. We can begin by constructing a confidence region for the true choice probabilities $\mathcal{P}$ by collecting all probabilities $P = (P_1^1, ..., P_J^1, ..., P_J^K)' \in S_J^K$ that pass a goodness-of-fit test:

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

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:

$$W(P,\hat{P}) = n\sum_{j,k}\hat{P}^k\frac{\left(\hat{P}^k_j - P^k_j\right)^2}{P^k_j}.$$

Stage 2. To construct confidence regions for marginal effects and any other structural parameters we project each $P \in CR_{1-\alpha}(\mathcal{P})$ onto $\Xi = \{P : \exists\beta \in \mathbb{B} \text{ with } \mathcal{F}_k(\beta, P) \neq \varnothing, \forall k = 1, ..., K\}$, the space of conditional choice probabilities that are compatible with the model. We obtain this projection $P^*(P)$ by solving the minimum distance problem:

$$P^*(P) = \arg\min_{\tilde{P}\in\Xi} W(\tilde{P}, P), \quad W(\tilde{P}, P) = n\sum_{j,k}\hat{P}^k\frac{(P^k_j - \tilde{P}^k_j)^2}{\tilde{P}^k_j}. \tag{7}$$

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 $P \in CR_{1-\alpha}(\mathcal{P})$ we solve

$$B^*(P) = \left\{\beta \in \mathbb{B} : \exists\tilde{P} \in P^*(P) \text{ with } \mathcal{F}_k(\beta, \tilde{P}) \neq \varnothing, k = 1, ..., K\right\}. \tag{8}$$

Denote the resulting confidence region as

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

We may interpret this set as a confidence region for the set B^* of β that are compatible with a best approximating model. Under correct specification, this will be a confidence region for the identified set B.

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

$$\begin{aligned}
\Delta^k_\ell(P) &= \min_{\beta\in B^*(P), F_k\in\mathcal{F}_k(\beta,P^*(P))}\int\Delta(\alpha,\beta)F_k(d\alpha),\\
\Delta^k_u(P) &= \max_{\beta\in B^*(P), F_k\in\mathcal{F}_k(\beta,P^*(P))}\int\Delta(\alpha,\beta)F_k(d\alpha).
\end{aligned}$$

Denote the resulting confidence regions as

$$CR_{1-\alpha}[\Delta^{k*}_\ell, \Delta^{k*}_u] = \{[\Delta^k_\ell(P), \Delta^k_u(P)] : P \in CR_{1-\alpha}(\mathcal{P})\}.$$

These sets are confidence regions for the sets $[\Delta^{k*}_\ell, \Delta^{k*}_u]$, where Δ^{k*}_ℓ and Δ^{k*}_u are the lower and upper bounds on the marginal effects induced by any best approximating model. Under correct specification, these will include the true upper and lower bounds on the marginal effect $[\Delta^k_\ell, \Delta^k_u]$ 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.

THEOREM 14: *If Assumptions 4, 5, and 8 are satisfied then for any* Π *satisfying Assumption 8*,

$$\lim_{n\to\infty} \Pr_{\Pi} \left\{ \begin{array}{lcl} \mathcal{P} & \in & CR_{1-\alpha}(\mathcal{P}) \\ B^* & \in & CR_{1-\alpha}(B^*) \\ [\Delta_\ell^{k*}, \Delta_u^{k*}] & \in & CR_{1-\alpha}[\Delta_\ell^{k*}, \Delta_u^{k*}], \forall k \end{array} \right\} = 1 - \alpha.$$

10.2 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 ATE is of interest. The estimators for parameters and ATE 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 modified projection method is well suited for performing simultaneous inference on all possible functionals of the parameter vector. In contrast, the perturbed bootstrap is better suited for performing inference on a given functional of the parameter vector, such as the average structural effect. In order to understand why the latter method can be much sharper than the former method in the case where a single functional is of interest, it suffices to think of how these methods perform in the simplest situation of inference about the mean of a multinomial distribution. In this case, the perturbed bootstrap will become asymptotically equivalent to the usual bootstrap, since the limit distribution is continuous with respect to the DGP in this example, and our local perturbations of DGP converge to the true DGP (note that, more generally, in cases with limit distributions being discontinuous with respect to the DGP, introduction of the local perturbations ensures that the resulting confidence interval possesses locally uniform coverage.). Therefore in this example perturbed bootstrap inference asymptotically becomes first-order equivalent to the t-statistic-based inference on the mean, and is efficient. Now compare that with the Scheffe-style projection based confidence interval, whereby one creates a confidence region for multinomial probabilities and projects it down to the confidence interval for the mean, a linear functional of these probabilities. It is clear that the latter is very conservative, and is much less sharp than the t-statistic based confidence interval. We refer the reader to Romano and Wolf (2000) for the pertinent discussion of this example in the context

of a closely related inference method.

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).

We consider the problem of performing inference on a real parameter θ^*. For example, θ^* can be an upper (or lower) bound on the conditional ATE Δ^k such as

$$\theta^*(P) = \max_{\beta \in B^*(P), F_k \in \mathcal{F}_k(\beta, P^*(P))} \int \Delta(\alpha, \beta) F_k \left(d\alpha \right),$$

where P^* denotes the projection of P onto the model space, as defined in (7), and $B^*(P)$ is the corresponding projection for the identified set of the parameter defined as in (8). Alternatively, θ^* can be an upper (or lower) bound on a scalar functional $c'\beta^*$ of the parameter β^*. Then we define

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

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

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

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:

$$W(P, \hat{P}) = n \sum_{j,k} \hat{P}^k \frac{\left(\hat{P}_j^k - P_j^k\right)^2}{P_j^k}.$$

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

$$\underline{G}_n^{-1}(\alpha, \mathcal{P}) / \overline{G}_n^{-1}(\alpha, \mathcal{P}) = \inf / \sup_{P \in CR_{1-\gamma}(\mathcal{P})} G_n^{-1}(\alpha, P), \tag{9}$$

where $G_n^{-1}(\alpha, P) = \inf\{s : G_n(s, P) \geq \alpha\}$ is the α-quantile of the distribution function $G_n(s, P)$. 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 $\alpha = \alpha_1 + \alpha_2$,

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

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: *If Assumptions 4, 5, and 7 are satisfied then for any true parameter value* Π *satisfying Assumption 8,*

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

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

1. Draw a potential DGP $P_r = (P'_{r1}, ..., P'_{rK})$, where $P_{rk} \sim \mathcal{M}(n\hat{P}^k, (\hat{P}_1^k, ..., \hat{P}_J^k))/(n\hat{P}^k)$ and $\mathcal{M}$ denotes the multinomial distribution.

2. Keep P_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 $G_n(s, P_r)$ of $S_n(P_r)$ by simulation under the DGP P_r.

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

5. Let $\hat{\underline{G}}_n^{-1}(\alpha, \mathcal{P})/\hat{\overline{G}}_n^{-1}(\alpha, \mathcal{P}) = \min / \max\{G_n^{-1}(\alpha, P_1), ..., G_n^{-1}(\alpha, P_R)\}$, 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 $\underline{\theta} = \hat{\theta} - \hat{\overline{G}}_n^{-1}(1 - \alpha_1, \mathcal{P})$, $\overline{\theta} = \hat{\theta} - \hat{\underline{G}}_n^{-1}(\alpha_2, \mathcal{P})$, and $\alpha_1 + \alpha_2 = \alpha$.

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

11 Empirical Examples

We illustrate the estimation and inference results with two empirical examples. One obtains nonparametric bounds for the effect of unions on earnings quantiles. The other compares nonparametric and semiparametric bounds for the effect of fertility on women's labor force participation.

11.1 Union Premium

We revisit the empirical question of how unions impact the wage structure using panel data. Our major contribution here is to estimate the effect without imposing the assumption that unobserved heterogeneity is some additive term that can be simply differenced out. In our model unobserved heterogeneity can have an almost unrestricted impact on the structural/causal response functions, with the time homogeneity serving as the only restriction.

Our analysis is motivated by previous empirical studies that find unobserved differences between union and nonunion workers. For instance, in an influential study, Chamberlain (1982) finds strong evidence of heterogeneity bias in the estimation of the union effect by comparing estimates of cross-sectional models and panel data models with additive heterogeneity. This finding demonstrates the important need of controlling for unobserved heterogeneity. On the other hand, Angrist and Newey (1991) reject the hypothesis that the unobserved heterogeneity acts solely in an additive fashion. This finding demonstrates the important need of controlling for unobserved heterogeneity acting non-additively.

We use data from the National Longitudinal Survey (Youth Sample). The sample consists of full-time young working males, 20 to 29 year-old in 1986, followed over the period 1986 to 1993. We exclude individuals who failed to provide sufficient information for each year, were in the active army forces or students any year, or reported too high (more than $500 per hour) or too low (less than $1 per hour) wages. The final sample includes 2,065 men. We use the union membership and the log hourly wage rate in 1980 dollars as the covariate and the outcome

variables. The union membership variable reflects whether or not the individual had his wage set in collective bargaining agreement. We report results for panels with 2, 4, 6, and 8 years, all starting in 1986. Vella and Verbeek (1998) also used data from the NLSY for different years and found evidence of important union effect heterogeneity with a random effects model.

In our analysis, we focus on estimating the union effect for the subpopulation of workers that became ever unionized within the sample. For this subpopulation, the union effect is not point-identified, since there are 13% of the workers that stayed always unionized between 1986 and 1993 (see Table 2). However, we hope to construct informative bounds on the union effect. We consider both a static model that allows for the union membership decisions to be strictly exogenous with respect to wage setting decisions, and a dynamic model that allows for the union membership decisions to be only predetermined with respect to wage setting decisions. We shall also report the estimates of the union effect for the subpopulation of workers who change the union status at least once within the sample. For this subpopulation, the effect is point-identified, that is, the bounds on the union effect collapse to a point. We shall not estimate the union effect for the entire population of workers, since the bounds are completely uninformative in this case. This happens because more than half of the workers are never unionized within the sample (see Table 2).

We begin by presenting the estimates of the union effect for the subpopulation of workers who change the union status at least once within the sample. In Figure 4 we compare our panel data estimates of quantile effects with pooled cross-section estimates. In the cross-sectional estimates, we see that the quantile effect of union is positive but declines sharply at the upper end of the distribution, which agrees with previous cross-sectional findings (Chamberlain, 1994). A common explanation for this phenomenon is that the high-skill workers at the lower end of the earning distribution tend to join the union, whereas the high-skill workers at the high end of the earning distribution tend not to join the union. The estimated quantile effect in the cross-section therefore captures this selection effect of unobserved skills. In the panel data estimates, which control for the unobserved skills, we see that the quantile effects of union become very flat across the quantile indices. Thus, by controlling for individual heterogeneity, we have eliminated the selection effect. Finally, our estimates of quantile effects are higher in the dynamic model than in the static model indicating a possible dynamic feedback between the wage setting and union membership decisions. Figure 5 shows that the results of the static model are not sensitive to the inclusion of location and scale effects, as described in Appendix B.

We next present estimated bounds on the union effect for the subpopulation of workers that became ever unionized within the sample. In Figures 6 and 7 we show these bounds for both static and dynamic models and for panels of lengths $T \in \{2, 4, 6, 8\}$. In both cases, the size of the bounds decreases substantially with T. The bounds for $T = 8$ are informative, and show

that the effect is positive for most of the quantile indices. The figures also show bounds obtained using the assumption of monotonic and positive union effect on earnings described in Appendix B. These bounds are also informative, and in fact are substantially tighter than the bounds obtained without the monotonicity assumption.

Figure 8 plots 90% uniform confidence bands for the identified union effect and quantile union effect on ever unionized workers in the static and dynamic models. They are constructed by bootstrap with 500 repetitions. These bands allow us to make visual simultaneous inference on the entire quantile functions. For example, we cannot reject that the identified union effect is constant and positive for all the quantiles. For the ever unionized, the quantile union effect is positive for a large range of quantiles. The bands are narrower in the static model because this model uses more observations in the estimation of the quantile functions.

11.2 Female Labor Force Participation

For an application of the semiparametric bounds we consider a binary choice panel model of female labor force participation. We focus on the relationship between participation and the presence of young children in the household. Other studies that estimate similar models of participation in panel data include Heckman and MaCurdy (1980, 1982), Heckman and MaCurdy (1982), Chamberlain (1984), Hyslop (1999), Chay and Hyslop (2000), Carrasco (2001), Carro (2007), and Fernández-Val (2009).

The empirical analysis is based on a sample of married women from the National Longitudinal Survey of Youth 1979 (NLSY79). 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 3. 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

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

where α_i is an individual specific effect. The parameters of interest are β^* and the ATE of fertility on participation. We compute nonparametric and semiparametric probit and logit bounds for these parameters. We also obtain linear and nonlinear fixed effects estimates, together with large-T analytical bias corrected estimates and conditional fixed effects logit estimates.[1] The nonparametric bounds impose monotonicity of the effects. For the semiparametric bounds, we use the algorithm described in Section 9 with penalty $\lambda_n = 1/(n \log n)$ and iterate the quadratic program 3 times with initial weights $\hat{w}_j^k = \hat{P}^k$. This iteration makes the estimates insensitive to the penalty and weighting. We search over discrete distributions with $M = 23$ support points at $\{-\infty, -4, -3.6, ..., 3.6, 4, \infty\}$ for the parameter β^*, and with $M = 163$ support points at $\{-\infty, -8, -7.9, ..., 7.9, 8, \infty\}$ for the ATE. The estimates are based on panels of 2 and 3 time periods, both of them starting in 1990.

Table 4 reports estimates and 95% confidence regions for the parameters of interest. The confidence regions for the nonparametric bounds are constructed using the normal approximation (95% N) and nonparametric bootstrap with 200 repetitions (95% B). The confidence regions for the semiparametric bounds are obtained using the procedures described in Section 10. For the modified projection method (95% MP), 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. For the perturbed bootstrap method (95% PB) we use $R = 100$, $\gamma = .01$, $\alpha_1 = \alpha_2 = .02$, and 200 simulations from each DGP to approximate the distribution of the statistic. Together the modified projection and the perturbed bootstrap took several days to compute on personal computer. We also include confidence intervals obtained by a canonical projection method (95% CP) that intersects the nonparametric confidence interval for $\mathcal{P}$ with the space of probabilities compatible with the semiparametric model Ξ:

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

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 nonparametric bound estimates and confidence regions are too wide to provide informative evidence about the relationship between participation and fertility. The semiparametric bounds offer a good compromise between producing more informative results without adding too much structure to the model. Thus, these estimates are always inside the confidence regions of the nonparametric 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

[1] The analytical corrections use the estimators of the bias based on expected quantities in Fernández-Val (2009).

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.

The semiparametric intervals for the ATE cover the -9.6% estimate of Chamberlain (1984) for the expected effect of having an additional young child on the participation probability. He obtained this estimate from a correlated random coefficient probit model, a richer specification that includes education and fertility covariates, and a different sample from the PSID.

12 Possible Extensions

Our analysis is mainly confined to models with only discrete explanatory variables. It would be interesting to extend the analysis to models with continuous explanatory variables.

13 Appendix

13.1 Proofs

Proof of Theorem 1: By Assumption 2, for $\tilde{\alpha} = X$,

$$\begin{aligned} E[Y_{it}|X_i, \tilde{\alpha}_i] &= E[g_0(X_{it}, \alpha_i, \varepsilon_{it})|X_i] = \int g_0(X_{it}, \alpha, \varepsilon) F(d\alpha, d\varepsilon|\tilde{\alpha}_i) = m_0(X_{it}, \tilde{\alpha}_i), \\ \int m_0(x, \tilde{\alpha}) F(d\tilde{\alpha}) &= \int g_0(x, \alpha, \varepsilon) F(d\alpha, d\varepsilon|\tilde{\alpha}) F(d\tilde{\alpha}) = \mu(x). \end{aligned}$$

Similarly, Assumption 3 implies, for $\tilde{\alpha}_i = (\alpha_i, X_{1i})$,

$$\begin{aligned} E[Y_{it}|X_{it}, ..., X_{i1}, \tilde{\alpha}_i] &= \int g_0(X_{it}, \alpha_i, \varepsilon) F(d\varepsilon|X_{it}, ..., X_{i1}, \alpha_i) \\ &= \int g_0(X_{it}, \alpha_i, \varepsilon) F(d\varepsilon|\tilde{\alpha}_i) = m_0(X_{it}, \tilde{\alpha}_i), \\ \int m_0(x, \tilde{\alpha}) F(d\tilde{\alpha}) &= \int g_0(x, \alpha, \varepsilon) F(d\varepsilon|\alpha, X_1) F(d\alpha, dX_1) \\ &= \int g_0(x, \alpha, \varepsilon) F(d\varepsilon, d\alpha, dX_1) = \mu(x). Q.E.D. \end{aligned}$$

Proof of Theorem 2: Note that $X_{it} - \bar{X}_i = (1 - r_i) d_{it}(1) - r_i d_{it}(0)$ and $r_i = \sum_{t=1}^{T} d_{it}(1)/T = 1 - \sum_t d_{it}(0)/T$. Then

$$\begin{aligned} E[\sum_{t=1}^{T} (X_{it} - \bar{X}_i) Y_{it}]/T &= E[\sum_{t=1}^{T} (X_{it} - \bar{X}_i) E[Y_{it}|X_i]]/T \\ &= E[\sum_{t=1}^{T} \{(1 - r_i) d_{it}(1) \mu(1|X_i) - r_i d_{it}(0) \mu(0|X_i)\}]/T = E[\sigma_i^2 \Delta(X_i)]. \end{aligned}$$

Also, $\sum_{t=1}^{T}(X_{it}-\bar{X}_i)^2/T=\sigma_i^2$ by definition. Then by the law of large numbers,

$$\sum_{i,t}(X_{it}-\bar{X}_i)Y_{it}/n \xrightarrow{p} TE[\sigma_i^2\Delta(X_i)], \sum_{i,t}(X_{it}-\bar{X}_i)^2/n \xrightarrow{p} TE[\sigma_i^2]>0,$$

so the conclusion follows by the continuous mapping theorem. Q.E.D.

Proof of Theorem 3: By the law of large numbers and iterated expectations,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} D_i\{\frac{\sum_{t=1}^{T} d_{it}(\tilde{x})Y_{it}}{T_i(\tilde{x})}-\frac{\sum_{t=1}^{T} d_{it}(\bar{x})Y_{it}}{T_i(\bar{x})}\} \xrightarrow{p} E[D_i\{\frac{\sum_{t=1}^{T} d_{it}(\tilde{x})Y_{it}}{T_i(\tilde{x})}-\frac{\sum_{t=1}^{T} d_{it}(\bar{x})Y_{it}}{T_i(\bar{x})}\}] \\
= \;& E[D_i\{\frac{\sum_{t=1}^{T} d_{it}(\tilde{x})E[Y_{it}|X_i]}{T_i(\tilde{x})}-\frac{\sum_{t=1}^{T} d_{it}(\bar{x})E[Y_{it}|X_i]}{T_i(\bar{x})}\}] \\
= \;& E[D_i\{\frac{\sum_{t=1}^{T} d_{it}(\tilde{x})\mu(\tilde{x}|X_i)}{T_i(\tilde{x})}-\frac{\sum_{t=1}^{T} d_{it}(\bar{x})\mu(\bar{x}|X_i)}{T_i(\bar{x})}\}] = E[D_i\{\mu(\tilde{x}|X_i)-\mu(\bar{x}|X_i)\}]
\end{aligned}$$

and $\sum_{i=1}^{n} D_i/n \xrightarrow{p} E[D_i]$. Dividing numerator and denominator in $\hat{\delta}$ by n and applying the continuous mapping theorem gives the result. Q.E.D.

Proof of Theorem 4: Note that

$$\mu(x)=E[1(T_i(x)>0)g_0(x,\alpha_i,\varepsilon_{i1})]+E[1(T_i(x)=0)g_0(x,\alpha_i,\varepsilon_{i1})].$$

Since $B_\ell \le g_0(x,\alpha_i,\varepsilon_{i1}) \le B_u$, multiplying through these inequalities by $1(T_i(x)=0)$ and taking expectations gives

$$B_\ell \bar{\mathcal{P}}(x) \le E[1(T_i(x)=0)g_0(x,\alpha_i,\varepsilon_{i1})] \le B_u \bar{\mathcal{P}}(x). \tag{10}$$

Also, by iterated expectations and $d_{it}(x)E[Y_{it}|X_i]=d_{it}(x)\mu(x|X_i)$,

$$\begin{aligned}
E[1(T_i(x) \;&>\; 0)\frac{\sum_{t=1}^{T} d_{it}(x)Y_{it}}{T_i(x)}] = E[1(T_i(x)>0)\frac{\sum_{t=1}^{T} d_{it}(x)E[Y_{it}|X_i]}{T_i(x)}] \\
&= E[1(T_i(x)>0)\frac{\sum_{t=1}^{T} d_{it}(x)\mu(x|X_i)}{T_i(x)}] = E[1(T_i(x)>0)\mu(x|X_i)] \\
&= E[1(T_i(x)>0)g_0(x,\alpha_i,\varepsilon_{it})].
\end{aligned}$$

Combining this equality with the bounds for $E[1(T_i(x)=0)g_0(x,\alpha_i,\varepsilon_{i1})]$ gives the first conclusion.

To show sharpness, let $\tilde{\alpha}_i=(\alpha_i,X_i)$ and $B_\ell \le C \le B_u$ be a constant. Define

$$g_0^C(x,\tilde{\alpha}_i,\varepsilon_{it})=1(T_i(x)>0)g_0(x,\alpha_i,\varepsilon_{it})+C\cdot 1(T_i(x)=0).$$

Note that $T_i(X_{it})>0$ with probability one, so that $g_0^C(X_{it},\tilde{\alpha}_i,\varepsilon_{it})=g_0(X_{it},\alpha_i,\varepsilon_{it})=Y_{it}$. Hence the conditional distribution of $(Y_{i1},...,Y_{iT})'$ given X_i is the same for g_0^C and $\tilde{\alpha}_i$ as for g_0 and

α_i. Also, because (α_i, X_i) is a one-to-one function of $(\tilde{\alpha}_i, X_i)$ it follows that Assumption 2 is satisfied with $\tilde{\alpha}_i$ replacing α_i. When $C = B_\ell$ the lower bound is attained and when $C = B_u$ the upper bound is attained. *Q.E.D.*

PROOF OF THEOREM 5: The proof of the bounds proceeds exactly as in the proof of Theorem 4 with $1(Y_{it} \leq y)$ replacing Y_{it} and $1(g_0(x, \alpha_i, \varepsilon_{it}) \leq y)$ replacing $g_0(x, \alpha_i, \varepsilon_{it})$. Note that

$$G(y, x) = E[1(T_i(x) > 0)1(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y)] + E[1(T_i(x) = 0)1(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y)].$$

Since $0 \leq 1(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y) \leq 1$, multiplying both sides by $1(T_i(x) = 0)$ and taking expectations gives

$$0 \leq E[1(T_i(x) = 0)1(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y)] \leq \bar{\mathcal{P}}(x).$$

Also, by iterated expectations and $d_{it}(x)E[1(Y_{it} \leq y)|X_i] = d_{it}(x)G(y, x|X_i)$ for $G(y, x|X_i) = E[1(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y)|X_i]$,

$$\begin{aligned} E[1(T_i(x) &> 0)\frac{\sum_{t=1}^T d_{it}(x)1(Y_{it} \leq y)}{T_i(x)}] = E[1(T_i(x) > 0)\frac{\sum_{t=1}^T d_{it}(x)E[1(Y_{it} \leq y)|X_i]}{T_i(x)}] \\ &= E[1(T_i(x) > 0)\frac{\sum_{t=1}^T d_{it}(x)G(y, x|X_i)}{T_i(x)}] = E[1(T_i(x) > 0)G(y, x|X_i)] \\ &= E[1(T_i(x) > 0)1(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y)]. \end{aligned}$$

Combining this equation with the bounds for $E[1(T_i(x) = 0)1(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y)]$ and inverting gives the first conclusion.

To show sharpness, define $\tilde{\alpha}_i$ and $g_0^C(x, \alpha_i, \varepsilon_{it})$ as in the previous proof, but now for any $C \in \mathbb{R}$. Let $G^C(y, x) = E[1(g_0^C(x, \alpha_i, \varepsilon_{it}) \leq y)]$ and note that

$$G^C(y, x) = G_\ell(y, x) + 1(y \geq C)\bar{\mathcal{P}}(x).$$

Since $G_\ell(y, x)$ is strictly increasing, the quantile structural function for g_0^C is

$$q^C(\lambda, x) = \begin{cases} q_u(\lambda, x), \lambda < G_\ell(C, x), \\ C, G_\ell(C, x) \leq \lambda \leq G_\ell(C, x) + \bar{\mathcal{P}}(x), \\ q_\ell(\lambda, x), \lambda > G_\ell(C, x) + \bar{\mathcal{P}}(x). \end{cases}$$

For $\bar{\mathcal{P}}(x) < \lambda < 1 - \bar{\mathcal{P}}(x)$, we have $q^C(\lambda, x) = q_\ell(\lambda, x)$ for C small enough that $G_\ell(C, x) + \bar{\mathcal{P}}(x) < \lambda$ and $q^C(\lambda, x) = q_u(\lambda, x)$ for C big enough. For $\lambda \leq \bar{\mathcal{P}}(x)$ we have $q^C(\lambda, x) = q_u(\lambda, x)$ for all C big enough and $\lim_{C \longrightarrow -\infty} q^C(\lambda, x) = -\infty$. For $\lambda \geq 1 - \bar{\mathcal{P}}(x)$ we have $q^C(\lambda, x) = q_\ell(\lambda, x)$ for all C small enough and $\lim_{C \longrightarrow \infty} q_C(\lambda, x) = +\infty$. Thus the bounds are sharp. *Q.E.D.*

PROOF OF THEOREM 6: Note that exactly one of the of dummy variables $d_{i1}^{\mathcal{X}}(x), ..., d_{iT}^{\mathcal{X}}(x), 1(T_i(x) = 0)$ is one for each X_i. Therefore,

$$\mu(x) = E[\sum_{t=1}^T d_{it}^{\mathcal{X}}(x)g_0(x, \alpha_i, \varepsilon_{i1})] + E[1(T_i(x) = 0)g_0(x, \alpha_i, \varepsilon_{i1})].$$

It follows similarly to eq. (10) that

$$B_\ell \bar{\mathcal{P}}(x) \leq E[1(T_i(x)=0)g_0(x,\alpha_i,\varepsilon_{iT})] \leq B_u \bar{\mathcal{P}}(x).$$

Also, by eq. (5),

$$\begin{aligned} E[\sum_{t=1}^{T} d_{it}^{\mathcal{X}}(x)Y_{it}] &= \sum_{t=1}^{T} E[d_{it}^{\mathcal{X}}(x)Y_{it}] = \sum_{t=1}^{T} E[d_{it}^{\mathcal{X}}(x)g_0(x,\alpha_i,\varepsilon_{iT})] \\ &= E[\sum_{t=1}^{T} d_{it}^{\mathcal{X}}(x)g_0(x,\alpha_i,\varepsilon_{iT})]. \end{aligned}$$

Combining this equality with the bounds for $E[1(T_i(x) = 0)g_0(x,\alpha_i,\varepsilon_{iT})]$ and noting that $E[g_0(x,\alpha_i,\varepsilon_{iT})] = E[g_0(x,\alpha_i,\varepsilon_{i1})]$ gives the first conclusion. The second conclusion follows by replacing Y_{it} and $g_0(x,\alpha_i,\varepsilon_{iT})$ by $1(Y_{it} \leq y)$ and $g_0(x,\alpha_i,\varepsilon_{iT})$ respectively, similarly to the proof of Theorem 5. *Q.E.D.*

Proof of Theorem 7: Let $\overset{as|\alpha_i}{\longrightarrow}$ denote almost sure convergence as $T \longrightarrow \infty$ conditional on α_i. Recall that $T_i(x) = \sum_{t=1}^{T} d_{it}(x)$. By the ergodic theorem and by the conditional support given α_i being equal to the marginal support, we have

$$T_i(x)/T \overset{as|\alpha_i}{\longrightarrow} E[d_{it}(x)|\alpha_i] = \Pr(X_{it} = x \mid \alpha_i) > 0.$$

Therefore $1(T_i(x) > 0) \overset{as|\alpha_i}{\longrightarrow} 1$ for any x in the support of X_{it}. Then for any $\tilde{x}$ and $\bar{x}$ in the support of X_{it},

$$D_i = 1(T_i(\tilde{x}) > 0)1(T_i(\bar{x}) > 0) \overset{as|\alpha_i}{\longrightarrow} 1.$$

Sincer $T_i(x)$ is an exchangeable function it follows that the distribution of $D_i Y_{it}$ is the same for each t conditional on α_i, so that E

$$\begin{aligned} &= E\left[D_i \sum_{t=1}^{T} d_{it}(\tilde{x})E[Y_{it}|X_i]/T_i(\tilde{x})\right] = E\left[D_i \mu(\tilde{x}|X_i)\right] \\ &= E[D_i E[g(\tilde{x},\alpha_i,\varepsilon_{i1})|X_i]] = E[D_i g(\tilde{x},\alpha_i,\varepsilon_{i1})]. \end{aligned}$$

Since $E[|g(\tilde{x},\alpha_i,\varepsilon_{i1})|] < \infty$ it follows that $E[|g(\tilde{x},\alpha_i,\varepsilon_{i1}||\alpha_i] < \infty$ with probability one, so by the dominated convergence theorem (DCT), $E[D_i g(\tilde{x},\alpha_i,\varepsilon_{i1})|\alpha_i] \longrightarrow E[g(\tilde{x},\alpha_i,\varepsilon_{i1})|\alpha_i]$ a.s. as $T \longrightarrow \infty$. Then by $|E[D_i g(\tilde{x},\alpha_i,\varepsilon_{i1})|\alpha_i]| \leq E[|g(\tilde{x},\alpha_i,\varepsilon_{i1}||\alpha_i]$ and another application of DCT we have

$$E[D_i g(\tilde{x},\alpha_i,\varepsilon_{i1})] = E[E[D_i g(\tilde{x},\alpha_i,\varepsilon_{i1})|\alpha_i]] \longrightarrow E[E[g(\tilde{x},\alpha_i,\varepsilon_{i1})|\alpha_i]] = \mu(\tilde{x}).$$

It follows similarly that $E\left[D_i \sum_{t=1}^{T} d_{it}(\bar{x})Y_{it}/T_i(\bar{x})\right] \longrightarrow \mu(\bar{x})$. Also, $E[D_i] \longrightarrow 1$ by the DCT, giving the first conclusion.

Next note that by the DCT,

$$\bar{\mathcal{P}}(x) = E[1(T_i(x) = 0)] \longrightarrow 0.$$

The first conclusion then follows by Theorem 4 or 6.

Next, for notational convenience, suppress the x argument. It follows as previously with $1(g_0(x, \alpha_i, \varepsilon_{it}) \leq y)$ replacing Y_{it} that for all y, as $T \longrightarrow \infty$

$$G_u(y) - G_\ell(y) \leq \bar{\mathcal{P}} \longrightarrow 0.$$

Consider any $0 < \lambda < 1$. Let T be large enough so that $\lambda < 1 - \bar{\mathcal{P}}$. Then $q_u(\lambda)$ is finite and $G_\ell(q_u(\lambda)) = \lambda = G(q(\lambda))$. It follows by $q_u(\lambda) \geq q(\lambda)$ that

$$0 \leq G(q_u(\lambda)) - G(q(\lambda)) = G(q_u(\lambda)) - G_\ell(q_u(\lambda)) \leq \bar{\mathcal{P}} \longrightarrow 0.$$

Since $G(y)$ is strictly monotonic in a neighborhood of $q(\lambda)$ and $q_u(\lambda) \geq q(\lambda)$, it follows that $q_u(\lambda) \longrightarrow q(\lambda)$. An analogous argument shows that $q_\ell(\lambda) \longrightarrow q(\lambda)$.*Q.E.D.*

Proof of Theorem 8: Let $\Pi_{t=1}^{T} 1(X_{it} \neq x)$ be the indicator function for the event that none of the elements of X_i is equal to x so that $\bar{\mathcal{P}}(x) = E[\Pi_{t=1}^{T} 1(X_{it} \neq x)]$. By iterated expectations, for $T > J$,

$$\begin{aligned}
\bar{\mathcal{P}}(x) &= E[E[\Pi_{t=1}^{T} 1(X_{it} \neq x)]] = E[\Pi_{t=1}^{T-1} 1(X_{it} \neq x) E[1(X_{iT} \neq x) | X_{i,T-1}, ..., X_{i1}, \alpha_i]] \\
&= E[\{\Pi_{t=1}^{T-1} 1(X_{it} \neq x)\} \Pr(X_{iT} \neq x | X_{i,T-1}, ..., X_{i,T-J}, \alpha_i)] \leq (1 - \varepsilon) E[\Pi_{t=1}^{T-1} 1(X_{it} \neq x)].
\end{aligned}$$

Repeating the argument for $T - 1, ..., J$ gives

$$\bar{\mathcal{P}}(x) \leq (1 - \varepsilon)^{T-J} E[\Pi_{t=1}^{J-1} 1(X_{it} \neq x)] \leq (1 - \varepsilon)^{T-J}.$$

The first conclusion then follows by Theorem 4 or 6.

Next suppress the x argument and proceed as in the proof of Theorem 7. Note that $G'(y) > D_x$ for all y in a neighborhood of $q(\lambda)$ and that $G(q_u(\lambda)) - G(q(\lambda)) \leq \bar{\mathcal{P}}$ for large enough T. Using these and previous bounds and a mean value expansion gives

$$(1 - \varepsilon)^{T-J} \geq \bar{\mathcal{P}} \geq G(q_u(\lambda)) - G(q(\lambda)) = G'(\bar{q}(\lambda))[q_u(\lambda) - q(\lambda)] \geq D_x[q_u(\lambda) - q(\lambda)] \geq 0,$$

where $\bar{q}(\lambda)$ lies between $q_u(\lambda)$ and $q(\lambda)$. Dividing by D_x then gives

$$D_x^{-1}(1 - \varepsilon)^{T-J} \geq q_u(\lambda) - q(\lambda) \geq 0.$$

An analogous argument gives $D_x^{-1}(1 - \varepsilon)^{T-J} \geq q(\lambda) - q_\ell(\lambda)$, so adding these inequalities gives the second conclusion. *Q.E.D.*

PROOF OF LEMMA 9: Let the vector of model probabilities for $(Y^1, ..., Y^J)$ be

$$\mathcal{L}^k(\alpha,\beta) \equiv \left(\mathcal{L}_1^k(\alpha,\beta), ..., \mathcal{L}_J^k(\alpha,\beta)\right)'.$$

Let $\Gamma_k(\beta) \equiv \left\{\mathcal{L}^k(\alpha,\beta) : \alpha \in \Upsilon\right\}$ and $\breve{\Gamma}_k(\beta)$ be the convex hull of $\Gamma_k(\beta)$. By Lemma 3 of Chamberlain (1987), $\breve{\Gamma}_k(\beta) = \{\int \mathcal{L}^k(\alpha,\beta) F(d\alpha) : F \text{ is a CDF on } \Upsilon\}$. Therefore, $\int \mathcal{L}^k(\alpha,\beta) F_k(d\alpha) \in \breve{\Gamma}_k(\beta)$. Note that $\Gamma_k(\beta)$ is contained in the unit simplex and so has dimension $J-1$. By the Carathéodory Theorem there exist J vectors $\mathcal{L}^k\left(\alpha_m^k,\beta\right), (m = 1, ..., J)$ and $0 \leq \pi_m^k \leq 1$ with $\sum_{m=1}^J \pi_m^k = 1$ such that

$$\int \mathcal{L}^k(\alpha,\beta) F_k(d\alpha) = \sum_{m=1}^J \pi_m^k \mathcal{L}^k\left(\alpha_m^k,\beta\right),$$

giving the conclusion for the discrete distribution F_k^J with J support points at $(\alpha_1^k, ..., \alpha_J^k)$ and probabilities $(\pi_1^k, ..., \pi_J^k)$.

Next, for any $\epsilon > 0$ let $\beta \in B$ and $F_{k\beta} \in \mathcal{F}_k(\beta, \mathcal{P})$ satisfy

$$\Delta_u^k - \epsilon < \int \Delta(\alpha,\beta) F_{k\beta}(d\alpha) \equiv \bar{\Delta}(\beta).$$

Similarly to the previous paragraph, let $\Gamma_k^\Delta(\beta) \equiv \left\{(\mathcal{L}^k(\alpha,\beta)', \Delta(\alpha,\beta))' : \alpha \in \Upsilon\right\}$ and $\breve{\Gamma}_k^\Delta(\beta)$ be the convex hull of $\Gamma_k^\Delta(\beta)$. Then $(\mathcal{P}_1^k, ..., \mathcal{P}_J^k, \bar{\Delta}(\beta))' \in \breve{\Gamma}_k^\Delta(\beta)$, so by Caratheodory's Theorem there exists a discrete distribution $F_{k\beta}^{J+1}$ with $J+1$ support points $(\alpha_1^k, ..., \alpha_{J+1}^k)$ and probabilities $\pi_1^k, ..., \pi_{J+1}^k$ such that $F_{k\beta}^{J+1} \in \mathcal{F}_k(\beta, \mathcal{P})$ and $\int \Delta(\alpha,\beta) F_{k\beta}^{J+1}(d\alpha) = \bar{\Delta}(\beta)$.

We now show that it suffices to have mass over just J points. Consider the problem of allocating $\pi_1^k, ..., \pi_{J+1}^k$ among $\left(\alpha_1^k, ..., \alpha_{J+1}^k\right)$ in order to solve

$$\begin{aligned} & \max_{(\pi_1^k,...,\pi_{J+1}^k)} \sum_{m=1}^{J+1} \Delta(\alpha_m^k,\beta)\pi_m^k, s.t. \\ \sum_{m=1}^{J+1} \pi_m^k \mathcal{L}_j^k\left(\alpha_m^k,\beta\right) &= \mathcal{P}_j^k, \sum_{m=1}^{J+1} \pi_m^k = 1, \pi_m^k \geq 0, (\ m = 1, ..., J+1). \end{aligned}$$

This is a linear program of the form

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

and any basic feasible solution to this program has $J+1$ active constraints, of which at most $rank(A) + 1$ can be equality constraints. This means that at least $J - \text{rank}(A)$ of active constraints are the form $\pi_m^k = 0$, see, e.g., Theorem 2.3 and Definition 2.9 (ii) in Bertsimas and Tsitsiklis (1997). Since $rank(A) \leq J-1$, a basic solution to this linear programming problem

will have at least one zero, that is at most J strictly positive π_m^k's.[2] Thus, we have shown that there exists a distribution $F_{k\beta}^J \in \mathcal{F}_k(\beta, \mathcal{P})$ with just J points of support such that

$$\Delta_u^k - \epsilon < \int \Delta(\alpha, \beta) F_{k\beta}^{J+1}(d\alpha) \leq \int \Delta(\alpha, \beta) F_{k\beta}^{J}(d\alpha).$$

This construction works for every $\epsilon > 0$. *Q.E.D.*

PROOF OF THEOREM 10: β^* is identified for logit so $B = \{\beta^*\}$. Consider here $k = 1$ where $X^k = (0, ..., 0)'$ and let $\mathcal{F}_1 = \mathcal{F}_k(\beta^*, \mathcal{P})$. The result for $X^k = (1, ..., 1)'$ will follow similarly. Let $Z = H(\alpha)$ and $G_1(z)$ be the CDF of Z when $F_1(\alpha)$ is the CDF of α. By $(Y_{i1}, ..., Y_{iT})$ mutually independent, for all $F_1 \in \mathcal{F}_1$,

$$\mathcal{P}_j^1 = \int z^{\sum_t Y_t^j} [1 - z]^{T - \sum_t Y_t^j} dG_1(z), (j = 1, ..., J).$$

Since $\sum_t Y_t^j$ takes on integer values, there are known functions M_t of the cell probabilities $\mathcal{P}_j^1$ such that for all $F_1 \in \mathcal{F}_1$,

$$M_t = \int z^t dG_1(z), (t = 1, ..., T).$$

Now consider a T^{th} order polynomial $P(z, T) = b_0 + b_1 z + ... + b_T z^T$ in z. Note that

$$\int P(z, T) dG_1(z) = b_0 + \sum_{t=1}^{T} b_t M_t$$

does not depend on $F_1 \in \mathcal{F}_1$. Similarly, $\int z dG_1(z) = M_1$ does not depend on $F_1 \in \mathcal{F}_1$. Define the function $h(z) = H(\beta^* + H^{-1}(z)) = ze^{\beta^*}/1 - (1 - e^{\beta^*})z$, so that the ATE is $\int [h(z) - z] dG_1^*(z)$. For any polynomial $P(z, t)$ let $R(z, t) = h(z) - P(z, t)$ be the remainder. Then we have

$$\begin{aligned} \Delta_u^k - \Delta_\ell^k &= \sup_{F_1 \in \mathcal{F}_1} \int [h(z) - z] dG_1(z) - \inf_{F_1 \in \mathcal{F}_1} \int [h(z) - z] dG_1(z) \qquad (11) \\ &= \sup_{F_1 \in \mathcal{F}_1} \int [P(z, T) + R(z, T)] dG_1(z) - \inf_{F_1 \in \mathcal{F}_1} \int [P(z, T) + R(z, T)] dG_1(z) \\ &= \sup_{F_1 \in \mathcal{F}_1} \int R(z, T) dG_1(z) - \inf_{F_1 \in \mathcal{F}_1} \int R(z, T) dG_1(z) \leq 2 \sup_{0 \leq z \leq 1} |R(z, T)|. \end{aligned}$$

The function $h(z)$ is continuously differentiable of order r for every r with

$$\left| \frac{d^r h(z)}{dz^r} \right| \leq r! e^{|\beta^*|} (e^{|\beta^*|} - 1|)^{r-1}.$$

[2]Note that $rank(A) \leq J - 1$, since $\sum_{j=1}^J \mathcal{L}_j^k(\alpha, \beta) = 1$. The exact rank of A depends on the sequence X^k, the parameter β, the form of $\mathcal{L}_j^k(\alpha, \beta)$, and T. For example in the model of Assumption 6 with $T = 2$ and X binary, $rank(A) = J - 2 = 2$ when $x_1 = x_2$, $\beta = 0$, or H is the logistic distribution; whereas $rank(A) = J - 1 = 3$ for $X_1^k \neq X_2^k$, $\beta \neq 0$, and H is any continuous distribution different from the logistic.

Then by Jackson's Theorem (e.g. Judd (1998) Chap. 3) there exists $P(z,T)$ such that for $\gamma = \pi(e^{|\beta^*|}-1|)/4$

$$\begin{aligned}\sup_{0\le z\le 1}|R(z,T)| &\le \frac{(T-r)!}{T!}\left(\frac{\pi}{4}\right)^r \sup_{0\le z\le 1}\left|\frac{d^r h(z)}{dz^r}\right| \\ &\le \frac{(T-r)!r!}{T!}\left(\frac{\pi}{4}\right)^r e^{|\beta^*|}(e^{|\beta^*|}-1|)^{r-1} \le C\left(\frac{r\gamma}{T}\right)^r.\end{aligned}$$

This inequality continues to hold if γ is replaced by $\max\{\gamma,1\}$, so we can assume $\gamma > 1$. Then choose r equal to $T/\gamma e$, so that

$$\sup_{0\le z\le 1}|R(z,T)| \le Ce^{-T/\gamma e}.$$

The conclusion then follows by eq. (11). Q.E.D.

Proof of Lemma 11: Consider the set $\bar{\Re} = (-\infty,+\infty)\cup\{-\infty,+\infty\}$. By Assumption 6 $H(v)$ is strictly monotonic and continuous on $\bar{\Re}$ with $H(-\infty)=0$ and $H(+\infty)=1$. Let $H^{-1}(u)$ be the inverse function defined on $[0,1]$. Let $\bar{v} = \max_{X^k\in\{X^1,\ldots,X^K\},\beta\in B}|X_t^{k\prime}\beta|$ and define the function

$$T(u) = \begin{cases} \bar{v} + H^{-1}(u),\ \frac{3}{4}\le u\le 1 \\ (4u-2)\left[\bar{v}+H^{-1}(\frac{3}{4})\right],\ \frac{1}{4}<u<\frac{3}{4} \\ -\bar{v}+H^{-1}(u),\ 0\le u\le\frac{1}{4}\end{cases}$$

This function is continuous and differentiable except at $u=\frac{1}{4}$ and $u=\frac{3}{4}$. At $u=\frac{1}{4}$ the left derivative is $\left[h(H^{-1}\left(\frac{1}{4}\right))\right]^{-1}$ and the right derivative is $4\left[\bar{v}+H^{-1}\left(\frac{3}{4}\right)\right]$.

Consider the function $H(v+T(u))$. By the chain rule, $H(v+T(u))$ is differentiable everywhere on $[-\bar{v},\bar{v}]\times\left(\frac{1}{4},\frac{3}{4}\right)$ and right differentiable at $\left(v,\frac{1}{4}\right)$ and left differentiable at $\left(v,\frac{3}{4}\right)$ with derivative (right or left) equal to

$$h(v+T(u))4\left[\bar{v}+H^{-1}(\frac{3}{4})\right].$$

This derivative is uniformly bounded on $[-\bar{v},\bar{v}]\times\left(\frac{1}{4},\frac{3}{4}\right)$ by h uniformly bounded. Also $H(v+T(u))$ is differentiable everywhere on $[-\bar{v},\bar{v}]\times\left\{\left(\frac{3}{4},\infty\right)\cup\left(-\infty,\frac{1}{4}\right)\right\}$, right differentiable at $[-\bar{v},\bar{v}]\times\left\{\frac{3}{4}\right\}$ and left differentiable at $[-\bar{v},\bar{v}]\times\left\{\frac{1}{4}\right\}$. For $u\in[3/4,1]$ the (right) derivative is

$$\frac{\partial}{\partial u}H(v+T(u)) = H'(v+T(u))T'(u) = \frac{h(v+\bar{v}+H^{-1}(u))}{h(H^{-1}(u))} \le \frac{h(H^{-1}(u))}{h(H^{-1}(u))} = 1$$

where the inequality holds by $\bar{v}+v\ge 0$ (implied by $v\ge-\bar{v}$) and by $H^{-1}(u)>0$. It follows similarly that $\partial H(v+T(u))/\partial u$ is uniformly bounded by 1 on $[-\bar{v},\bar{v}]\times[0,\frac{1}{4}]$. It follows that there is a constant C such that for all $v\in[-\bar{v},\bar{v}]$ and $u,\tilde{u}\in[0,1]$,

$$|H(v+T(\tilde{u}))-H(v+T(u))|\le C|\tilde{u}-u|.$$

Note that $T^{-1}(\alpha)$ is a strictly monotonic increasing function on $\bar{\Re}$. Define $d(\tilde{\alpha},\alpha) = |T^{-1}(\tilde{\alpha}) - T^{-1}(\alpha)|$. Note that $d(\tilde{\alpha},\alpha) \geq 0$ with equality if and only if $\tilde{\alpha} = \alpha$, and for any three points $\bar{\alpha}$, $\tilde{\alpha}$, and α, the triangle inequality implies

$$d(\tilde{\alpha},\alpha) = |T^{-1}(\tilde{\alpha}) - T^{-1}(\alpha)| \leq |T^{-1}(\tilde{\alpha}) - T^{-1}(\bar{\alpha})| + |T^{-1}(\bar{\alpha}) - T^{-1}(\alpha)| = d(\tilde{\alpha},\bar{\alpha}) + d(\bar{\alpha},\alpha).$$

Therefore $d(\tilde{\alpha},\alpha)$ is a metric. Also, for $\tilde{u} = T^{-1}(\tilde{\alpha})$ and $u = T^{-1}(\alpha)$, we have

$$\sup_{v\in[-\bar{v},\bar{v}]} |H(v+\tilde{\alpha}) - H(v+\alpha)| \leq C|T^{-1}(\tilde{\alpha}) - T^{-1}(\alpha)| = Cd(\tilde{\alpha},\alpha).$$

Also, by $|X_t^{k\prime}\beta| \leq \bar{v}$, and $0 \leq H(X_t^{k\prime}\beta+\alpha) \leq 1$, for all t, k, and $\beta \in \mathbb{B}$,

$$\begin{aligned}
\left|\mathcal{L}_j^k\left(\tilde{\alpha},\tilde{\beta}\right) - \mathcal{L}_j^k\left(\alpha,\beta\right)\right| &\leq \left|\mathcal{L}_j^k\left(\tilde{\alpha},\tilde{\beta}\right) - \mathcal{L}_j^k\left(\alpha,\tilde{\beta}\right)\right| + \left|\mathcal{L}_j^k\left(\alpha,\tilde{\beta}\right) - \mathcal{L}_j^k\left(\alpha,\beta\right)\right| \\
&\leq Cd(\tilde{\alpha},\alpha) + \sup_{\alpha,t,k} |H(X_t^{k\prime}\tilde{\beta}+\alpha) - H(X_t^{k\prime}\beta+\alpha)| \\
&\leq Cd(\tilde{\alpha},\alpha) + \sup_v h(v) \sup_{t,k} \left\|X_t^k\right\| \left\|\tilde{\beta}-\beta\right\| \\
&\leq C[d(\tilde{\alpha},\alpha) + \left\|\tilde{\beta}-\beta\right\|].
\end{aligned}$$

Finally, for every M let $\bar{\alpha}_{mM} = T((m-1)/(M-1)), (m = 1,...,M)$. Then

$$\eta(M) = \sup_{\alpha\in\bar{\Re}} \min_{\tilde{\alpha}\in\Upsilon_M} d(\alpha,\tilde{\alpha}) = \sup_{u\in[0,1]} \min_{\tilde{u}\in\{0,1/(M-1),2/(M-1),...,1\}} |u-\tilde{u}| = 1/(M-1). Q.E.D.$$

Proof of Theorem 12: This is proof is omitted because it is very similar (but easier) than the proof of Theorem 13 to follow.

Proof of Theorem 13: For notational convenience we here denote the probabilities associated with the fixed grid $\{\bar{\alpha}_{1M},...,\bar{\alpha}_{MM}\}$ by $\bar{\pi}^k$. Let $\bar{\pi} = (\bar{\pi}^{1\prime},...,\bar{\pi}^{K\prime})'$ be a $KM \times 1$ vector with each $\bar{\pi}^k$ in the M-dimensional unit simplex $\mathcal{S}_M$. Also, let the probabilities associated with a variable grid $\{\alpha_1^k,...,\alpha_{J+1}^k\}$ be π^k so that $\pi = (\pi^{1\prime},...,\pi^{K\prime})'$ is a $[(J+1)K] \times 1$ vector of probabilities with each π^k in the $J+1$-dimensional unit simplex $\mathcal{S}_{J+1}$. Let $\alpha^k = (\alpha_1^k,...,\alpha_{J+1}^k)'$, $\alpha = (\alpha^{1\prime},...,\alpha^{K\prime})'$, $\gamma = (\alpha',\pi')'$, $\theta = (\beta',\gamma')'$, $\tilde{P}_j^k(\theta) = \sum_{\ell=1}^{J+1} \mathcal{L}_j^k\left(\alpha_\ell^k,\beta\right)\pi_\ell^k$, $\Delta^k(\theta) = \sum_{\ell=1}^{J+1} \Delta\left(\alpha_\ell^k,\beta\right)\pi_\ell^k$, $\Theta = \mathbb{B}\times\Upsilon^{(J+1)K} \times \mathcal{S}_{J+1}^K$, and

$$\hat{Q}(\theta) = \sum_{j,k} \hat{w}_j^k \left[\hat{P}_j^k - \tilde{P}_j^k(\theta)\right]^2, Q(\theta) = \sum_{j,k} w_j^k \left[\mathcal{P}_j^k - \tilde{P}_j^k(\theta)\right]^2.$$

By applying the Caratheodory Theorem as in the proof of Lemma 12, for every $\bar{\pi}$ there is $\theta(\bar{\pi},\beta) = (\beta',\gamma(\bar{\pi},\beta)')'$ with

$$\Delta^k(\theta(\bar{\pi},\beta)) = \sum_{m=1}^M \Delta(\bar{\alpha}_{mM},\beta)\bar{\pi}_m^k, \tilde{P}_j^k(\theta(\bar{\pi},\beta)) = P_j^k(\beta,\bar{\pi},M), (j = 1,...,J; k = 1,...,K).$$

Let $\Theta_I = \{\theta : Q(\theta) = 0\}$,

$$\tilde{\Theta} = \{\theta(\bar{\pi}, \beta) : \hat{Q}(\theta(\bar{\pi}, \beta)) + \lambda_n \bar{\pi}'\bar{\pi} \leq \epsilon_n\}, \Theta_M = \{\theta(\bar{\pi}, \beta) : \bar{\pi} \in \mathcal{S}_M^K, \beta \in \mathbb{B}\}.$$

By construction the projection of $\tilde{\Theta}$ on $\mathbb{B}$ coincides with $\hat{B}$ and the projection of $\Theta_{\mathcal{I}}$ on $\mathbb{B}$ coincides with B. Also the identified set of marginal effects is $\{\Delta^k(\theta) : \theta \in \Theta_I\}$, $\Delta^k(\theta)$ is a continuous function of θ, and $\hat{D}^k = \{\Delta^k(\theta) : \theta \in \tilde{\Theta}\}$. Since the minimum and maximum of a set are continuous in the Hausdorff metric, it suffices to show that $d_H(\tilde{\Theta}, \Theta_I) \xrightarrow{p} 0$.

Let $d(\theta, \tilde{\theta}) = \max_{j,k} \max\{d(\alpha_j^k, \tilde{\alpha}_j^k), |\pi_j^k - \tilde{\pi}_j^k|, \left\|\beta - \tilde{\beta}\right\|\}$. From Assumption 7 and $\hat{M} \xrightarrow{p} \infty$ we have

$$\sup_{\alpha \in \Upsilon} \min_{\tilde{\alpha} \in \Upsilon_{\hat{M}}} d(\alpha, \tilde{\alpha}) \leq \eta(\hat{M}) \xrightarrow{p} 0.$$

Therefore for every $\alpha \in \Upsilon$ there is $\bar{\alpha}_{m(\alpha),\hat{M}}$ with $d(\alpha, \bar{\alpha}_{m(\alpha),\hat{M}}) \leq \eta(\hat{M})$, so that for any $\theta \in \Theta$ there are $\bar{\alpha}_{m(\alpha_\ell^k),\hat{M}}$ with $\max_{1 \leq \ell \leq J+1, k}\{d(\alpha_\ell^k, \bar{\alpha}_{m(\alpha_\ell^k),\hat{M}})\} \leq \eta(\hat{M})$. Let $\alpha^k(\theta) = (\bar{\alpha}_{m(\alpha_1^k),\hat{M}}, ..., \bar{\alpha}_{m(\alpha_{J+1}^k),\hat{M}})'$, $\alpha(\theta) = (\alpha^1(\theta)', ..., \alpha^K(\theta)')'$, and $\bar{\theta}(\theta) = (\beta', \alpha(\theta)', \pi')'$. By construction, $\bar{\theta}(\theta) \in \Theta_M$ and $d(\bar{\theta}(\theta), \theta) \leq \eta(\hat{M})$. Thus,

$$\sup_{\theta \in \Theta} \inf_{\tilde{\theta} \in \Theta_{\hat{M}}} d(\theta, \tilde{\theta}) \leq \eta(\hat{M}).$$

Also, by Assumption 7,

$$|\tilde{P}_j^k(\theta) - \tilde{P}_j^k(\tilde{\theta})| \leq \sum_{\ell=1}^{J} \left| \mathcal{L}_j^k\left(\alpha_\ell^k, \beta\right) \pi_\ell^k - \mathcal{L}_j^k\left(\tilde{\alpha}_\ell^k, \tilde{\beta}\right) \tilde{\pi}_\ell^k \right| \leq C d(\theta, \tilde{\theta}).$$

It then follows by standard calculations that there is $\hat{C} = O_p(1)$ such that

$$|\hat{Q}(\theta) - \hat{Q}(\tilde{\theta})| \leq \hat{C} d(\theta, \tilde{\theta}) \text{ for all } \theta, \tilde{\theta} \in \Theta.$$

Therefore we have

$$\sup_{\theta \in \Theta} \inf_{\tilde{\theta} \in \Theta_{\hat{M}}} |\hat{Q}(\theta) - \hat{Q}(\tilde{\theta})| \leq \hat{C} \eta(\hat{M}).$$

Also note that

$$\sup_{\theta \in \Theta_I} \hat{Q}(\theta) = \sum_{j,k} \hat{w}_j^k [\hat{P}_j^k - \mathcal{P}_j^k]^2 = O_p(n^{-1}).$$

Next let $\delta > 0$ be any positive constant and define the events

$$\mathcal{E}_1 = \left\{\eta(\hat{M}) < \delta\right\}, \mathcal{E}_2 = \left\{\hat{C}\eta(\hat{M}) < \frac{\epsilon_n}{3}\right\}, \mathcal{E}_3 = \left\{\sup_{\theta \in \Theta_I} \hat{Q}(\theta) < \frac{\epsilon_n}{3}\right\}, \mathcal{E}_4 = \sup_{\bar{\pi} \in \mathcal{S}_M^K} \lambda_n \bar{\pi}'\bar{\pi} < \frac{\epsilon_n}{3}.$$

By $(n^{-1} + \eta(\hat{M}) + \lambda_n)/\epsilon_n \xrightarrow{p} 0$ it follows that

$$\begin{aligned} \Pr(\mathcal{E}_1) &\longrightarrow 1, \Pr(\mathcal{E}_2) = \Pr\left(\hat{C} < \frac{\eta(\hat{M})^{-1}\epsilon_n}{3}\right) \longrightarrow 1, \\ \Pr(\mathcal{E}_3) &= \Pr\left(n \sup_{\theta \in \Theta_I} \hat{Q}(\theta) < \frac{n\epsilon_n}{3}\right) \longrightarrow 1, \Pr(\mathcal{E}_4) \geq \Pr(\lambda_n K \leq \frac{\epsilon_n}{3}) \longrightarrow 1. \end{aligned}$$

It follows that $\Pr(\cap_{r=1}^4 \mathcal{E}_r) \longrightarrow 1$. When $\cap_{r=1}^4 \mathcal{E}_r$ occurs then for every $\theta \in \Theta_I$ there is $\bar{\pi}$ with $\theta_M = \theta(\bar{\pi}, \beta) \in \Theta_M$ such that $d(\theta, \bar{\theta}) < \delta$ and

$$\begin{aligned} \hat{Q}(\bar{\theta}) + \lambda_n \bar{\pi}'\bar{\pi} &\leq \hat{Q}(\bar{\theta}) + \frac{\epsilon_n}{3} \leq \hat{Q}(\theta) + \hat{Q}(\bar{\theta}) - \hat{Q}(\theta) + \frac{\epsilon_n}{3} \\ &\leq \sup_{\theta \in \Theta_I} \hat{Q}(\theta) + \hat{C}\hat{\eta}(M) + \frac{\epsilon_n}{3} \leq \epsilon_n, \end{aligned}$$

i.e. $\bar{\theta} \in \tilde{\Theta}$. Thus, with probability approaching one,

$$\sup_{\theta \in \Theta_I} \inf_{\tilde{\theta} \in \tilde{\Theta}} d(\theta, \tilde{\theta}) \leq \delta.$$

Next, note that $\hat{Q}(\theta) \xrightarrow{p} Q(\theta)$ so it follows by Theorem 2.1 of Newey (1991) that $\sup_{\theta \in \Theta} \left| \hat{Q}(\theta) - Q(\theta) \right| \xrightarrow{p} 0$. Define $\Theta_I^\delta = \left\{ \theta : \inf_{\tilde{\theta} \in \Theta_I} d(\theta, \tilde{\theta}) < \delta \right\}$. Note that Θ_I^δ is open so that $\Theta \backslash \Theta_I^\delta$ is compact, so by continuity of $Q(\theta)$, $\inf_{\Theta \backslash \Theta_I^\delta} Q(\theta) = \rho > 0$. It follows by uniform convergence that $\inf_{\Theta \backslash \Theta_I^\delta} \hat{Q}(\theta) > \frac{\rho}{2}$ with probability approaching 1 (w.p.a. 1). By $\epsilon_n \to 0$,

$$\sup_{\theta \in \tilde{\Theta}} \hat{Q}(\theta) \leq \sup_{\bar{\pi}} \{ \hat{Q}(\theta(\bar{\pi}, \beta)) + \lambda_n \bar{\pi}'\bar{\pi} \leq \epsilon_n \} < \rho/2,$$

so that $\tilde{\Theta} \subseteq \Theta_I^\delta$. Therefore w.p.a.1 for all $\tilde{\theta} \in \tilde{\Theta}$ there exists $\theta \in \Theta_I$ such that $d(\tilde{\theta}, \theta) < \delta$, i.e. $\sup_{\tilde{\theta} \in \tilde{\Theta}} \inf_{\theta \in \Theta_I} d(\theta, \tilde{\theta}) \leq \delta$. It follows that with w.p.a.1, $d_H(\tilde{\Theta}, \Theta_I) \leq \delta$. Since $\delta > 0$ is arbitrary, it follows that $d_H(\tilde{\Theta}, \Theta_I) \xrightarrow{p} 0$.*Q.E.D.*

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

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

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

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

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

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

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

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

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

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

Q.E.D.

14 Appendix B: Supplementary Results

We give here some supplementary results that may also be useful and are referred to in the paper.

14.1 Monotonicity for Nonparametric Models

When properties of g_0 are known it should be possible to tighten the bounds. An example is monotonicity, as imposed in the following condition.

ASSUMPTION B1: *For some $\tilde{x}$ and $\bar{x}$, $g_0(\tilde{x}, \alpha_i, \varepsilon_{it}) \geq g_0(\bar{x}, \alpha_i, \varepsilon_{it})$.*

This condition leads to tighter bounds for the ASF and QSF. The following result gives bounds for the static model.

THEOREM B1: *Suppose that Assumptions 1, 2, 4, and B1 are satisfied. If $E[|g_0(x, \alpha_i, \varepsilon_{it})|] < \infty$ for $x \in \{\tilde{x}, \bar{x}\}$ then*

$$\mu(\tilde{x}) - \mu(\bar{x}) \geq E[D_i\{\frac{\sum_{t=1}^T d_{it}(\tilde{x})Y_{it}}{T_i(\tilde{x})} - \frac{\sum_{t=1}^T d_{it}(\bar{x})Y_{it}}{T_i(\bar{x})}\}] = \delta E[D_i].$$

Also, if $G^*_u(y, \tilde{x})$ and $G^*_\ell(y, \bar{x})$ are continuous and strictly increasing on the interior of their range for $\tilde{1}_i = 1(T_i(\tilde{x}) > 0)$, $\bar{1}_i = 1(T_i(\bar{x}) > 0), \bar{\mathcal{P}}(\bar{x}, \tilde{x}) = E[(1 - \tilde{1}_i)(1 - \bar{1}_i)]$,

$$\begin{aligned} G^*_u(y, \tilde{x}) &= E[\tilde{1}_i \frac{\sum_{t=1}^T d_{it}(\tilde{x})1(Y_{it} \leq y)}{T_i(\tilde{x})} + (1 - \tilde{1}_i)\bar{1}_i \frac{\sum_{t=1}^T d_{it}(\bar{x})1(Y_{it} \leq y)}{T_i(\bar{x})}] \\ &\quad + \bar{\mathcal{P}}(\bar{x}, \tilde{x}), \\ G^*_\ell(y, \bar{x}) &= E[\bar{1}_i \frac{\sum_{t=1}^T d_{it}(\bar{x})1(Y_{it} \leq y)}{T_i(\bar{x})} + (1 - \bar{1}_i)\tilde{1}_i \frac{\sum_{t=1}^T d_{it}(\tilde{x})1(Y_{it} \leq y)}{T_i(\tilde{x})}], \end{aligned}$$

then $q(\lambda,\tilde{x}) \geq Q(\lambda, G_u^*(\cdot,\tilde{x}))$ *and* $q(\lambda,\bar{x}) \leq Q(\lambda, G_\ell^*(\cdot,\bar{x}))$, *so that*

$$q(\lambda,\tilde{x}) - q(\lambda,\bar{x}) \geq Q(\lambda, G_u^*(\cdot,\tilde{x})) - Q(\lambda, G_\ell^*(\cdot,\bar{x})).$$

PROOF: Let $\tilde{1}_i = 1(T_i(\tilde{x}) > 0)$ and $\bar{1}_i = 1(T_i(\bar{x}) > 0)$ and note that $1 = \tilde{1}_i + (1-\tilde{1}_i)\bar{1}_i + (1-\tilde{1}_i)(1-\bar{1}_i)$. Also let $\tilde{Y}_i = \tilde{1}_i \sum_{t=1}^T d_{it}(\tilde{x})Y_{it}/T_i(\tilde{x})$ and $\bar{Y}_i = \bar{1}_i \sum_{t=1}^T d_{it}(\bar{x})Y_{it}/T_i(\bar{x})$. It follows as in the proof of Theorem 4 that

$$E[\tilde{1}_i g_0(\tilde{x},\alpha_i,\varepsilon_{it})] = E[\tilde{1}_i\tilde{Y}_i], E[\bar{1}_i g_0(\bar{x},\alpha_i,\varepsilon_{it})] = E[\bar{1}_i\bar{Y}_i].$$

Then by monotonicity we have

$$\begin{aligned}
\mu(\tilde{x}) &= E[g_0(\tilde{x},\alpha_i,\varepsilon_{it})] \geq E[\{\tilde{1}_i + (1-\tilde{1}_i)(1-\bar{1}_i)\}g_0(\tilde{x},\alpha_i,\varepsilon_{it})] \\
&\quad + E[(1-\tilde{1}_i)\bar{1}_i g_0(\bar{x},\alpha_i,\varepsilon_{it})] \\
&= E[\tilde{1}_i\tilde{Y}_i + (1-\tilde{1}_i)\bar{1}_i\bar{Y}_i + (1-\tilde{1}_i)(1-\bar{1}_i)g_0(\tilde{x},\alpha_i,\varepsilon_{it})].
\end{aligned}$$

Similarly we have

$$\mu(\bar{x}) \leq E[\bar{1}_i\bar{Y}_i + (1-\bar{1}_i)\tilde{1}_i\tilde{Y}_i + (1-\tilde{1}_i)(1-\bar{1}_i)g_0(\bar{x},\alpha_i,\varepsilon_{it})].$$

Subtracting this inequality from the previous one, and noting that $\tilde{1}_i - (1-\bar{1}_i)\tilde{1}_i = \bar{1}_i\tilde{1}_i = D_i$ and $-\bar{1}_i + (1-\tilde{1}_i)\bar{1}_i = -D_i$, we have

$$\begin{aligned}
\mu(\tilde{x}) - \mu(\bar{x}) &\geq E[D_i(\tilde{Y}_i - \bar{Y}_i)] + E[(1-\tilde{1}_i)(1-\bar{1}_i)\{g_0(\tilde{x},\alpha_i,\varepsilon_{it}) - g_0(\bar{x},\alpha_i,\varepsilon_{it})\}] \\
&\geq E[D_i(\tilde{Y}_i - \bar{Y}_i)] = \delta E[D_i],
\end{aligned}$$

giving the first conclusion.

Next, similarly to above, for $\hat{G}_i(y,x) = 1(T_i(x) > 0)T_i(x)^{-1}\sum_{t=1}^T d_{it}(x)1(Y_{it} \leq y)$

$$\begin{aligned}
G(y,\tilde{x}) &= E[\{\tilde{1}_i + (1-\tilde{1}_i)(1-\bar{1}_i) + (1-\tilde{1}_i)\bar{1}_i\}1(g_0(\tilde{x},\alpha_i,\varepsilon_{it}) \leq y)] \\
&\leq E[\tilde{1}_i\hat{G}_i(y,\tilde{x})] + E[(1-\tilde{1}_i)\bar{1}_i\hat{G}_i(y,\bar{x})] + \bar{\mathcal{P}}(\bar{x},\tilde{x}) = G_u^*(y,\tilde{x}). \\
G(y,\bar{x}) &\geq G_\ell^*(y,\bar{x}).
\end{aligned}$$

Inverting gives the second conclusion. *Q.E.D.*

Turning now to the dynamic model, to sharpen the bounds for the monotonic case we use different partitions than in Section 5. Define $\mathcal{Y}_T(x) = \{X_i : X_{iT} = x\}$. The partition we use here to derive a lower bound for $\mu(\tilde{x})$ is

$$\{\mathcal{X}_t(\tilde{x}), t = 1,...,T; \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x}); \bar{\mathcal{X}}(\tilde{x}) \cap \bar{\mathcal{X}}(\bar{x}); \bar{\mathcal{X}}(\tilde{x}) \cap [\bar{\mathcal{X}}(\bar{x}) \cup \mathcal{Y}_T(\bar{x})]^c\},$$

where the superscript c denotes the complement of a set, i.e., $A^c = \{X : X \notin A\}$. The partition we use to derive an upper bound for $\mu(\bar{x})$ is the same with $\tilde{x}$ and $\bar{x}$ interchanged. For the QTE, we consider coarser partitions that do not include $\bar{\mathcal{X}}(\tilde{x}) \cap \bar{\mathcal{X}}(\bar{x})$ separately. The partition we use to derive a lower bound for $q(\lambda, \tilde{x})$ is

$$\{\mathcal{X}_t(\tilde{x}), t = 1, ..., T; \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x}); \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x})^c\}.$$

The partition we use to derive an upper bound for $q(\lambda, \bar{x})$ is the same with $\tilde{x}$ and $\bar{x}$ interchanged.

THEOREM B2: *Suppose that Assumptions 1, 3, 4, and B1 are satisfied. If* $B_\ell \leq g_0(x, \alpha_i, \varepsilon_{it}) \leq B_u$ *for* $x \in \{\tilde{x}, \bar{x}\}$ *then*

$$\begin{aligned}\mu(\tilde{x}) - \mu(\bar{x}) \geq \tilde{\delta} &+ E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x})\}Y_{iT}] - E[1\{X_i \in \bar{\mathcal{X}}(\bar{x}) \cap \mathcal{Y}_T(\tilde{x})\}Y_{iT}] \\ &+ E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap [\bar{\mathcal{X}}(\bar{x}) \cup \mathcal{Y}_T(\bar{x})]^c\}]B_\ell - E[1\{X_i \in \bar{\mathcal{X}}(\bar{x}) \cap [\bar{\mathcal{X}}(\tilde{x}) \cup \mathcal{Y}_T(\tilde{x})]^c\}]B_u,\end{aligned}$$

where $\tilde{\delta} = \sum_{t=1}^T E[(1\{X_i \in \mathcal{X}_t(\tilde{x})\} - 1\{X_i \in \mathcal{X}_t(\bar{x})\})Y_{it}]$. *If* $\tilde{G}^*_u(y, \tilde{x})$ *and* $\tilde{G}^*_\ell(y, \bar{x})$ *are continuous and strictly increasing on the interior of their range for*

$$\begin{aligned}\tilde{G}^*_u(y, \tilde{x}) &= \sum_{t=1}^T E[1\{X_i \in \mathcal{X}_t(\tilde{x})\}1\{Y_{it} \leq y\}] + E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x})\}1\{Y_{iT} \leq y\}] \\ &\quad + E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x})^c\}], \\ \tilde{G}^*_\ell(y, \bar{x}) &= \sum_{t=1}^T E[1\{X_i \in \mathcal{X}_t(\bar{x})\}1\{Y_{it} \leq y\}] + E[1\{X_i \in \bar{\mathcal{X}}(\bar{x}) \cap \mathcal{Y}_T(\tilde{x})\}1\{Y_{iT} \leq y\}],\end{aligned}$$

then $q(\lambda, \tilde{x}) \geq Q(\lambda, \tilde{G}^*_u(\cdot, \tilde{x}))$ *and* $q(\lambda, \bar{x}) \leq Q(\lambda, \tilde{G}^*_\ell(\cdot, \bar{x}))$, *so that*

$$q(\lambda, \tilde{x}) - q(\lambda, \bar{x}) \geq Q(\lambda, \tilde{G}^*_u(\cdot, \tilde{x})) - Q(\lambda, \tilde{G}^*_\ell(\cdot, \bar{x})).$$

PROOF: By monotonicity we have

$$\begin{aligned}E[g_0(\tilde{x}, \alpha_i, \varepsilon_{iT})] &\geq \sum_{t=1}^T E[1\{X_i \in \mathcal{X}_t(\tilde{x})\}Y_{it}] + E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x})\}Y_{iT}] \\ &\quad + E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap \bar{\mathcal{X}}(\bar{x})\}g_0(\tilde{x}, \alpha_i, \varepsilon_{iT})] + E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap [\bar{\mathcal{X}}(\bar{x}) \cup \mathcal{Y}_T(\bar{x})]^c\}]B_\ell.\end{aligned}$$

By the analogous equation with $\bar{x}$ and $\tilde{x}$ interchanged,

$$\begin{aligned}E[g_0(\bar{x}, \alpha_i, \varepsilon_{iT})] &\leq \sum_{t=1}^T E[1\{X_i \in \mathcal{X}_t(\bar{x})\}Y_{it}] + E[1\{X_i \in \bar{\mathcal{X}}(\bar{x}) \cap \mathcal{Y}_T(\tilde{x})\}Y_{iT}] \\ &\quad + E[1\{X_i \in \bar{\mathcal{X}}(\bar{x}) \cap \bar{\mathcal{X}}(\tilde{x})\}g_0(\bar{x}, \alpha_i, \varepsilon_{iT})] + E[1\{X_i \in \bar{\mathcal{X}}(\bar{x}) \cap [\bar{\mathcal{X}}(\tilde{x}) \cup \mathcal{Y}_T(\tilde{x})]^c\}]B_u.\end{aligned}$$

Subtracting these two inequalities and using monotonicity gives the first conclusion.

Also, by monotonicity it follows similarly to above that

$$\begin{aligned} G(y,\tilde{x}) &= \sum_{t=1}^{T} E[1\{X_i \in \mathcal{X}_t(\tilde{x})\}1\{g_0(\tilde{x},\alpha_i,\varepsilon_{iT}) \leq y\}] \\ &\quad + E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x})\}1\{g_0(\tilde{x},\alpha_i,\varepsilon_{iT}) \leq y\}] \\ &\quad + E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x})^c\}1\{g_0(\tilde{x},\alpha_i,\varepsilon_{iT}) \leq y\}] \leq \tilde{G}^*_u(y,\tilde{x}), \\ G(y,\bar{x}) &\geq \tilde{G}^*_\ell(\bar{x},y). \end{aligned}$$

The conclusion follows by inverting. Q.E.D.

If $X_{it} \in \{0,1\}$, the lower bound for $\mu(\tilde{x}) - \mu(\bar{x})$ simplifies to

$$\tilde{\delta} + E[1\{X_i \in \bar{\mathcal{X}}(\tilde{x}) \cap \mathcal{Y}_T(\bar{x})\}Y_{iT}] - E[1\{X_i \in \bar{\mathcal{X}}(\bar{x}) \cap \mathcal{Y}_T(\tilde{x})\}Y_{iT}],$$

which does not depend on B_ℓ and B_u. When the regressor takes on more than two values we can get tighter bounds if a monotonicity restriction holds for every possible pair of values. For example, if x were a scalar and $g_0(\tilde{x},\alpha_i,\epsilon_{it}) \geq g_0(\bar{x},\alpha_i,\epsilon_{it})$ for *every* $\tilde{x}$ and $\bar{x}$ with $\tilde{x} > \bar{x}$ then we could obtain improved bounds on the ASF and QSF.

14.2 Time Effects for Static Nonparametric Models

In static models it is possible to let g_0 depend on t through location and scale time effects. These effects can even be allowed to depend on x, though we focus here on the case where they do not.

ASSUMPTION B3: *There are vectors α_i and $\varepsilon_{it}, (t = 1, ..., T)$ of unobserved variables satisfying*

$$Y_{it} = g_{t0}(X_{it},\alpha_i,\varepsilon_{it}),\ g_{t0}(x,\alpha,\varepsilon) = \tau_t + s_t g_0(x,\alpha,\varepsilon),\ \tau_1 = 0,\ s_1 = 1.$$

Here τ_t and s_t are period specific location and scale effects. We impose the restrictions that $\tau_1 = 0$ and $s_1 = 1$, so that $g_{10} = g_0$. We also require that time homogeneity continues to hold as in Assumption 2. Now the ASF and QSF depend on t and are given by

$$\begin{aligned} \mu_t(x) &= \tau_t + s_t \int g_0(x,\alpha,\varepsilon)F(d\varepsilon,d\alpha), \qquad (12) \\ q_t(\lambda,x) &= \lambda^{th} \text{ quantile of } \tau_t + s_t g_0(x,\alpha_i,\varepsilon_{it}) \\ &= \tau_t + s_t \cdot \lambda^{th} \text{ quantile of } g_0(x,\alpha_i,\varepsilon_{it}). \end{aligned}$$

We use the fact that $E[g_0(x,\alpha_i,\varepsilon_{it})|X_i]$ does not depend on t to identify time effects. Different time periods with the same x provide identifying information. In particular, by eq. (1) for $t = 1$ and $t = t$,

$$\begin{aligned} d_{it}(x)d_{i1}(x)E[Y_{it}|X_i] &= d_{it}(x)d_{i1}(x)\{\tau_t + s_t E[g_0(x,\alpha_i,\varepsilon_{it})|X_i]\} \\ &= d_{it}(x)d_{i1}(x)\{\tau_t + s_t E[Y_{i1}|X_i]\}. \end{aligned}$$

It follows that

$$E[1(X_{it} = X_{i1})(Y_{it} - \tau_t - s_t Y_{i1})|X_i] = 0.$$

This is a conditional moment restriction that identifies τ_t and s_t as long as $E[Y_{i1}|X_i]$ varies over the set where $X_{it} = X_{i1}$. Bounds for the ASF and QSF for each t can then be formed by accounting for location and scale, as in the following result. Specifically, let Z_i be a dummy variable which is 1 for some X_{i1} values and zero for others.

THEOREM B4: *Suppose that Assumptions 2, 4, and B3 are satisfied,* $E[|Y_{it}|] < \infty$ *for all* t, *and for each* t, $\Pr(X_{it} = X_{i1}) > 0$ *and* $Var(E[Y_{i1}|X_i]|X_{it} = X_{i1}) > 0$. *Then there is a function* Z_{it} of X_i such that $Cov(Z_{it}, Y_{i1}|X_{it} = X_{i1}) \neq 0$ and

$$s_t = \frac{Cov(Z_{it}, Y_{it}|X_{it} = X_{i1})}{Cov(Z_{it}, Y_{i1}|X_{it} = X_{i1})}, \tau_t = E[Y_{it}|X_{it} = X_{i1}] - s_t E[Y_{i1}|X_{it} = X_{i1}], t = 2, ..., T.$$

If $B_\ell \leq g_0(x, \alpha_i, \varepsilon_{it}) \leq B_u$ *for constants* B_ℓ *and* B_u *and all* x, *then* $\mu_{t\ell}(x) \leq \mu_t(x) \leq \mu_{tu}(x)$ *where*

$$\begin{aligned}
\mu_{t\ell}(x) &= \tau_t + s_t E[1(T_i(x) > 0) T_i(x)^{-1} \sum_{t=1}^{T} d_{it}(x) \left(\frac{Y_{it} - \tau_t}{s_t} \right)] + s_t \bar{\mathcal{P}}(x) B_\ell, \\
\mu_{tu}(x) &= \mu_{t\ell}(x) + s_t \bar{\mathcal{P}}(x)(B_u - B_\ell).
\end{aligned}$$

Also if $G_\ell(y, x) = E[1(T_i(x) > 0) T_i(x)^{-1} \sum_{t=1}^{T} d_{it}(x) 1\left(\frac{Y_{it} - \tau_t}{s_t} \leq y\right)]$ *is continuous and strictly increasing on the interior of its range then* $q_{t\ell}(\lambda, x) \leq q_t(\lambda, x) \leq q_{tu}(\lambda, x)$ *where*

$$q_{t\ell}(\lambda, x) = \tau_t + s_t Q(\lambda, G_\ell(\cdot, x) + \bar{\mathcal{P}}(x)), q_{tu}(\lambda, x) = \tau_t + s_t Q(\lambda, G_\ell(\cdot, x)).$$

PROOF: By hypothesis $E[Y_{i1}|X_i]$ takes on more than one value when $X_{it} = X_{i1}$. Let $\tilde{d}_{it}$ denote a dummy variable that is equal to one when $E[Y_{i1}|X_i]$ takes on one of its distinct values and $Z_{it} = \tilde{d}_{it} - \Pr(\tilde{d}_{it} = 1|X_{it} = X_{i1})$. Note that Z_{it} is a function of X_i, so that by iterated expectations,

$$Cov(Z_{it}, Y_{i1}|X_{it} = X_{i1}) = E[Z_{it} Y_{i1}|X_{it} = X_{i1}] = E[Z_{it} E[Y_{i1}|X_i]|X_{it} = X_{i1}] \neq 0,$$

giving the first conclusion. The second conclusion follows by solving the usual population normal equations for instrumental variables conditional on $X_{it} = X_{i1}$.

Next, by Assumption B3 and Theorems 4 and 5 it follows that g_0 is bounded as in the conclusion of Theorems 4 and 5. The remainder of the proof follows by applying the location and scale transformation for each t. *Q.E.D.*

In general, there may be multiple instrumental variables Z_{it} that identify τ_t and s_t. For efficiency it would be desirable to estimate using all the avaiable instrumental variables and

optimal GMM . However, the small sample properties of this are likely to be poor because some data cells may have few observations, and so we focus on using a single instrumental variable.

The QSF bounds are unusual in that the quantile time effects are identified from expectations. This approach depends crucially on τ_t and s_t being constant (i.e. nonrandom). The ASF bounds will also apply when τ_t and s_t are random and independent of the data, but the QSF bounds will not.

14.3 Consistency of Fixed Effects Estimator of Identified Marginal Effect When $T = 2$.

In some models fixed effect (FE) estimators of the ATE appear to have small biases; e.g. see Hahn and Newey (2004) and Fernandez-Val (2009). Here we show consistency of FE for the ATE conditional on X_i values where the ATE is nonparametrically identified, in binary choice with binary regressors and $T = 2$. To describe this result, note that the FE estimator of the ASF conditional on $X_i = X^k$ is

$$\begin{aligned} \hat{\mu}_k^{FE}(x) &= \sum_{i=1}^{n} 1(X_i = X^k) H(x'\hat{\beta}_{FE} + \hat{\alpha}_i)/nP^k, \\ \hat{\beta}_{FE}, \hat{\alpha}_1, ..., \hat{\alpha}_n &= \arg \max_{\beta, \alpha_1, ..., \alpha_n} \sum_{i,t} \ln\{H(X_{it}'\beta + \alpha_i)^{Y_{it}}[1 - H(X_{it}'\beta + \alpha_i)]^{1-Y_{it}}\}. \end{aligned}$$

Let β_T denote the limit of $\hat{\beta}_{FE}$. In the multinomial choice model $\hat{\alpha}_i$ will have a limit distribution conditional on $X_i = X^k$ that is discrete with J support points $\alpha_j^k(\beta_T)$ and $\Pr(\alpha = \alpha_j^k(\beta_T)) = \mathcal{P}_j^k$, $(j = 1, ..., J)$. These limits will satisfy

$$\begin{aligned} \beta_T &= \text{argmax}_\beta \sum_{k=1}^{K} \mathcal{P}^k \sum_{j=1}^{J} \mathcal{P}_j^k \log \mathcal{L}_j^k \left(\alpha_j^k(\beta), \beta\right), \\ \alpha_j^k(\beta) &= \text{argmax}_\alpha \mathcal{L}_j^k(\alpha, \beta), (j = 1, ..., J; k = 1, ..., K). \end{aligned} \tag{13}$$

The corresponding limit of $\hat{\mu}_k^{FE}(x)$ is then given by

$$\mu_k^T(x) = \sum_{j=1}^{J} \mathcal{P}_j^k H(x'\beta_T + \alpha_j^k(\beta_T)).$$

As before with binary X_{it} and $T = 2$ we have $K = 4$. Let $X^1 = (0, 0)$, $X^2 = (0, 1)$, $X^3 = (1, 0)$, and $X^4 = (1, 1)$, so that the identified effect equals $\delta = \sum_{k=2}^{3} \mathcal{P}^k \Delta^k / \sum_{k=2}^{3} \mathcal{P}^k$.

THEOREM B5: *If* $H'(x) > 0$, $H(-x) = 1 - H(x)$, $X_{it} \in \{0, 1\}$, $T = 2$ *and* $\mathcal{P}_2 + \mathcal{P}_3 > 0$ *then*

$$\sum_{k=2}^{3} \mathcal{P}^k [\mu_k^T(1) - \mu_k^T(0)] / \sum_{k=2}^{3} \mathcal{P}^k = \delta.$$

Proof: Let $Y^1=(0,0)', Y^2=(0,1)', Y^3=(1,0)', Y^4=(1,1)'$ and $X^1=(0,0)'$, $X^2=(0,1)'$, $X^3=(1,0)'$, $X^4=(1,1)'$. The identified effect is

$$\begin{aligned}\delta &= \left\{\mathcal{P}^2 E[Y_{i2}-Y_{i1}|X_i=X^2]+\mathcal{P}^3 E[Y_{i1}-Y_{i2}|X_i=X^2]\right\}/(\mathcal{P}^2+\mathcal{P}^3)\\ &= \left[\mathcal{P}^2(\mathcal{P}_2^2-\mathcal{P}_3^2)+\mathcal{P}^3(\mathcal{P}_3^3-\mathcal{P}_2^3)\right]/(\mathcal{P}^2+\mathcal{P}^3).\end{aligned}$$

Next, the symmetry$H(-x)=1-H(x)$ implies that $\alpha_j^k(\beta)$ take the form

$$\alpha_j^k(\beta)=\begin{cases}-\infty, & j=1,\\ -\beta(X_1^k+X_2^k)/2, & j=2,3,\\ \infty, & j=4.\end{cases}$$

Note that for $k=2$ or $k=3$ we have $X_1^k+X_2^k=1$, so that $\alpha_j^k(\beta)=-\tilde{\beta}$ for $\tilde{\beta}=\beta/2$. Thus,

$$H(\beta+\alpha_j^k(\beta))-H(\alpha_j^k(\beta))=H(\tilde{\beta})-H(-\tilde{\beta})=2H(\tilde{\beta})-1.$$

Therefore the limit of the fixed effects estimator of the identified effect is

$$A[2H(\tilde{\beta})-1], A=\left[\mathcal{P}^2(\mathcal{P}_2^2+\mathcal{P}_3^2)+\mathcal{P}^3(\mathcal{P}_2^3+\mathcal{P}_3^3)\right]/(\mathcal{P}^2+\mathcal{P}^3).$$

Next, the limit of the concentrated log likelihood is

$$2\mathcal{P}^2[\mathcal{P}_2^2\ln H(\tilde{\beta})+\mathcal{P}_3^2\ln H(-\tilde{\beta})]+2\mathcal{P}^3[\mathcal{P}_2^3\ln H(-\tilde{\beta})+\mathcal{P}_3^3\ln H(\tilde{\beta})].$$

The first-order conditions for maximization of this object are

$$0=2\mathcal{P}^2[\mathcal{P}_2^2\lambda(\tilde{\beta})-\mathcal{P}_3^2\lambda(-\tilde{\beta})]+2\mathcal{P}^3[-\mathcal{P}_2^3\lambda(-\tilde{\beta})+\mathcal{P}_3^3\lambda(\tilde{\beta})],$$

where $\lambda(x)=H'(x)/H(x)$. By symmetry, $H'(-\tilde{\beta})=H'(\tilde{\beta})$. Divide the first order conditions by $H'(\tilde{\beta})$ and multiply by $H(\tilde{\beta})H(-\tilde{\beta})$ to obtain

$$\begin{aligned}0 &= 2\mathcal{P}^2[\mathcal{P}_2^2H(-\tilde{\beta})-\mathcal{P}_3^2H(\tilde{\beta})]+2\mathcal{P}^3[-\mathcal{P}_2^3H(\tilde{\beta})+\mathcal{P}_3^3H(-\tilde{\beta})]\\ &= 2(\mathcal{P}^2+\mathcal{P}^3)[\delta-A(2H(\tilde{\beta})-1)].Q.E.D.\end{aligned}$$

In numerical examples this same result continues to hold for $T=3$ and $T=4$. It would be interesting to extend this result to larger T but it is beyond the scope of this paper to do so. Unfortunately this result does not extend to the overall ATE.

References

[1] Altonji, J., and R. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors", *Econometrica* 73, 1053-1102.

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

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

[6] Bester, A.C., and C. Hansen (2008), "Flexible correlated random effects estimation in panel models with unobserved heterogeneity," working paper, GSB, Univ. of Chicago.

[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] 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.

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

[10] 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.

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

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

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

[14] Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34, 305-334.

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

[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] Fernandez-Val, I. (2009), "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models," *Journal of Econometrics* 150(1), pp. 71-85.

[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] 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.

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

[29] 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.

[30] Honore, B.E. (1992): "Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica* 60, 533-565

[31] Honore, 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] Imbens, G. and W.K. Newey (2009), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica* 77, 1481-1512.

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

[35] Lehmann, E. L. (1974): Nonparametrics: Statistical Methods Based on Ranks. San Francisco, CA: Holden-Day.

[36] Lindsay, B.G. (1983), "The Geometry of Mixture Likelihoods: A General Theory," *Annals of Statistics* 11, 86-94.

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

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

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

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

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

[42] Vella, F. and M. Verbeek (1998), "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," *Journal of Applied Econometrics,* 13, 163-183.

[43] 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.

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

[45] Yitzhaki, S. (1996) "On Using Linear Regressions in Welfare Economics," *Journal of Business & Economic Statistics* 14, 478-486.

[46] Yu, K. and M.C. Jones (1998): "Local Linear Quantile Regression," *Journal of the American Statistical Association* 93, 228-237.

Table 1: Biases of linear probability model estimator

T	$100(\delta_w-\Delta)/\Delta$			$100(\delta-\Delta)/\Delta$			$Pr(X_i = (0,\ldots,0)) + Pr(X_i = (1,\ldots,1))$		
	U	T	N	U	T	N	U	T	N
				A. $Pr(X_{it} = 1) = 0.10$					
4	-20	-33	-40	-19	-32	-38	0.70	0.74	0.75
8	-20	-33	-40	-16	-26	-32	0.54	0.61	0.64
16	-20	-33	-40	-11	-18	-23	0.38	0.48	0.52
				B. $Pr(X_{it} = 1) = 0.25$					
4	-10	-13	-12	-10	-12	-12	0.43	0.49	0.53
8	-10	-12	-12	-6	-7	-7	0.24	0.32	0.37
16	-10	-13	-12	-3	-4	-3	0.12	0.20	0.25
				C. $Pr(X_{it} = 1) = 0.50$					
4	2	9	14	2	8	13	0.26	0.35	0.40
8	2	8	14	1	5	10	0.08	0.17	0.22
16	2	8	14	0	3	7	0.01	0.07	0.12
				D. $Pr(X_{it} = 1) = 0.75$					
4	12	18	24	11	18	23	0.43	0.49	0.53
8	11	18	24	8	15	20	0.24	0.32	0.37
16	11	18	24	5	12	17	0.12	0.20	0.25
				E. $Pr(X_{it} = 1) = 0.90$					
4	16	17	14	16	17	14	0.70	0.73	0.75
8	16	16	13	15	18	18	0.54	0.61	0.63
16	16	16	14	13	19	21	0.38	0.48	0.52

Notes: probit model with $Y_{it} = 1(X_{it} + \alpha_i > \varepsilon_{it})$, $X_{it} = 1(\xi + \alpha_i > \eta_{it})$, $\varepsilon_{it} \sim N(0,1)$, and $\eta_{it} \sim N(0,1)$. Three distributions for α_i: uniform(-1,1) (U), triangular(-2,0,2) (T), and normal(0,1) (N). The value of ξ is calibrated to obtain the values of $Pr(X_{it} = 1)$ shown in the table. δ_w is the probability limit of the linear fixed effects estimator with constant slopes, δ is the probability limit of the average of the linear fixed effects estimators with individual specific slopes, and Δ is the ATE. The probabilities $Pr(Y_{it} = 1)$ are about 0.51, 0.55, 0.62, 0.70 and 0.75 in panels A, B, C, D, and E, respectively. Probability limits simulated numerically with random samples of 500,000 individuals.

Table 2: Empirical probabilities of union sequences

	Full sample		Ever unionized
	Never unionized	Always unionized	Always unionized
T = 2	0.69	0.13	0.42
T = 4	0.61	0.08	0.22
T = 6	0.56	0.07	0.16
T = 8	0.53	0.06	0.13

Source: NLSY79 1986-1993, 2,065 men. All the panels start in 1986

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

	Nonparametric model	Semiparametric model							Linear model
		Logit	FE-Logit	BC-Logit	CMLE	Probit	FE-Probit	BC-Probit	
					T = 2				
β*		-.36	-.78	-.36	-.39	[-.411, -.409]	-.88	-.51	
(95% N)			(-1.11, -.46)	(-.67, -.05)	(-.70, -.08)		(-1.24, -.52)	(-.86, -.16)	
(95% CP)		(-.75, .02)				(-.85, .03)			
(95% MP+)		(-.85, .02)				(-.88, .04)			
(95% PB^)		(-.88, .08)				(-1.06, .10)			
ATE	[-.49, -.02]	[-.06, -.05]	-.06	-.04		[-.07, -.05]	-.06	-.05	-.07
(95% N)	(-.53, .00)		(-.08, -.04)	(-.06, -.02)			(-.08, -.04)	(-.07, -.02)	(-.11, -.03)
(95% B*)	(-.52, -.01)								
(95% CP)		(-.15, .00)				(-.17, .00)			
(95% MP+)		(-.17, .00)				(-.18, .01)			
(95% PB^)		(-.19, .01)				(-.19, .02)			
					T = 3				
β*		-.42	-.71	-.46	-.46	[-.462, -.460]	-.78	-.55	
(95% N)			(-.90, -.52)	(-.64, -.28)	(-.65, -.28)		(-.99, -.57)	(-.75, -.35)	
(95% CP)		(-)				(-)			
(95% MP+)		(-.76, -.07)				(-.74, -.17)			
(95% PB^)		(-.74, -.12)				(-.73, -.16)			
ATE	[-.40, -.04]	[-.07, -.07]	-.08	-.07		[-.08, -.07]	-.08	-.07	-.08
(95% N)	(-.46, .00)		(-.09, -.06)	(-.09, -.05)			(-.09, -.06)	(-.09, -.05)	(-.11, -.06)
(95% B*)	(-.41, -.02)								
(95% CP)		(-)				(-)			
(95% MP+)		(-.13, -.01)				(-.14, -.03)			
(95% PB^)		(-.13, -.02)				(-.14, -.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, 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); BC denotes bias corrected FEMLE; CMLE denotes conditional logit FEMLE; Linear 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.

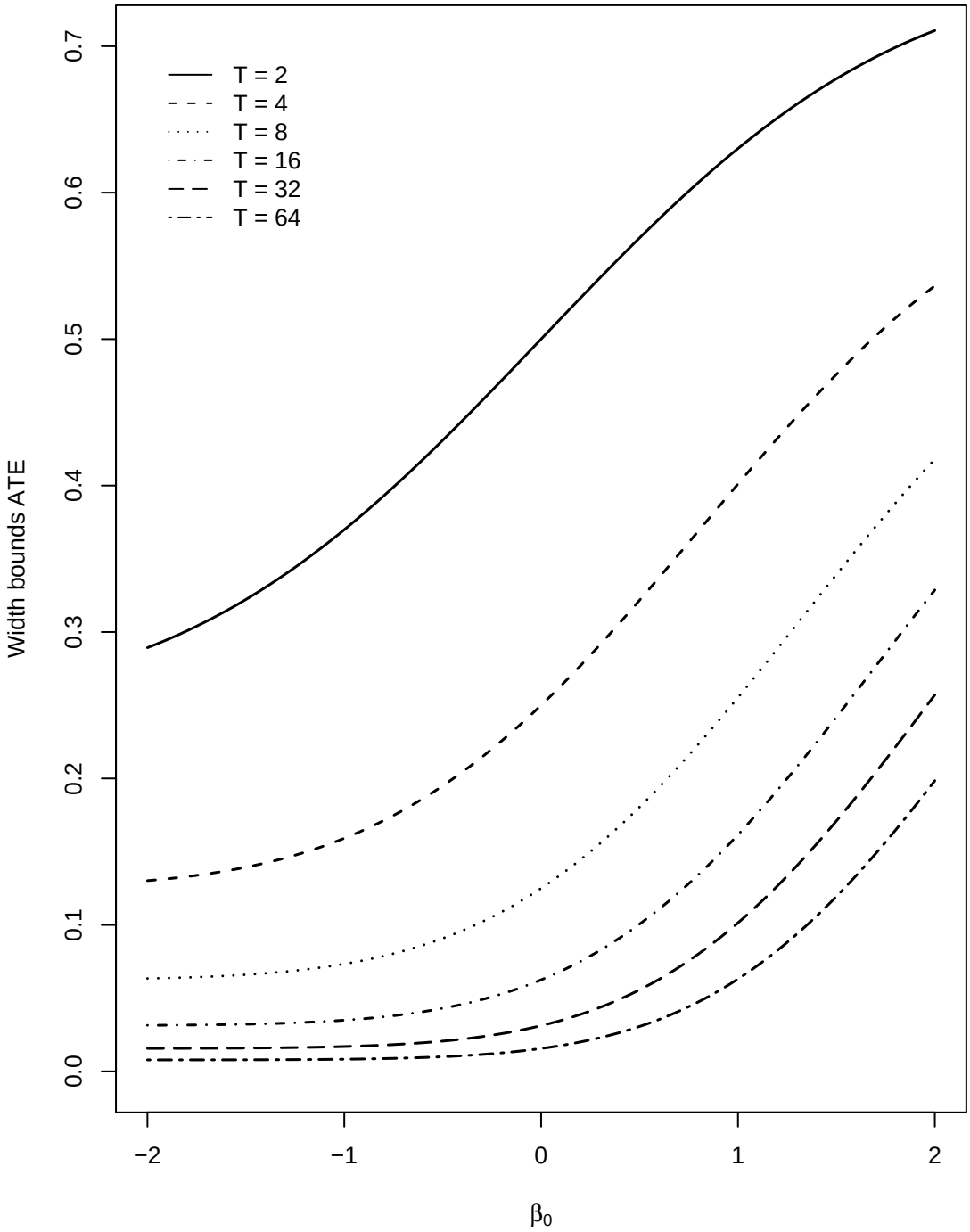

Figure 1: Width of nonparametric bounds for the ATE in dynamic binary choice probit models with $Y_{it} = 1(\beta_0 Y_{i,t-1} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\Pr(Y_{i0} = 1) = .5$, $\beta_0 \in [-2, 2]$, and $T \in \{2, 4, 8, 16, 32, 64\}$.

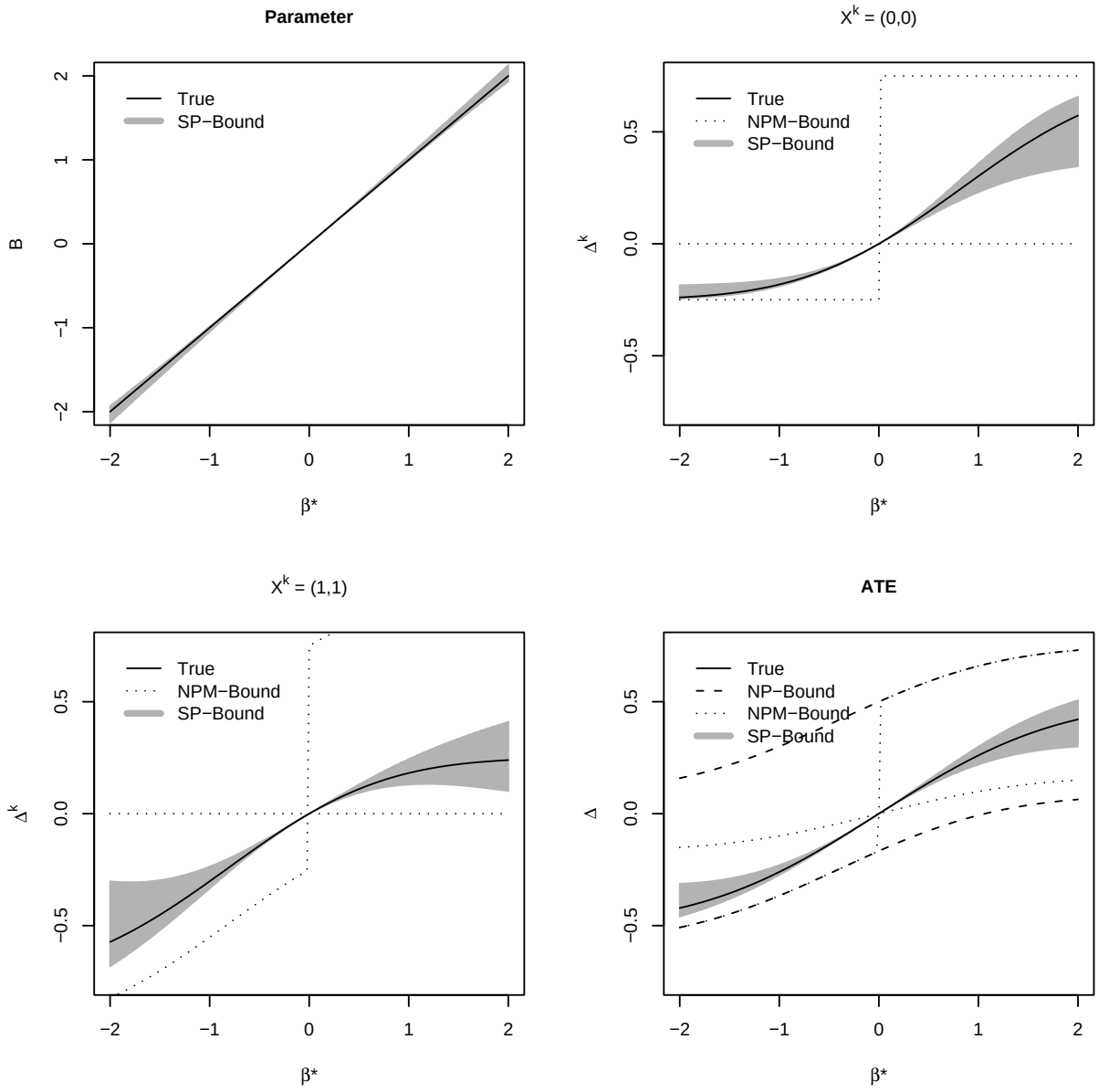

Figure 2: Identified set for parameter and ATEs in binary choice probit models with $Y_{it} = 1(\beta^* X_{it} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim N(0,1)$, $X_{it} = 1(\alpha_i \geq \eta_{it})$, $\eta_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\beta^* \in [-2,2]$, and $T = 2$.

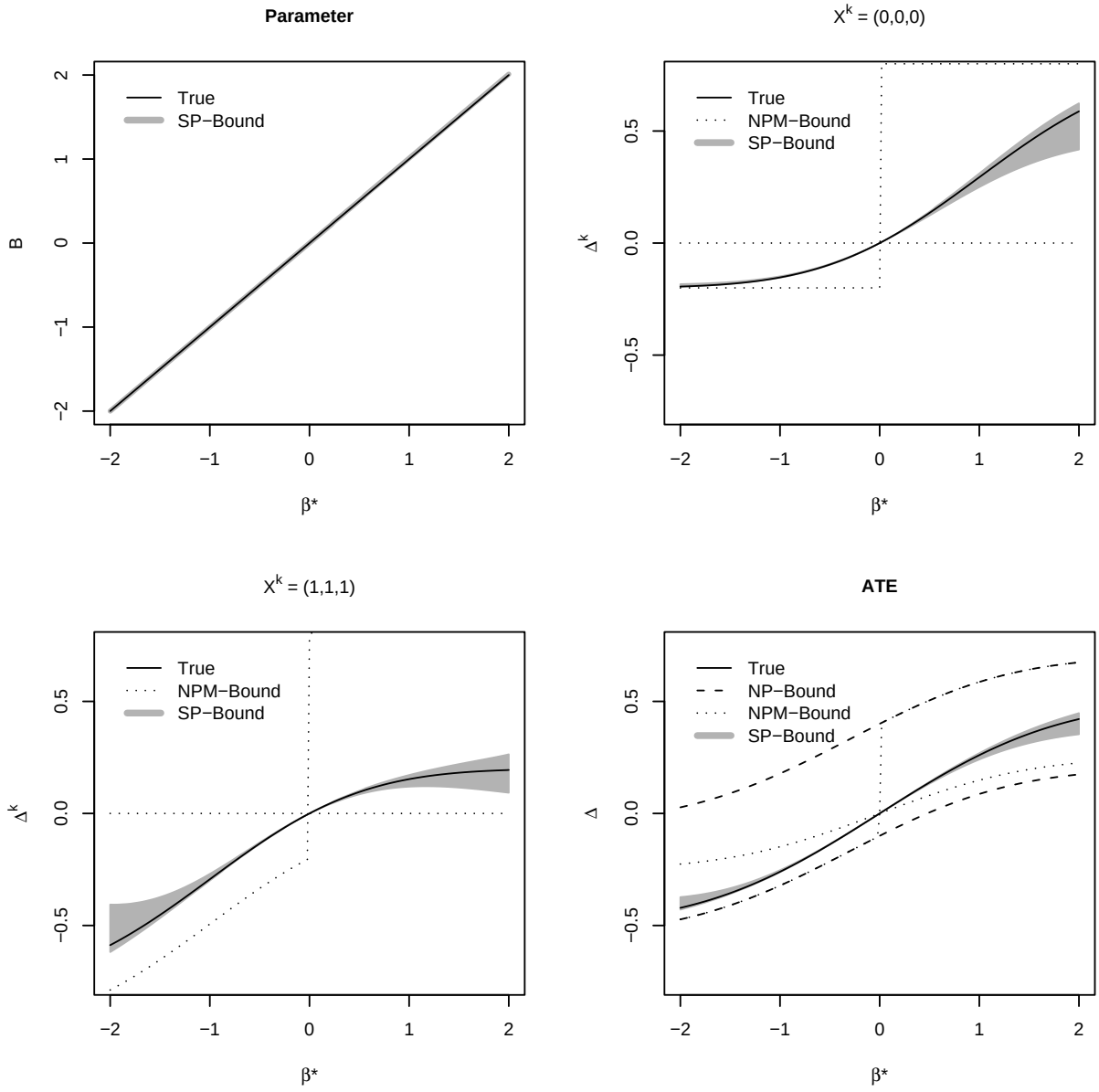

Figure 3: Identified set for parameter and ATEs in binary choice probit models with $Y_{it} = 1(\beta^* X_{it} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim N(0,1)$, $X_{it} = 1(\alpha_i \geq \eta_{it})$, $\eta_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\beta^* \in [-2,2]$, and $T = 3$.

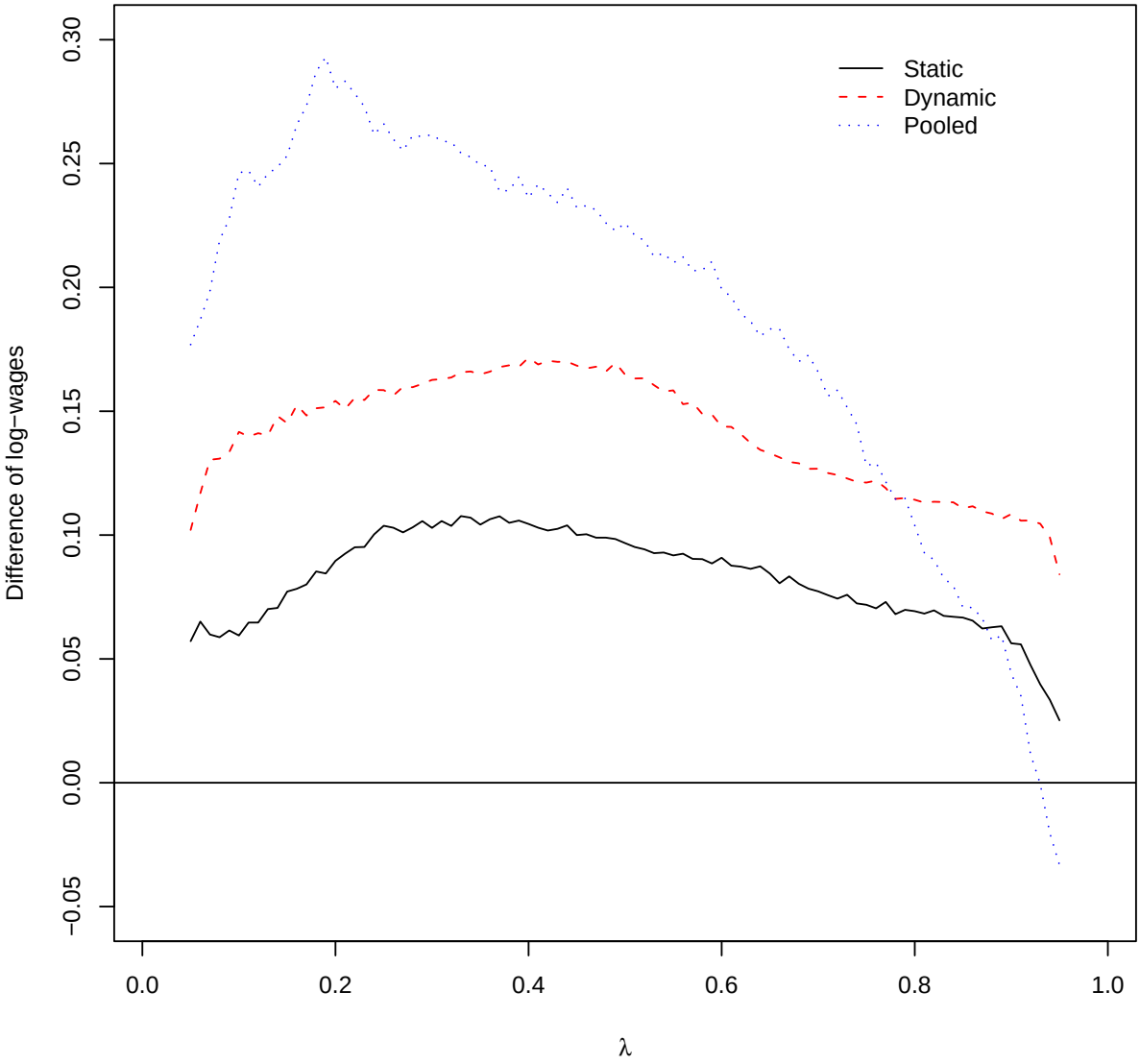

Figure 4: Identified quantile union effect. Estimates based on the entire panel 1986–1993.

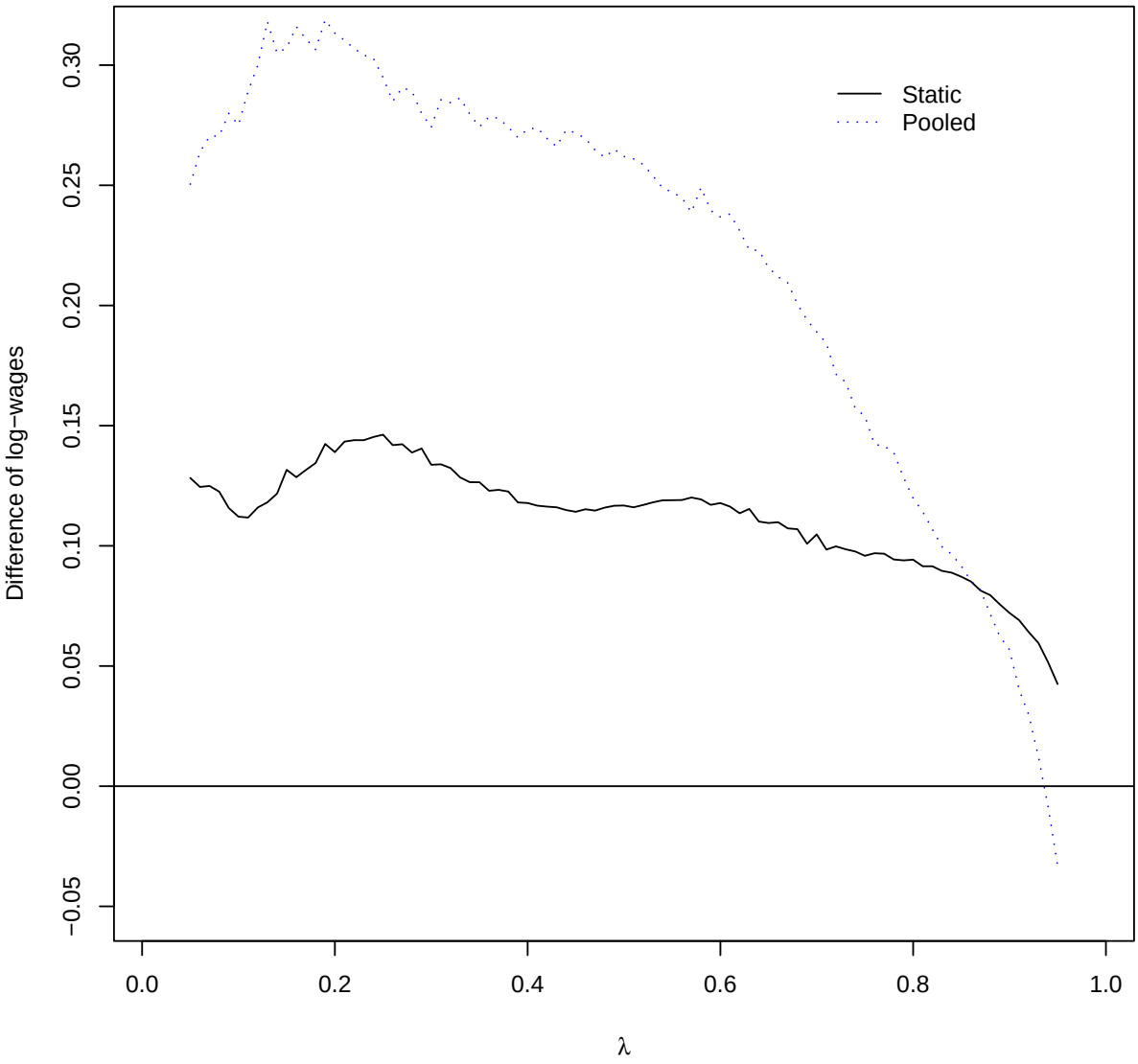

Figure 5: Identified quantile union effect with location and scale time effects. Estimates based on the entire panel 1986–1993.

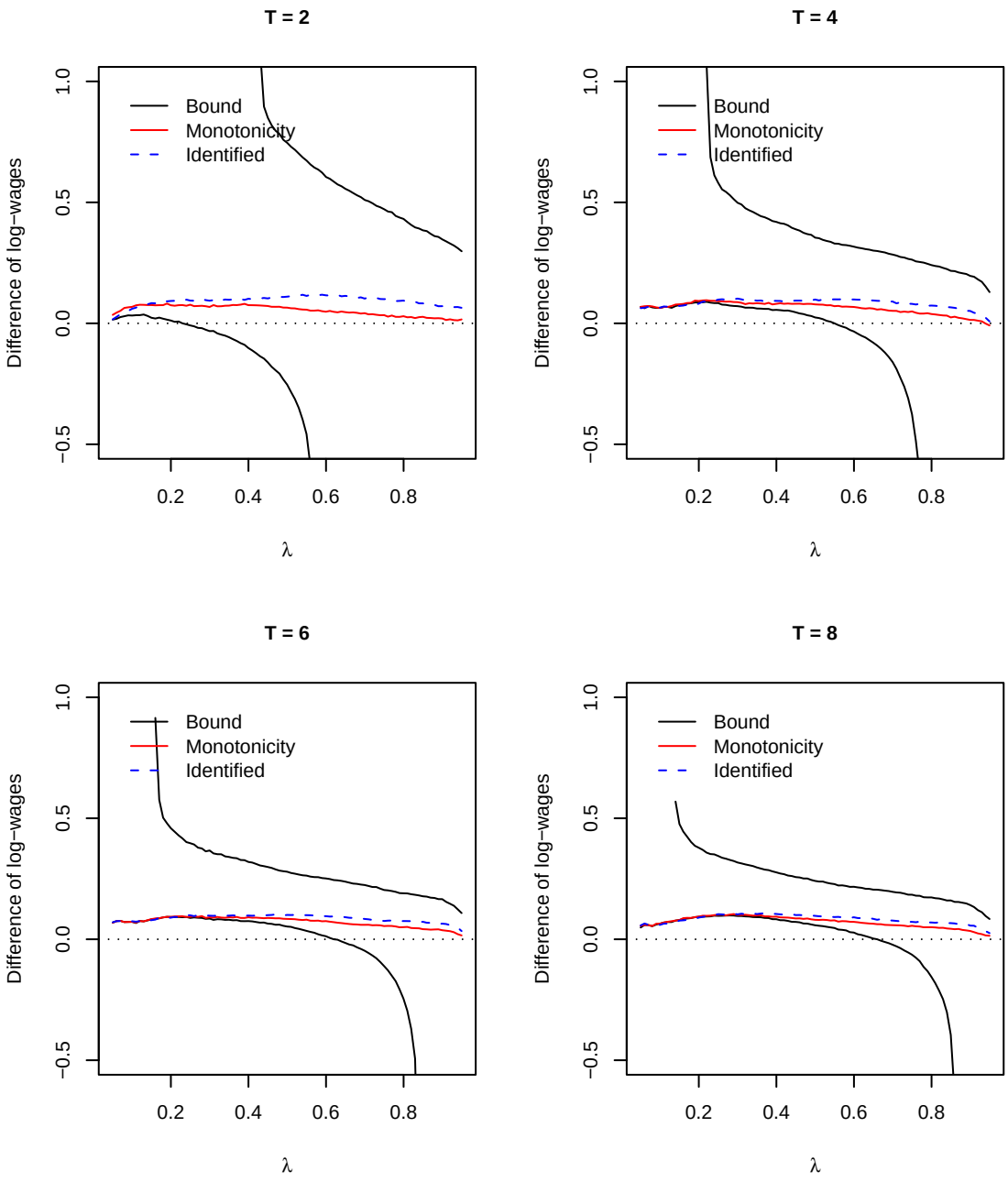

Figure 6: Bounds for quantile union effect on ever unionized. Static model.

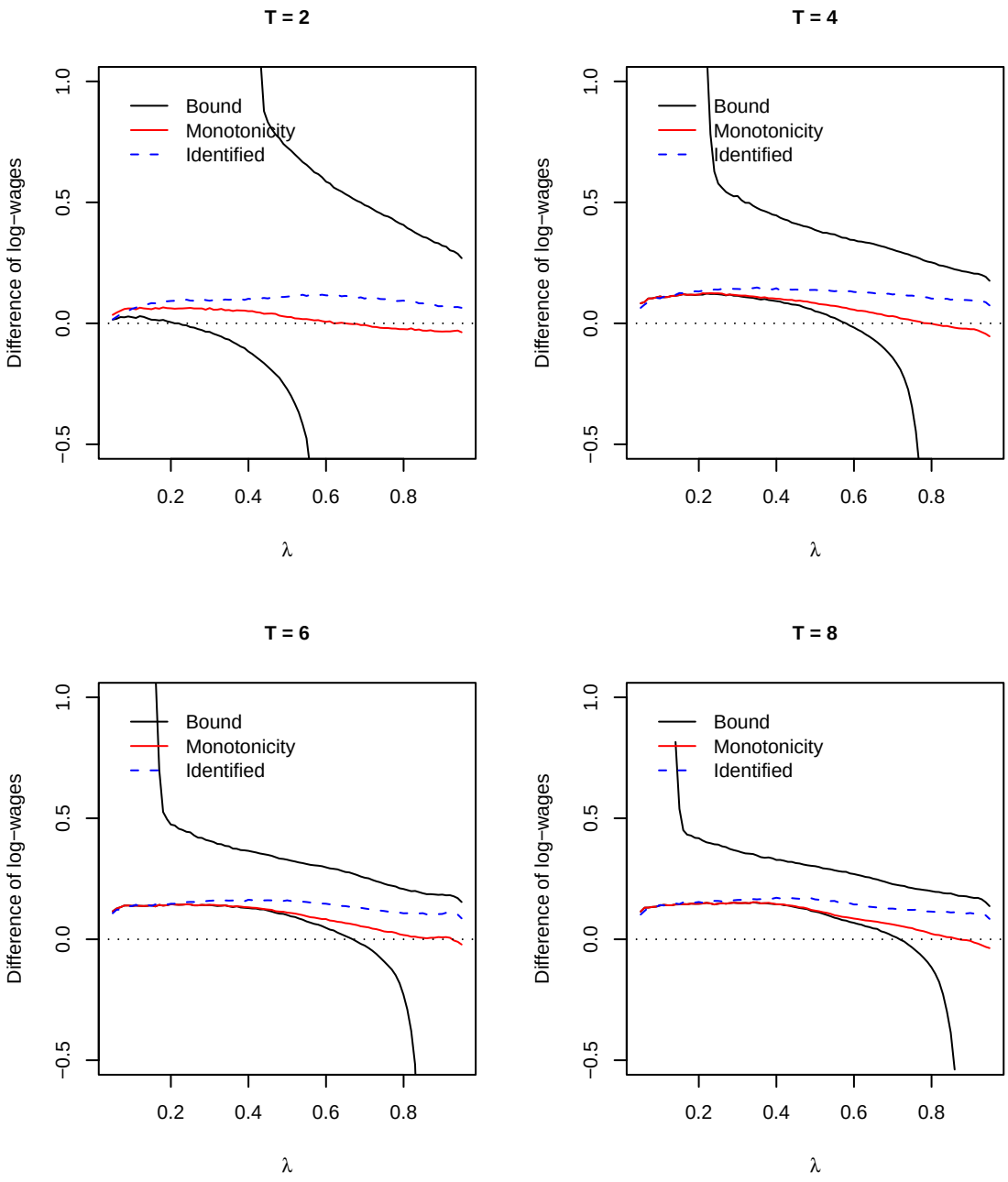

Figure 7: Bounds for quantile union effect on ever unionized. Dynamic model.

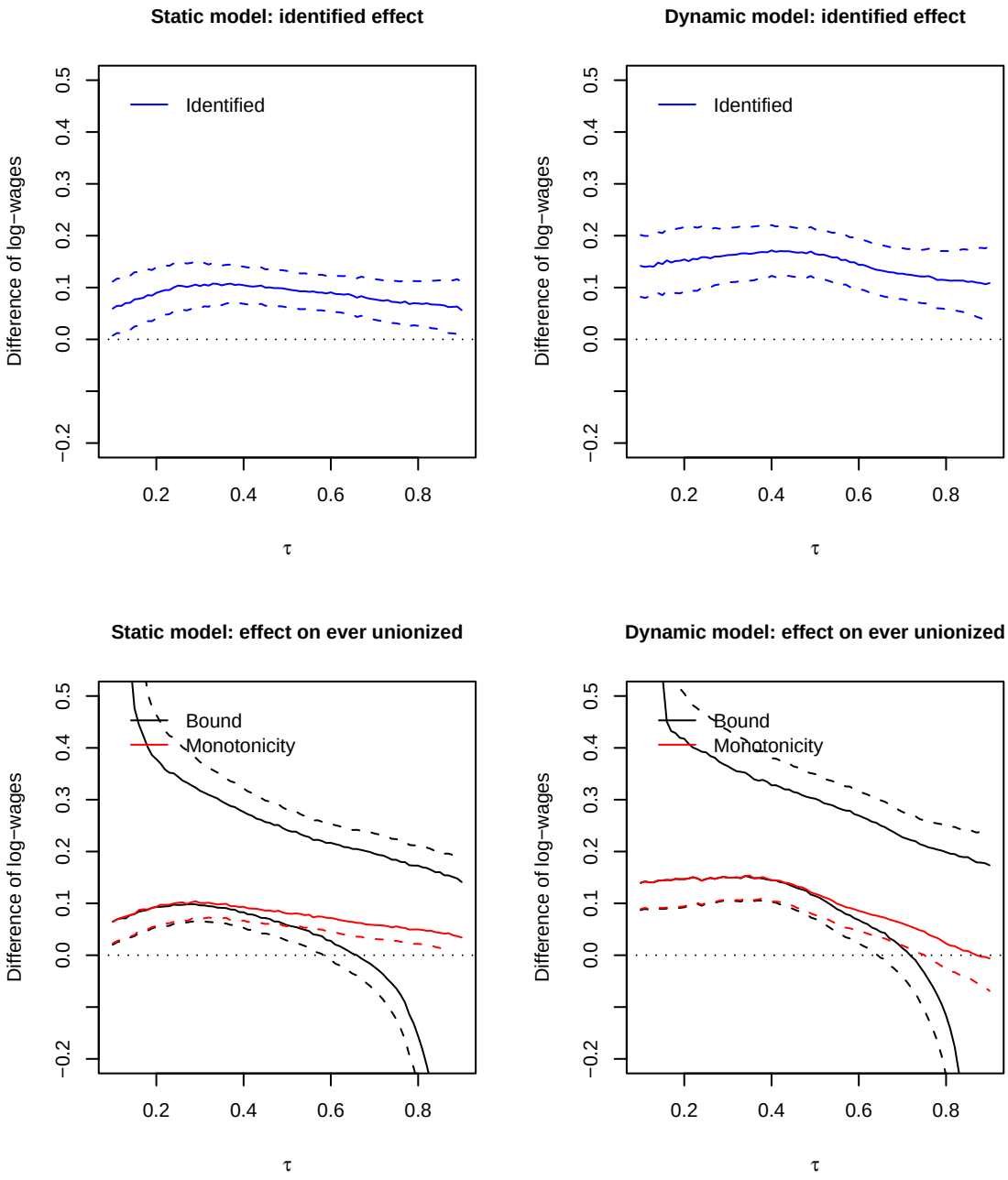

Figure 8: 90% bootstrap uniform confidence bands for the identified union effect and union effect on ever unionized (dashed lines). Estimates based on the entire panel 1986–1993.

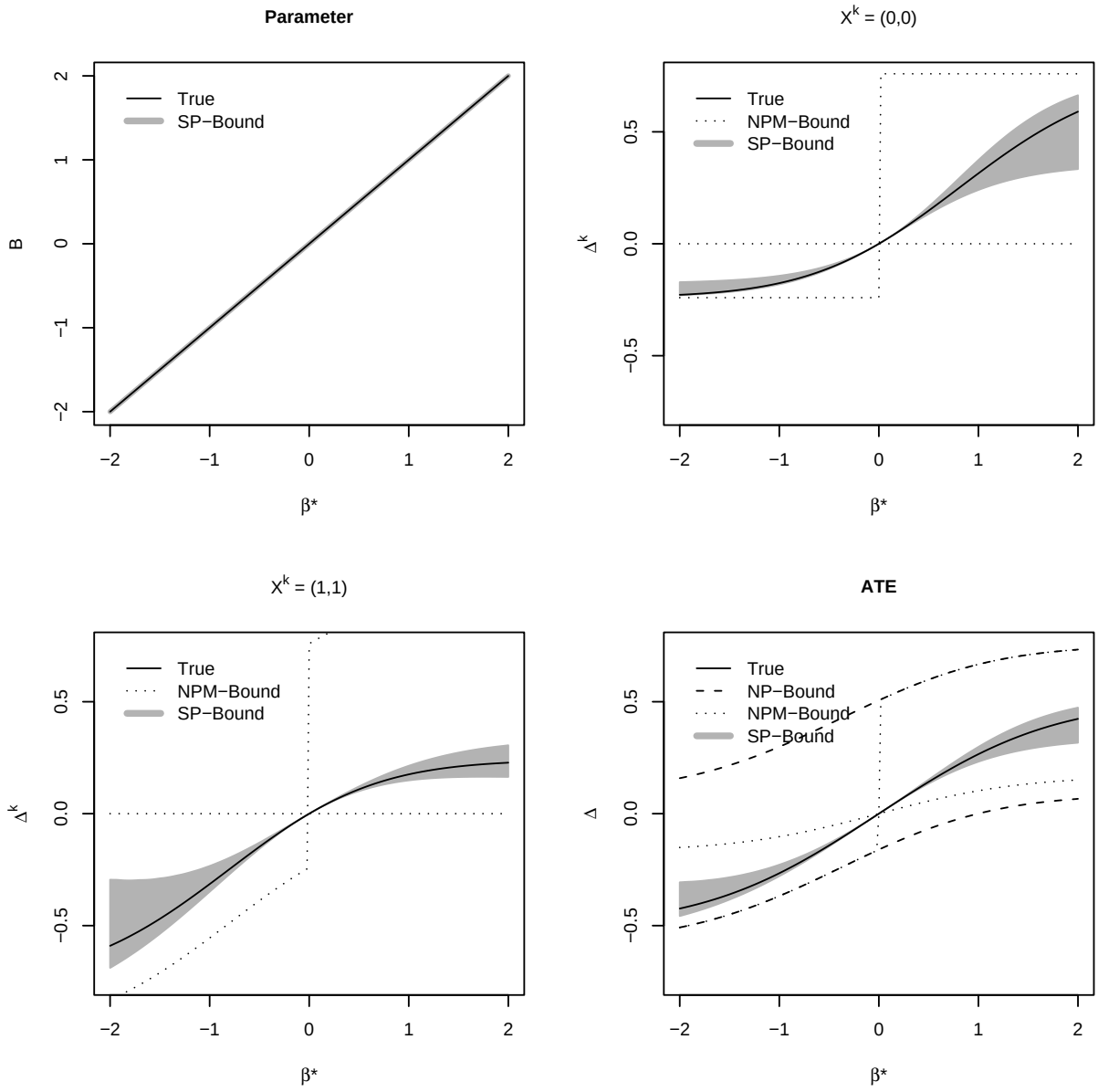

Figure 9: Identified set for parameter and ATEs in binary choice logit models with $Y_{it} = 1(\beta^* X_{it} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim L(0,1)$, $X_{it} = 1(\alpha_i \geq \eta_{it})$, $\eta_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\beta^* \in [-2,2]$, and $T = 2$.

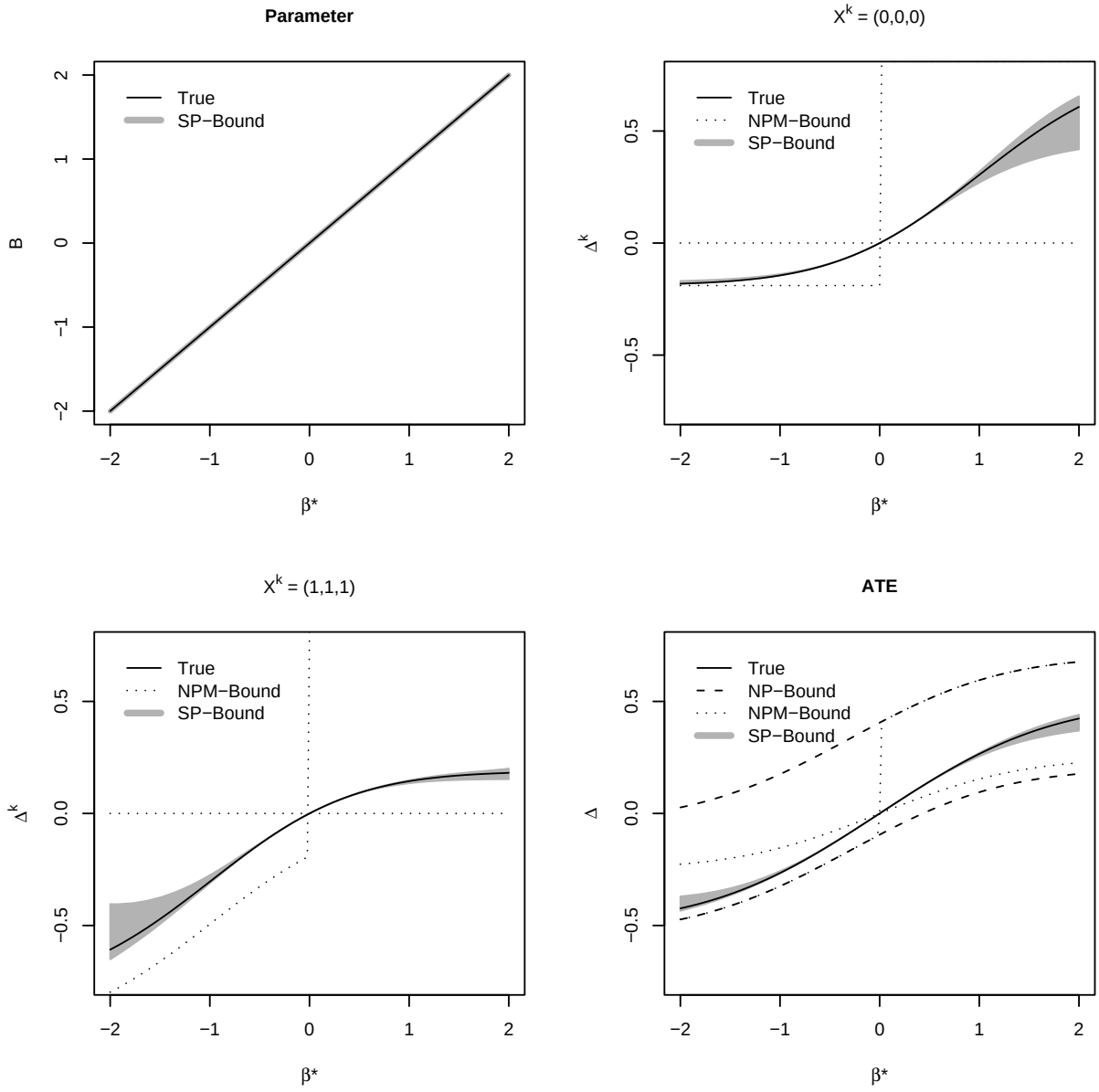

Figure 10: Identified set for parameter and ATEs in binary choice logit models with $Y_{it} = 1(\beta^* X_{it} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim L(0,1)$, $X_{it} = 1(\alpha_i \geq \eta_{it})$, $\eta_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\beta^* \in [-2,2]$, and $T = 3$.