Average and Quantile Effects in Nonseparable Panel Models

Average and Quantile Effects in Nonseparable Panel Models[1]

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

March 27, 2013

[1]We thank J. Angrist, G. Chamberlain, D. Chetverikov, B. Frandsen, B. Graham, J. Hausman, and many seminar participants for comments. Brad Larsen and Seongyeon Chang 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 in between. We gratefully acknowledge research support from the NSF.

Abstract

Nonseparable panel models are important in a variety of economic settings, including discrete choice. This paper gives identification and estimation results for nonseparable models under time homogeneity conditions that are like "time is randomly assigned" or "time is an instrument." Partial identification results for average and quantile effects are given for discrete regressors, under static or dynamic conditions, in fully nonparametric and in semiparametric models, with time effects. It is shown that the usual, linear, fixed-effects estimator is not a consistent estimator of the identified average effect, and a consistent estimator is given. A simple estimator of identified quantile treatment effects is given, providing a solution to the important problem of estimating quantile treatment effects from panel data. Bounds for overall effects in static and dynamic models are given. The dynamic bounds provide a partial identification solution to the important problem of estimating the effect of state dependence in the presence of unobserved heterogeneity. The impact of T, the number of time periods, is shown by deriving shrinkage rates for the identified set as T grows. We also consider semiparametric, discrete-choice models and find that semiparametric panel bounds can be much tighter than nonparametric bounds. Computationally-convenient methods for semiparametric models are presented. We propose a novel inference method that applies in panel data and other settings and show that it produces uniformly valid confidence regions in large samples. We give empirical illustrations.

1 Introduction

Interesting empirical questions are often formulated in terms 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 binary-choice logit model with an individual location effect. However these are rare and dependent on specific models or distributions. For example, the slope parameter of the binary-choice model with a time dummy is identified only for logit as shown by Chamberlain (2010), and the average treatment effect is not identified even for logit without a time dummy, as shown below.

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 condition is like "time being randomly assigned" or "time is an instrument." 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 nonseparable models, i.e. for models that are not additively separable in unobserved factors. We allow for multidimensional heterogeneity, as motivated by models where effects of interest, such as price and income elasticities, are distributed among individuals in unrestricted ways; see Altonji and Matzkin (2005), Browning and Carro (2007), and Fernandez-Val and Lee (2010), among others. We also weaken the strict time homogeneity conditions to allow 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 many effects, though some conditional effects are identified. This paper considers partial identification and estimation of average and quantile effects, under static or dynamic conditions, in fully nonparametric and in semiparametric models, with time effects.

For the nonparametric, static model we give simple estimators of the identified average effect of x on y, conditional on x varying over time. These estimators extend Chamberlain (1982, pp. 10-17) to multiple regressors with location and scale time effects. We also find that linear, fixed-effects estimate a variance-weighted average effect instead of the average effect. For bounded y we move beyond the analysis of identified effects and give simple estimators of sharp bounds for average effects. These bounds provide nonparametric, partial-identification estimates of average effects in important cases, such as binary choice in panel data.

The quantile estimators given here are more novel than the average-effect estimators. They provide simple estimators of the effect of x on quantiles of y, conditional on x varying over time, that allow for location and scale time effects. Estimators of sharp bounds are also provided for the unconditional, overall quantile effect. The estimators allow for multidimensional heterogeneity, for example for both location and slope to vary across individuals in an unrestricted way. In this way we provide a solution for the important problem of nonparametric quantile regression in panel data with individual effects, for discrete regressors. Graham, Hahn and Powell (2009) also consider quantile effects in linear, heterogenous coefficients models, but impose conditions which essentially restrict the heterogeneity to be one-dimensional, and focus on identification of the distribution of coefficients.

Dynamics is often an important feature of economic models with intertemporal choice. Here we give a dynamic, nonseparable, panel model that nests the static one. Simple estimators of bounds on average and quantile effects are provided. We show that these results provide a partial-identification solution to the important problem of distinguishing state dependence from heterogeneity.

This paper shows the impact of the number of time periods T on identification. We find that the identified set of effects shrinks to a point exponentially quickly as T grows, when individual effects are bounded and time period disturbances are not, and that the rate is some power of T^{-1} more generally. In a nonparametric, dynamic, binary-choice model we find that the rate is faster the larger the variance of the period-specific disturbance relative to the variance of the individual effect.

In numerical examples we find that the nonparametric bounds can be quite wide, motivating more informative models. Semiparametric models that specify the distribution of the outcome given regressors and individual effect is an important class of more informative models. Here we describe both static and dynamic semiparametric models. When restrictions are imposed on the heterogeneity, like only some coefficients varying across individuals, semiparametric models can have substantially tighter bounds than nonparametric models. We find that in the important binary-logit model with just a location effect the average effect bounds shrink exponentially quickly as T grows, in both dynamic and static models, even when the nonparametric bounds shrink slowly. This result quantifies the gain in information of a semiparametric model with just a location effect over the nonparametric model. We also find quite tight bounds for semiparametric models relative to nonparametric ones in numerical examples.

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 (2003, 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 the *perturbed bootstrap*, that is described in the paper, and another called *modified projection,* that is described in the Supplementary Material. These methods produce uniformly valid inference in large samples and may be of substantial independent interest.

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

Recent research has considered nonseparable panel models with time homogeneity and continuous regressors. Graham and Powell (2011) give estimators of the average effect in a linear model with heterogeneous slopes. Hoderlein and White (2011) give estimators of the average derivative conditional on equality of regressors across time periods.

Chamberlain (1980, 1984), Altonji and Matzkin (2005), Bester and Hansen (2008), and others have used control functions for panel data estimation. We focus instead on time homogeneity with unrestricted dependence between individual effects and regressors. 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 depend on large T for consistency while we estimate identified effects and bounds for fixed T.

Section 2 describes the models and effects we consider. Section 3 discusses estimation of identified effects. Sections 4 and 5 derive bounds for the static and dynamic nonparametric models respectively. Section 6 describes the impact of T. Section 7 describes and gives results for semiparametric, discrete-choice models. Section 8 gives computationally convenient methods for semiparametric models and numerical examples. Section 9 considers estimation and inference for semiparametric models. Section 10 gives empirical examples. The Supplementary Material Chernozhukov et. al. (2012) includes a variety of omitted discussions and results along with the proofs of results stated in the paper.

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 ε_{it} 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 model.

Assumption 2: $\varepsilon_{it}|X_i, \alpha_i \overset{d}{=} \varepsilon_{i1}|X_i, \alpha_i$, *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 $\tilde{\varepsilon}_{it}|X_i \overset{d}{=} \tilde{\varepsilon}_{i1}|X_i$ 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 "time is randomly assigned" or "time is an instrument" 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. Assumption 2 also turns out to be a natural strengthening of linear model conditions, as shown in Theorem A1 and the associated discussion in the Supplementary Material.

A dynamic 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 that is nested within the static model of Assumptions 1 and 2, as shown in Theorem A2 of the Supplementary Material. Here the conditional distribution given only current and lagged regressors must be time invariant. It also implies that the conditional distribution of ε_{it} given current and lagged regressors only depends on X_{i1}. Here ε_{it} can be thought of as additional information that is independent of the past regressors. A conditional-mean version of this condition arises in rational-expectations models that implies disturbances have mean zero conditional on past information. Here the stronger conditional independence restriction is imposed as seems needed for a nonseparable model. The conditioning on X_{i1} is a way to account for the initial conditions of this dynamic model. Bhargava and Sargan (1983) adopted this approach in a linear model as have Honore and Tamer (2006) and Browning and Carro (2007) in a likelihood setting.

If X_{it} includes lagged Y_{it} then Assumption 3 specifies that the model is "dynamically complete," ruling out $Y_{it} = g_0(X_{it}, \alpha_i, \varepsilon_{it})$ as one equation of a dynamic system. 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 $Y_{it} \in \{0,1\}$ is binary, representing state dependence, with α_i representing unobserved heterogeneity. This example is treated in Section 5.

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

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

where throughout the paper F denotes the cumulative distribution function (CDF) of a random vector that appears as the arguments of F. 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 x^b (before) to x^a (after) is

$$\Delta = \mu(x^a) - \mu(x^b).$$

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 CDF of $g_0(x, \alpha_i, \varepsilon_{it})$,

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

In the treatment-effects literature the λ^{th} quantile treatment effect (QTE) of changing x from x^b to x^a is

$$\Delta_\lambda = q(\lambda, x^a) - q(\lambda, x^b),$$

as in Lehmann (1974). This effect does not give the quantile of the treatment effect but does quantify the shift in the distribution of Y_{it} that is due to a change in x. It accounts for multidimensional individual effects that may be correlated with x.

The static model implies a conditional-mean model that has been considered by Chamberlain (1982), Hahn (2001), Wooldridge (2005), and Chernozhukov et. al. (2007). This conditional-mean model specifies that there is an α_i and $m_0(x, \alpha)$ such that $E[Y_{it}|X_i, \alpha_i] = m_0(X_{it}, \alpha_i)$. A conditional mean ATE, as in Wooldridge (2005), is $\int [m_0(x^a, \alpha) - m_0(x^b, \alpha)] dF(\alpha)$. This model and effect differ from those we consider in specifying conditional-mean restrictions, while we specify conditional distribution restrictions. In Theorem A3 of the Supplementary Material we show that the conditional-mean model is implied by Assumptions 1 and 2, or 1 and 3, and that the conditional mean ATE is equal to the ATE we consider. Thus all results we give for the ATE, including bounds, apply to the conditional mean models, such as that of Chernozhukov et. al. (2007).

To help explain the relationship between the conditional-mean model and the model of our paper, and to illustrate other results, it is useful to consider examples. Binary choice is a very important model for panel data, as it has many applications. For this reason we use binary choice as a 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

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

for scalar ε_{it} and an unknown parameter vector β^*. In this example $g_0(x, \alpha, \varepsilon) = 1(x'\beta^* + \alpha \geq \varepsilon)$ and the ATE is

$$\Delta = \int [1(x^{a\prime}\beta^* + \alpha \geq \varepsilon) - 1(x^{b\prime}\beta^* + \alpha \geq \varepsilon)]dF(\varepsilon, \alpha).$$

This is an unusual object in the binary choice literature but is equal to a conditional mean ATE. In particular, if ε_{it} is independent of (X_i, α_i) with CDF $H(\varepsilon)$ for each t. Then $E[Y_{it}|X_i, \alpha_i] = \Pr(Y_{it} = 1|X_i, \alpha_i) = H(X_{it}'\beta^* + \alpha_i)$ and

$$\Delta = \int [H(x^{a\prime}\beta^* + \alpha) - H(x^{b\prime}\beta^* + \alpha)]dF(\alpha).$$

Thus the ATE is also the effect of changing x on the choice probabilities averaged over the individual effect, i.e. the conditional mean ATE.

Our model also includes binary choice with individual-specific slopes as a special case. Economic motivation for varying slopes is provided by Browning and Carro (2007, 2009) who point out that with constant slopes the sign of the treatment effect is the same for every individual and give empirical examples where varying slopes are important. A general model with varying slopes is $Y_{it} = 1(X_{it}'\alpha_i \geq \varepsilon_{it})$ where X_{it} now includes a constant and ε_{it} is independent of (X_i, α_i) with CDF $H(\varepsilon)$. In this model

$$\Delta = \int [H(x^{a\prime}\alpha) - H(x^{b\prime}\alpha)]dF(\alpha),$$

accounting for individual specific slopes. When X_{it} is discrete and fully saturated (e.g. consists of a full set of dummies, one for every discrete outcome) this model is actually equivalent to the general static model. It will be more restrictive when the distribution of α is restricted in some way, such as having some components of α be constant. In the semiparametric analysis described below we show how to impose such restrictions.

Time effects are clearly important in practice but identification of treatment effects will preclude including t among the regressors X_{it} in the nonparametric model of Assumptions 1 - 3. Identification will be based on variation over time in X_{it}, and if t is a regressor then $g_0(X_{it}, \alpha_i, \varepsilon_{it})$ has unrestricted variation over time, precluding identification of the effect of any other regressor. Some time effects can be allowed for by restricting the way t enters g_0. Below we will describe how this is done in semiparametric, discrete-choice models. With continuous Y_{it} one can allow for location and scale time effects that are relatively easy to estimate.

Assumption 4: *There is a function* $g_0(x,\alpha,\varepsilon)$*, vectors* α_i *and* ε_{it} *of random variables, and constants* $\tau_t, s_t, (t=2,...,T)$ *such that for* $\tau_1 = 0,\ s_1 = 1,$

$$Y_{it} = g_{t0}(X_{it},\alpha_i,\varepsilon_{it}),\ \ g_{t0}(x,\alpha,\varepsilon) = \tau_t + s_t g_0(x,\alpha,\varepsilon), (i=1,...,n; t=1,...,T).$$

This condition allows the mean and variance of Y_{it} to vary over time in an unrestricted way. The condition could be generalized to allow for other time effects, but we leave that to future work. It does not apply to Y_{it} with fixed, discrete support because Assumption 4 does not make sense in that case. There t must be included "inside" g_0, as we do in the semiparametric analysis described below.

With these time effects the ASF and QSF can depend on t. The ASF and QSF for the first period will be $\mu(x)$ and $q(\lambda, x)$ as given above, and for the other periods are

$$\mu_t(x) = \tau_t + s_t\mu(x), q_t(\lambda,x) = \tau_t + s_t q(\lambda,x), (t=2,...,T).$$

Corresponding period-specific and time-averaged ATE and QTE are given by

$$\begin{aligned}\mu_t(x^a) - \mu_t(x^b) &= s_t[\mu(x^a) - \mu(x^b)], q_t(\lambda,x^a) - q_t(\lambda,x^b) = s_t[q(\lambda,x^a) - q(\lambda,x^b)], \qquad (1)\\ & \left(\frac{\sum_{t=1}^T s_t}{T}\right)[\mu(x^a) - \mu(x^b)], \left(\frac{\sum_{t=1}^T s_t}{T}\right)[q(\lambda,x^a) - q(\lambda,x^b)],\end{aligned}$$

where $s_1 = 1$.

In the rest of this paper we will focus on discrete regressors, imposing the following condition from here on:

Assumption 5: *The support of* X_i *is finite.*

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.

3 Identified Effects in the Nonparametric Static Model

The analysis of identification in the static model is quite simple. This simplicity is a virtue, leading to estimators of identified effects and bounds on unidentified effects that are easy to calculate in a very general model. For example, this approach gives a simple solution to the important problem of identification of quantile treatment effects in panel data. The idea is based on Assumption 2, which states that, conditional on X_i, the distribution of unobservables does not vary over time. Therefore, conditional on X_i where both x^b and x^a occur for some time periods, one can identify effects from the changes in Y_{it} across those time periods. For the ATE, the identified conditional effects can be averaged to identify effects conditional on X_i being in subsets where both x^b and x^a occur for some time period. This idea is a slight extension of Chamberlain (1982, pp. 10-17) to discrete regressors that are not binary. For the QTE the distribution functions can be averaged and inverted to identify corresponding quantile effects. This idea appears to be novel.

There is a simple approach to allowing for covariates. Suppose $x = (x_1, x_2)$, and one is interested in the effect of x_1 holding x_2 fixed. Then one can take $x^b = (x_1^b, x_2)$ and $x^a = (x_1^a, x_2)$, so that the effect of changing from x^b to x^a is then the effect of interest. Furthermore, one could average these effects over x_2 to identify an effect that is averaged over covariates. We explicitly allow for covariates in the semiparametric models given below. Because we are already attempting to cover so much ground here, we leave averaging over covariates in the nonparametric model to future work.

To describe identified effects and their estimators we will focus on the ATE and QTE conditional on both x^a and x^b appearing in X_i for some time period. We could also consider effects conditional on smaller subsets of X_i but postpone this until later in order to keep the exposition relatively simple. We need a little more notation to give a precise description. Let $1(X_{it} = x)$ denote the indicator function that is equal to one when $X_{it} = x$ and zero otherwise and let $T_i(x) = \sum_{t=1}^{T} 1(X_{it} = x)$. Here we let the subscript i denote a random variable that may depend on X_i and Y_i. Let $D_i = 1(T_i(x^a) > 0)1(T_i(x^b) > 0)$ be the indicator for the event that X_i includes both x^a and x^b for some time period. Define

$$\delta = E[g_0(x^a, \alpha_i, \varepsilon_{i1}) - g_0(x^b, \alpha_i, \varepsilon_{i1})|D_i = 1]. \tag{2}$$

This δ is the ATE for those individuals where both x^b and x^a occur for some time period. This effect 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 average 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 changed union status at some time. This is δ. Or we might even be interested in the effect for those who were ever in a union. Bounds for such an effect are described below.

A simple estimator of the conditional ATE δ is

$$\hat{\delta} = \frac{\sum_{i=1}^{n} D_i[\bar{Y}_i(x^a) - \bar{Y}_i(x^b)]}{\sum_{i=1}^{n} D_i}, \bar{Y}_i(x) = \begin{cases} T_i(x)^{-1} \sum_{t=1}^{T} 1(X_{it} = x)Y_{it}, T_i(x) > 0 \\ 0, T_i(x) = 0 \end{cases}. \tag{3}$$

Consistency of this estimator results from

$$E[D_i\{\bar{Y}_i(x^a) - \bar{Y}_i(x^b)\}] = E[D_i\{g_0(x^a, \alpha_i, \varepsilon_{i1}) - g_0(x^b, \alpha_i, \varepsilon_{i1})\}],$$

see Lemma A5 of the Supplementary Material. Intuitively, this equation follows from time being randomly assigned, so that we can estimate the effect by comparing Y_{it} where $X_{it} = x^a$ with Y_{is} where $X_{is} = x_b$.

Since $\bar{Y}_i(x^a) - \bar{Y}_i(x^b)$ is a difference of means it can be interpreted as a coefficient of $1(X_{it} = x^a)$ in a regression of Y_{it} on that dummy and on $1(X_{it} = x^a) + 1(X_{it} = x^b)$. Thus, $\hat{\delta}$ is an average of least-squares estimates for each i with $D_i = 1$. From this interpretation we see that $\hat{\delta}$ extends Chamberlain's (1982, p. 12) estimator to discrete regressors that are not binary. A consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\delta} - \delta)$ is $n^{-1}\sum_{i=1}^{n} \hat{\psi}_i^2$ where $\hat{\psi}_i = nD_i[\bar{Y}_i(x^a) - \bar{Y}_i(x^b) - \hat{\delta}]/\sum_{i=1}^{n} D_i$. For brevity we leave the asymptotic theory to the Supplementary Material (see Theorem A6) and efficiency results to future work.

We can also identify and estimate a conditional QTE. Let $G(y, x|D_i = 1) = \Pr(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y|D_i = 1)$ denote the CDF of $g_0(x, \alpha_i, \varepsilon_{i1})$ conditional on $D_i = 1$. The QTE conditional on $D_i = 1$ is

$$\delta_\lambda = G^{-1}(\lambda, x^a|D_i = 1) - G^{-1}(\lambda, x^b|D_i = 1).$$

An estimator of this effect can be constructed using a CDF $\Phi(u)$ and a scalar bandwidth h. An estimator of $G(y, x|D_i = 1)$ is given by

$$\hat{G}(y, x|D_i = 1) = \frac{\sum_{i=1}^{n} D_i \bar{G}_i(y, x)}{\sum_{i=1}^{n} D_i}, \bar{G}_i(y, x) = \begin{cases} T_i(x)^{-1}\sum_{t=1}^{T} 1(X_{it} = x)\Phi(\frac{y - Y_{it}}{h}), T_i(x) > 0, \\ 0, T_i(x) = 0. \end{cases}.$$

In this estimator the indicator function $1(Y_{it} < y)$ has been replaced by a smoothed approximation $\Phi(\frac{y-Y_{it}}{h})$, as suggested by Yu and Jones (1998) for estimating a conditional CDF. An estimator of δ_λ is then

$$\hat{\delta}_\lambda = \hat{q}_\lambda^a - \hat{q}_\lambda^b, \hat{q}_\lambda^a = \hat{G}^{-1}(\lambda, x^a|D_i = 1), \hat{q}_\lambda^b = \hat{G}^{-1}(\lambda, x^b|D_i = 1).$$

Note here that we first average, then invert, and then difference. This estimator solves an important problem of estimating panel quantile effects and appears to be novel.

A consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\delta}_\lambda - \delta_\lambda)$ is $n^{-1}\sum_{i=1}^{n} \hat{\psi}_{\lambda i}^2$ for

$$\hat{\psi}_{\lambda i} = -\frac{nD_i}{\sum_{i=1}^{n} D_i}\left[\frac{\bar{G}_i(\hat{q}^a, x^a) - \lambda}{\hat{G}'(\hat{q}^a, x^a|D_i = 1)} - \frac{\bar{G}_i(\hat{q}^b, x^b) - \lambda}{\hat{G}'(\hat{q}^b, x^b|D_i = 1)}\right],$$

where $\hat{G}'(y, x|D_i = 1) = \partial\hat{G}(y, x|D_i = 1)/\partial y$. Here the denominator terms are actually kernel density estimates. For this reason one might use different bandwidths h in the numerator and denominator, with the denominator chosen to be appropriate for density estimation. Asymptotic theory for this estimator is given in the Supplementary Material (see Theorem A8). Alternatively, one could simply use the bootstrap to construct a confidence interval for $\hat{\delta}_\lambda$.

A helpful example is the binary regressor case where $X_{it} \in \{0,1\}$. Here X_{it} could be thought of as a treatment variable where $X_{it} = 1$ for treated and $X_{it} = 0$ for untreated. Let $Y_{it}(0) = g_0(0, \alpha_i, \varepsilon_{it})$ and $Y_{it}(1) = g_0(1, \alpha_i, \varepsilon_{it})$. Assumption 2 is equivalent to the assumption that the conditional distribution of $(Y_{it}(0), Y_{it}(1))$ given X_i does not vary with t. This is the key assumption that identifies treatment effects from time variation in treatment. In this context $\delta = E[Y_{it}(1) - Y_{it}(0)|D_i = 1]$ is the ATE for individuals where both treatment and nontreatment occurs during the observation period. Similarly, δ_λ is the difference between the λ quantile of the distribution of $Y_{it}(1)$ and the λ quantile for $Y_{it}(0)$ conditional on $D_i = 1$. The ATE and QTE are not identified for those individuals that either receive treatment in every time period or receive no treatment in every time period.

In general the usual panel data within (linear fixed effects) estimator is not a consistent estimator of δ. This inconsistency results because the within estimator constrains the slope coefficient to be the same for each i when the slope is actually varying with i. For simplicity we demonstrate this inconsistency in the binary X_{it} example. The within estimator $\hat{\delta}_w$ is given by

$$\hat{\delta}_w = \frac{\sum_{i=1}^n \sum_{t=1}^T (X_{it} - \bar{X}_i) Y_{it}}{\sum_{i=1}^n \sum_{t=1}^T (X_{it} - \bar{X}_i)^2}, \bar{X}_i = T^{-1} \sum_{t=1}^T X_{it}.$$

Let $\sigma_i^2 = (T-1)^{-1} \sum_{t=1}^T (X_{it} - \bar{X}_i)^2$ be the sample variance over time of X_{it}.

THEOREM 1: *If Assumptions 1 and 2 are satisfied,* $X_{it} \in \{0,1\}$, $E[Y_{it}^2] < \infty$, $(t = 1, ..., T)$, *and* $E[D_i \sigma_i^2] > 0$, *then* $\delta = E[D_i\{\bar{Y}_i(1) - \bar{Y}_i(0)\}]/E[D_i]$ *and*

$$\hat{\delta}_w \xrightarrow{p} \delta_w = \frac{E[\sigma_i^2 D_i \{\bar{Y}_i(1) - \bar{Y}_i(0)\}]}{E[\sigma_i^2 D_i]}. \tag{4}$$

Note that the limit of the within estimator is a weighted average of individual, least-squares estimates $\bar{Y}_i(1) - \bar{Y}_i(0)$ from equation (3). If $T \geq 4$ then the weights σ_i^2 vary over the positive σ_i^2 and so the limit δ_w of $\hat{\delta}_w$ is not the identified conditional ATE δ.

Theorem 1 is different than Yitzhaki (1996) and Angrist (1998), who gave weighted average interpretations of least squares in other, non-panel settings. Theorem 1 is also different from Hahn (2001), who found that $\hat{\delta}_w$ consistently estimates the ATE. Hahn (2001) considered $T = 2$ and assumed $X_i = (0,1)'$. As noted by Hahn (2001), those conditions are quite special. Theorem 1 is also different from Wooldridge (2005), who showed that if $b_i = E[Y_{it}(1) - Y_{it}(0)|\alpha_i]$ is mean independent of $X_{it} - \bar{X}_i$ for each t then linear fixed effects is a consistent estimator of δ. The problem is that the mean-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

$$E[b_i | X_i = (1,0)'] = E[b_i | X_i = (0,1)'] = E[b_i | X_i \in \{(0,0)', (1,1)'\}].$$

This is quite close to independence of b_i and X_i, which is not very interesting if we want to allow the treatment effect to vary with X_i.

The conditional ATE and QTE estimators can easily be modified to accommodate the time effects of Assumption 4. The changes in Y_{it} over time for fixed X_{it} can be used to identify and estimate the time effects that can then be included in the estimation of the ATE and QTE. To describe this approach, let $\hat{m}_t = \sum_{i=1}^n 1(X_{it} = X_{i1})Y_{it} / \sum_{i=1}^n 1(X_{it} = X_{i1})$ and

$$\hat{s}_t = \frac{\sum_{i=1}^n 1(X_{it} = X_{i1})X_{i1}(Y_{it} - \hat{m}_t)}{\sum_{i=1}^n 1(X_{it} = X_{i1})X_{i1}(Y_{i1} - \hat{m}_1)}, \hat{\tau}_t = \hat{m}_t - \hat{s}_t \hat{m}_1, t = 2, ..., T.$$

This $(\hat{\tau}_t, \hat{s}_t)'$ is an instrumental variables estimator where the residual is $Y_{it} - \tau_t - s_t Y_{i1}$, the instruments are $(1, X_{i1})'$, and the estimation is done on the subsample where $X_{it} = X_{i1}$. These estimators will be consistent and asymptotically normal as long as $Cov(X_{i1}, Y_{i1}|X_{it} = X_{i1}) \neq 0$ for each $t = 2, ..., T$. One could also use other functions of X_{i1} as instrumental variables to improve efficiency. We focus on just X_{i1} as an instrument for simplicity. Graham and Powell (2011) use a similar approach to identify time effects in a linear model with continuous regressors.

The time effects are accounted for in ATE and QTE estimation by removing time location and scale effects from all periods when estimating the first period effect, and then putting the scale effects back for other periods. Note first that under Assumption 4 δ is the conditional ATE for the first time period. Let $\tilde{Y}_{it} = (Y_{it} - \hat{\mu}_t)/\hat{s}_t$ be the t^{th} period observation with estimated location and scale removed. Replacing Y_{it} by $\tilde{Y}_{it}$ in the formula for $\hat{\delta}$ gives

$$\tilde{\delta} = \frac{\sum_{i=1}^n D_i[\tilde{Y}_i(x^a) - \tilde{Y}_i(x^b)]}{\sum_{i=1}^n D_i}, \tilde{Y}_i(x) = \begin{cases} T_i(x)^{-1} \sum_{t=1}^T 1(X_{it} = x)\tilde{Y}_{it}, T_i(x) > 0 \\ 0, T_i(x) = 0 \end{cases}.$$

The conditional ATE for the t^{th} time period is given by $s_t \delta$ and a time average by $(\sum_{t=1}^T s_t/T)\delta$ for $s_1 = 1$, analogously to equation (1). These can be estimated by $\hat{s}_t \tilde{\delta}$ and $\bar{s}\tilde{\delta}$, respectively for $\bar{s} = \sum_{t=1}^T \hat{s}_t/T$ and $\hat{s}_1 = 1$. These estimators will be consistent and asymptotically normal. Because of their multistage nature the bootstrap may provide the easiest approach to carrying out inference on these estimators, where one resamples from the empirical distribution of (Y_i, X_i), $(i = 1, ..., n)$ to form confidence intervals for the true parameter. For brevity we omit explicit results.

An analogous approach can be followed to account for time effects in the QTE. The interpretation of δ_λ now becomes QTE for the first time period conditional on $D_i = 1$. An estimator of $G(y, x|D_i = 1) = \Pr(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y|D_i = 1)$ that adjusts for time, location and scale is given by

$$\tilde{G}(y, x|D_i = 1) = \frac{\sum_{i=1}^n D_i \tilde{G}_i(y, x)}{\sum_{i=1}^n D_i}, \tilde{G}_i(y, x) = \begin{cases} T_i(x)^{-1} \sum_{t=1}^T 1(X_{it} = x)\Phi(\frac{y - \tilde{Y}_{it}}{h}), T_i(x) > 0 \\ 0, T_i(x) = 0 \end{cases}.$$

Let $\tilde{q}^a_\lambda = \tilde{G}^{-1}(\lambda, x^a|D_i = 1)$ and $\tilde{q}^b_\lambda = \tilde{G}^{-1}(\lambda, x^b|D_i = 1)$. Estimators for the conditional QTE for the first period, other periods, and a time average are given by $\tilde{\delta}_\lambda = \tilde{q}^a_\lambda - \tilde{q}^b_\lambda$, $\hat{s}_t\tilde{\delta}_\lambda, (t = 2, ..., T)$, and $\bar{s}\tilde{\delta}_\lambda$, respectively. Here again the bootstrap provides a convenient method for inference. One could also use quantiles to estimate the time effects, but we avoid that for simplicity.

4 Nonparametric Bounds in the Static Model

When $g_0(x, \alpha_i, \varepsilon_{it})$ is bounded we can estimate bounds for the ASF and corresponding bounds for the ATE. For the QSF and QTE we can also estimate bounds without any restriction on g_0, using the fact that there are known upper and lower bounds for the indicator function $1(g_0(x, \alpha_i, \varepsilon_{it}) \leq y)$. The idea of the bounds is an extension of the estimation of identified effects discussed in the previous Section. Time homogeneity allows us to use time averages to estimate the identified parts of the ASF or QSF when x is an element of X_i, i.e. $X_{it} = x$ for some t, and apply the lower or upper bounds when x does not appear in X_i.

We first describe bounds estimation for the ASF. These bounds depend on bounds on g_0 imposed in the following condition:

ASSUMPTION 6: *$B_\ell \leq g_0(x, \alpha_i, \varepsilon_{it}) \leq B_u$ for constants B_ℓ and B_u and all x.*

For example, in the binary-choice model, where $Y_{it} \in \{0, 1\}$, upper and lower bounds are $B_u = 1$ and $B_\ell = 0$ respectively. We could allow B_ℓ and B_u to depend on x and using that information could tighten the ATE bounds given below. To avoid further complication we do not allow this.

Let $T_i(x)$ and $\bar{Y}_i(x)$ be as in Section 3 and $\bar{P}(x) = \sum_{i=1}^n 1(T_i(x) = 0)/n$ be the sample frequency of x not occurring in any time period. Estimated lower and upper bounds for $\mu(x)$ are

$$\hat{\mu}_\ell(x) = n^{-1}\sum_{i=1}^n \bar{Y}_i(x) + \bar{P}(x)B_\ell, \hat{\mu}_u(x) = \hat{\mu}_\ell(x) + \bar{P}(x)(B_u - B_\ell).$$

Here $\bar{Y}_i(x)$ estimates the identified part of the ASF, corresponding to $T_i(x) > 0$, and the upper and lower bounds are applied for observations where $T_i(x) = 0$. Corresponding estimated lower and upper bounds for the ATE are $\hat{\Delta}_\ell = \hat{\mu}_\ell(x^a) - \hat{\mu}_u(x^b)$ and $\hat{\Delta}_u = \hat{\mu}_u(x^a) - \hat{\mu}_\ell(x^b)$. The width of these estimated bounds is

$$\hat{\Delta}_u - \hat{\Delta}_\ell = [\bar{P}(x^a) + \bar{P}(x^b)](B_u - B_\ell).$$

For example, for binary choice with a binary regressor, where $B_u = 1$ and $B_\ell = 0$, the width of the estimated bounds for the ATE is $\bar{P}(0) + \bar{P}(1)$, where $\bar{P}(0)$ and $\bar{P}(1)$ are the sample proportions of X_i with $X_{it} = 1$ for all t and $X_{it} = 0$ for all t, respectively

These estimators will be jointly asymptotically normal under i.i.d. (Y_i, X_i). The asymptotic variance can be estimated by $\hat{\Sigma} = \sum_{i=1}^n \hat{\Psi}_i \hat{\Psi}_i'/n$, where

$$\hat{\Psi}_i = \begin{pmatrix} \bar{Y}_i(x^a) - \bar{Y}_i(x^b) + B_\ell 1(T_i(x^a) = 0) - B_u 1(T_i(x^b) = 0) - \hat{\Delta}_\ell \\ \bar{Y}_i(x^a) - \bar{Y}_i(x^b) + B_u 1(T_i(x^a) = 0) - B_\ell 1(T_i(x^b) = 0) - \hat{\Delta}_u \end{pmatrix}.$$

Confidence intervals for the identified set can then be formed using results of Chernozhukov, Hong, and Tamer (2007) or Beresteanu and Molinari (2008, pp. 779-781) on estimators of intervals where the upper and lower endpoints are jointly asymptotically normal.

Turning to the bounds for the QSF, lower and upper estimated bounds for the $G(y, x) = \Pr(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y)$ are $\hat{G}_\ell(y, x) = \sum_{i=1}^n \bar{G}_i(y, x)/n$ and $\hat{G}_u(y, x) = \hat{G}_\ell(y, x) + \bar{P}(x)$ respectively. The idea of these bounds is similar to the ASF, with a known lower bound of 0 and upper bound of 1 for $1(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y)$. To obtain quantile bounds we need to invert these functions of y. For a strictly increasing function $G(y)$ with range contained in $[0, 1]$ let

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

This is a function with domain $[0, 1]$ and range equal to the extended real line that can be used to invert $\hat{G}_u(y, x)$ and $\hat{G}_\ell(y, x)$. Estimators of lower and upper bounds on the QSF are given by

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

Corresponding lower and upper bounds for the QTE are $\hat{\Delta}_{\lambda\ell} = \hat{q}_\ell^a - \hat{q}_u^b$ and $\hat{\Delta}_{\lambda u} = \hat{q}_u^a - \hat{q}_\ell^b$ where $\hat{q}_\ell^a = \hat{q}_\ell(\lambda, x^a)$, $\hat{q}_u^a = \hat{q}_u(\lambda, x^a)$, $\hat{q}_\ell^b = \hat{q}_\ell(\lambda, x^b)$, and $\hat{q}_u^b = \hat{q}_u(\lambda, x^b)$. The width of these bounds depends on the shape of the empirical distribution of Y_{it} and on $\bar{P}(x)$. The width of the bounds will be finite when

$$\max\{\bar{P}(x^a), \bar{P}(x^b)\} < \lambda < \min\{1 - \bar{P}(x^a), 1 - \bar{P}(x^b)\}, \tag{5}$$

and otherwise they are infinitely wide.

The bounds will be joint asymptotically normal under the following regularity condition:

Assumption 7: *$\Pr(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y | X_i)$ is twice continuously differentiable in y with uniformly bounded derivatives and $G_\ell(y, x) = E[E[1(T_i(x) > 0)|X_i] \Pr(g_0(x, \alpha_i, \varepsilon_{i1}) \leq y|X_i)]$ is strictly increasing in y on the interior of its range for all x. Also $nh^4 \longrightarrow 0$ and $nh^2 \longrightarrow \infty$.*

For λ satisfying equation (5) the asymptotic variance can be estimated by $\hat{\Sigma}_\lambda = \sum_{i=1}^n \hat{\Psi}_{\lambda i} \hat{\Psi}_{\lambda i}'/n$, where

$$\hat{\Psi}_{\lambda i} = \begin{pmatrix} \frac{\bar{G}_i(\hat{q}_\ell^a, x^a) + 1(T_i(x^a)=0) - \lambda}{\hat{G}_\ell'(\hat{q}_\ell^a, x^a)} - \frac{\bar{G}_i(\hat{q}_u^b, x^b) - \lambda}{\hat{G}_\ell'(\hat{q}_u^b, x^b)} \\ \frac{\bar{G}_i(\hat{q}_u^a, x^a) - \lambda}{\hat{G}_u'(\hat{q}_u^a, x^a)} - \frac{\bar{G}_i(\hat{q}_\ell^b, x^b) + 1(T_i(x^b)=0) - \lambda}{\hat{G}_\ell'(\hat{q}_\ell^b, x^b)} \end{pmatrix}.$$

As in estimation of the conditional quantile effect, one might want to use different bandwidths for numerators and denominators, or just bootstrap to estimate the asymptotic variance.

Here is a result for both ATE and QTE bounds:

THEOREM 2: *Suppose that Assumptions 1, 2, and 5 are satisfied. If Assumption 6 is satisfied then there are* Δ_ℓ, Δ_u, *and* Σ *such that*

$$\sqrt{n}[(\hat{\Delta}_\ell, \hat{\Delta}_u)' - (\Delta_\ell, \Delta_u)'] \xrightarrow{d} N(0, \Sigma), \hat{\Sigma} \xrightarrow{p} \Sigma.$$

where $\Delta_\ell \leq \Delta \leq \Delta_u$, *and these bounds are sharp. If Assumption 7 is satisfied then there are* $\Delta_{\lambda\ell}$, $\Delta_{\lambda u}$, *and* Σ_λ *such that*

$$\sqrt{n}[(\hat{\Delta}_{\lambda\ell}, \hat{\Delta}_{\lambda u})' - (\Delta_{\lambda\ell}, \Delta_{\lambda u})'] \xrightarrow{d} N(0, \Sigma_\lambda), \hat{\Sigma}_\lambda \xrightarrow{p} \Sigma_\lambda.$$

where $\Delta_{\lambda\ell} \leq \Delta_\lambda \leq \Delta_{\lambda u}$. *If* $G_\ell(y, x)$ *is also everywhere strictly increasing in* y *then these bounds are sharp.*

The sharpness conclusion of Theorem 2 for the ATE 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.

Similarly to the treatment-effects literature, we may be interested in the ATE or QTE, conditional on $X_i \in S$ for some set S. For example, if $X_{it} \in \{0, 1\}$ represents treatment then we might be interested in the effect of treatment conditional on ever treated, i.e. conditional on $X_i \neq (0, ..., 0)'$. Tighter bounds for such effects can be formed and in some cases the effects may be identified. These bounds can be estimated by replacing $1(X_{it} = x)$ by $1(X_i \in S)1(X_{it} = x)$ in the definition of $\bar{Y}_i(x)$ and $\bar{G}_i(y, x)$, $1(T_i(x) = 0)$ by $1(X_i \in S)1(T_i(x) = 0)$ in the definition of $\bar{P}(x)$, and dividing through by $\sum_{i=1}^n 1(X_i \in S)/n$. If $1(X_i \in S) \leq D_i$ for D_i from Section 3 the corresponding effects will be identified, and the upper and lower estimated bounds will be identical.

Time effects can easily be allowed for in quantile-effect bounds by adapting the approach used earlier. It is not clear that allowing for time effects in that way makes sense for bounds on the ATE, e.g. for binary choice models where the support of Y_{it} is fixed. Therefore we focus just on time effects in quantile bounds. For QTE bounds we can replace Y_{it} by $\tilde{Y}_{it} = (Y_{it} - \hat{\mu}_t)/\hat{s}_t$ in the formula for $\hat{G}_\ell(y, x)$ given above, and interpret $\hat{\Delta}_{\lambda\ell}$ and $\hat{\Delta}_{\lambda u}$ as estimators of the first period bounds. Estimators of t^{th} period lower and upper bounds for the QTE are then given by $\hat{s}_t\hat{\Delta}_{\lambda\ell}$ and $\hat{s}_t\hat{\Delta}_{\lambda u}$ respectively. Estimators of time average bounds are $\bar{s}\hat{\Delta}_{\lambda\ell}$ and $\bar{s}\hat{\Delta}_{\lambda u}$, where $\bar{s} = \sum_{t=1}^T \hat{s}_t/T$. These upper and lower bounds will be joint asymptotically normal, and their asymptotic variance can be estimated by the bootstrap.

5 Nonparametric Bounds in the Dynamic Model

Analysis of the dynamic model is more challenging than that of the static one. In the dynamic model of Assumption 3 only the first-period regressor is common to the conditioning sets for each time period. Consequently location and scale time effects are not identified, because the conditioning set is different for every time period. For this reason we do not consider time effects in the nonparametric dynamic model. Also, the identification and bounds analysis is limited to objects that are conditional on the first period or are unconditional. For example, we cannot identify or bound the ATE conditional on X_{it} changing over time because that event involves information about all time periods. We can bound unconditional objects and ones that are conditional on just X_{i1}. These bounds are simple and novel, for example in providing partial-identification results for the average effect of state dependence with heterogeneity in both location and slope when Y_{it} is binary and $X_{it} = Y_{it-1}$.

The model with a binary, lagged dependent variable has $Y_{it} = g_0(Y_{i,t-1}, \alpha_i, \varepsilon_{it})$, and under Assumption 3,

$$\begin{aligned} \Pr(Y_{it} = 1|X_{it}, ..., X_{i1}, \alpha_i) &= \int g(Y_{i,t-1}, \alpha_i, \varepsilon) dF(\varepsilon|\alpha_i, Y_{i0}) \\ &= \Pr(Y_{it} = 1|Y_{i,t-1}, \alpha_i, Y_{i0}), \end{aligned}$$

where $F(\varepsilon|\alpha_i, Y_{i0})$ denotes the conditional CDF of ε_{it} given α_i and Y_{i0}. Here $\Pr(Y_{it} = 1|Y_{i,t-1}, \alpha_i, Y_{i0})$ does not vary with t, and the model places no other restrictions on $\Pr(Y_{it} = 1|Y_{i,t-1}, \alpha_i, Y_{i0})$. Conditioning on Y_{i0} is present to account correctly for the initial condition, as in Honore and Tamer (2006) and Browning and Carro (2007, 2009). The probabilities can be distributed across individuals in any way at all through the individual effect α_i. That is we can think of the four conditional probabilities,

$$\Pr(Y_{it} = 1|1, \alpha_i, 1), \Pr(Y_{it} = 1|0, \alpha_i, 1) \Pr(Y_{it} = 1|1, \alpha_i, 0), \Pr(Y_{it} = 1|0, \alpha_i, 0),$$

as having an unrestricted distribution. Here the ATE is

$$\Delta = \int [\Pr(Y_{it} = 1|Y_{i,t-1} = 1, \alpha, Y_0) - \Pr(Y_{it} = 1|Y_{i,t-1} = 0, \alpha, Y_0)] dF(\alpha, Y_0).$$

This object quantifies the effect of state dependence in the presence of individual heterogeneity, an important problem posed by Feller (1943) and Heckman (1981). The dynamic bounds here provide a simple, estimable, identified set for this object. This model is considered by Browning and Carro (2007, 2009), who derive properties of various estimators and restrictions on α_i that lead to identification. We give nonparametric bounds.

A partition of X_i values that preserves the dynamic structure of Assumption 3 is used to obtain bounds for the ASF and QSF. For each x we partition X_i into realizations where the first occurrence of x is at time t and the set where x never occurs. This partition is given by $\{\bar{\mathcal{X}}(x), \mathcal{X}_1(x), ..., \mathcal{X}_T(x)\}$ where

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

Define $\hat{Y}_i(x) = \sum_{t=1}^T 1(X_i \in \mathcal{X}_t(x))Y_{it}$, which picks out the Y_{it} for the time period where x first occurs. Estimated lower and upper ASF bounds are

$$\hat{\mu}_\ell(x) = n^{-1}\sum_{i=1}^n \hat{Y}_i(x) + \bar{P}(x)B_\ell, \hat{\mu}_u(x) = \hat{\mu}_\ell(x) + \bar{P}(x)(B_u - B_\ell).$$

Corresponding lower and upper bounds for Δ are $\hat{\Delta}_\ell = \hat{\mu}_\ell(x^a) - \hat{\mu}_u(x^b)$ and $\hat{\Delta}_u = \hat{\mu}_u(x^a) - \hat{\mu}_\ell(x^b)$. A joint asymptotic-variance estimator $\hat{\Sigma}$ can be constructed exactly as for the static case with $\hat{Y}_i(x)$ replacing $\bar{Y}_i(x)$.

It is interesting to note that the width $\bar{P}(x)(B_u - B_\ell)$ of the estimated ASF bounds is the same for the dynamic and static models. Because the static model is a special case of the dynamic one we conjecture that the bounds for the dynamic model are sharp like the bounds for the static one, but have not yet been able to show this.

To construct estimated lower and upper bounds for the CDF of $g_0(x, \alpha_i, \varepsilon_{it})$ let $\hat{G}_i(y, x) = \sum_{t=1}^T 1(X_i \in \mathcal{X}_t(x))\Phi(\frac{y - Y_{it}}{h})$. The estimated CDF bounds are

$$\hat{G}_\ell(y, x) = \frac{1}{n}\sum_{i=1}^n \hat{G}_i(y, x), \hat{G}_u(y, x) = \hat{G}_\ell(y, x) + \bar{P}(x).$$

Estimated lower and upper bounds for the QSF are then given by

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

Corresponding lower and upper bounds for the QTE are $\hat{\Delta}_{\lambda\ell} = \hat{q}_\ell(\lambda, x^a) - \hat{q}_u(\lambda, x^b)$ and $\hat{\Delta}_{\lambda u} = \hat{q}_u(\lambda, x^a) - \hat{q}_\ell(\lambda, x^b)$. A joint asymptotic variance estimator $\hat{\Sigma}_\lambda$ can be constructed just as for the static case with $\hat{G}_i(y, x)$ replacing $\bar{G}_i(y, x)$.

THEOREM 3: *Suppose that Assumptions 1, 3, and 5 are satisfied. If Assumption 6 is satisfied then there are Δ_ℓ, Δ_u, and Σ such that*

$$\sqrt{n}[(\hat{\Delta}_\ell, \hat{\Delta}_u)' - (\Delta_\ell, \Delta_u)'] \xrightarrow{d} N(0, \Sigma), \hat{\Sigma} \xrightarrow{p} \Sigma.$$

where $\Delta_\ell \leq \Delta \leq \Delta_u$. Also if Assumption 7 is satisfied with X_{i1} replacing X_i then there are $\Delta_{\lambda\ell}$, $\Delta_{\lambda u}$, and Σ_λ such that

$$\sqrt{n}[(\hat{\Delta}_{\lambda\ell}, \hat{\Delta}_{\lambda u})' - (\Delta_{\lambda\ell}, \Delta_{\lambda u})'] \xrightarrow{d} N(0, \Sigma_\lambda), \hat{\Sigma}_\lambda \xrightarrow{p} \Sigma_\lambda.$$

where $\Delta_{\lambda\ell} \leq \Delta_\lambda \leq \Delta_{\lambda u}$.

Similarly to the static model we may be interested in effects conditional on $X_{i1} \in S_1$ for some set S_1. For example, if $X_{it} \in \{0,1\}$ represents treatment then we might be interested in the effect of treatment conditional on being treated in the first period, i.e. conditional on $X_{i1} = 1$. Tighter bounds for such effects can be estimated by replacing $1(X_i \in \mathcal{X}_t(x))$ by $1(X_{i1} \in S_1)1(X_i \in \mathcal{X}_t(x))$ in the definition of $\hat{Y}_i(x)$ and $\hat{G}_i(y,x)$, $1(T_i(x) = 0)$ by $1(X_{i1} \in S_1)1(T_i(x) = 0)$ in the definition of $\bar{P}(x)$, and dividing through by $\sum_{i=1}^n 1(X_{i1} \in S_1)/n$.

In the binary, lagged-dependent-variable example we have $B_\ell = 0$ and $B_u = 1$, so the bounds on the ATE are

$$\hat{\Delta}_\ell = \frac{1}{n}\sum_{i=1}^n [\hat{Y}_i(1) - \hat{Y}_i(0)] - \bar{P}(0), \hat{\Delta}_u = \hat{\Delta}_\ell + \bar{P}(1) + \bar{P}(0).$$

Here $\bar{P}(1) + \bar{P}(0)$ estimates the width of the bounds, providing a very simple measure of the severity of the problem of identifying state dependence in the presence of heterogeneity. The bounds will tend to be wide in short panels but more informative in long ones.

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

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

We consider different DGPs indexed by $\beta^* \in [-2, 2]$ and compute the width of the bounds for $T \in \{2, 4, 8, 16, 32, 64\}$. The width is asymmetric with respect to $\beta^* = 0$ because $\Pr(X_i = (1, ..., 1)')$ grows with β^*, whereas $\Pr(X_i = (0, ..., 0)')$ does not depend on β^*. The width growing with β^* may therefore be explained by having fewer switches of Y_{it} between one and zero when β^* is larger. It is presumably the changes that help identify the ATE. We find that the bounds can be substantially wide for high values of β^* even for large T, consistent with the width of the nonparametric bounds shrinking only at rate $1/T$, as shown in the next Section. Semiparametric bounds for this model that impose the constancy of β^* across individuals, will shrink much faster at T grows, as shown in Section 7.

6 The Impact of T

Increasing T improves identification, shrinking the estimated and population-identified sets for the objects of interest. The rate at which the identified set shrinks quantifies this improvement. Here we give rates for the ASF and, for brevity, leave the quantile results to the Supplementary Material.

The width of the population bounds for the ASF is $(B_u - B_\ell)\bar{\mathcal{P}}(x)$ where

$$\bar{\mathcal{P}}(x) = \Pr(X_{i1} \neq x, ..., X_{iT} \neq x).$$

Thus, the rate at which the identified set shrinks, that we will refer to as the identification rate, is the same as the rate at which $\bar{\mathcal{P}}(x)$ shrinks. Factors that determine this rate can be seen when X_{it} is i.i.d. conditional on α_i. In that case

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

The rate at which $\bar{\mathcal{P}}(x)$ goes to zero will be determined by how much probability mass of $\Pr(X_{it} \neq x|\alpha_i)$ is close to one. If $\Pr(X_{it} \neq x|\alpha_i) = 1$ with positive probability then $\bar{\mathcal{P}}(x)$ does not go to zero. This corresponds to nonidentification of the ASF, where x does not occur for some individuals as indexed by α_i (see Theorem A11 of the Supplementary Material). On the other hand, if $\Pr(X_{it} \neq x|\alpha_i)$ is bounded away from one then the identified set will shrink exponentially quickly, since $\Pr(X_{it} \neq x|\alpha_i)^T \leq (1-\varepsilon)^T$ for some $\varepsilon > 0$. In between the nonidentified and exponential rate cases there are a range of rates depending on how much of the distribution of $\Pr(X_{it} \neq x|\alpha_i)$ is close to 1. The following result shows the range of rates.

Theorem 4: *Suppose that Assumptions 1, 3, 5, and 6 are satisfied and $(X_{i1}, X_{i2}, ...)$ is stationary and Markov of order J conditional on α_i. If for some $\varepsilon > 0$, $\Pr(X_{it} = x|X_{i,t-1}, ..., X_{i,t-J}, \alpha_i) \geq \varepsilon$ a.s. then $\mu_u(x) - \mu_\ell(x) \leq (B_u - B_\ell)(1-\varepsilon)^{T-J}$. If X_{it} is i.i.d. conditional on α_i, $\Pr(X_{it} \neq x|\alpha_i)$ is continuously distributed with pdf $f_P(p)$, and*

$$f_P(p) \leq Cp^{\gamma-1}(1-p)^{v-1}, \gamma > 0, v > 0, \tag{6}$$

then $\mu_u(x) - \mu_\ell(x) = O(T^{-v})$.

The upper bound on the rate at which the pdf $f_P(p)$ of $\Pr(X_{it} \neq x|\alpha_i)$ grows or converges to zero as $p \longrightarrow 1$ provides an upper bound on the rate at which the identified set shrinks. For example, if $v = 1$ so that $f_P(p)$ is bounded as $p \longrightarrow 1$, then the identified set shrinks at rate $1/T$. All of the rates implied by this result are slower than the exponential rate, reflecting how having $\Pr(X_{it} \neq x|\alpha_i)$ close to 1 affects the rate. Also, γ has no effect on the convergence rate because that rate is determined by closeness of $\Pr(X_{it} \neq x|\alpha_i)$ to 1, and not to 0.

The dynamic, binary-choice model is an example where more explicit conditions can be given. Suppose $Y_{it} = 1(\alpha_{i1} + (\alpha_{i2} - \alpha_{i1})Y_{i,t-1} \geq \varepsilon_{it})$ and ε_{it} is i.i.d. and independent of $\alpha_i = (\alpha_{i1}, \alpha_{i2})$ with CDF $H(\varepsilon)$. Here $\Pr(Y_{it} = 1|Y_{i,t-1} = 0, \alpha_i) = H(\alpha_{i1})$ and $\Pr(Y_{it} = 1|Y_{i,t-1} = 1, \alpha_i) = H(\alpha_{i2})$. Unbounded α_i and bounded ε_{it} will correspond to the unidentified case. Bounded α_i and unbounded ε_{it} lead to an exponential convergence rate. The following result covers the in-between case. Let $f_\varepsilon(\varepsilon)$, $f_{\alpha_1}(\alpha)$, and $f_{\alpha_2}(\alpha)$ denote the pdfs of ε_{it}, α_{i1}, and α_{i2} respectively, all are assumed to be continuously distributed.

THEOREM 5: *If* $Y_{it} = 1(\alpha_{i1} + (\alpha_{i2} - \alpha_{i1})Y_{i,t-1} \geq \varepsilon_{it})$, *where* $\varepsilon_{it}, (t = 1, ..., T)$ *is i.i.d. and independent of* $(\alpha_{i1}, \alpha_{i2})$ *and there is* $v, C > 0$ *such that for all* ε

$$\max_{j=1,2} f_{\alpha_j}(\varepsilon) \leq CH(\varepsilon)^{v-1}[1 - H(\varepsilon)]^{v-1} f_\varepsilon(\varepsilon), \tag{7}$$

then $\Delta_u - \Delta_\ell = O(T^{-v})$.

Here we see that the identification rate in the nonparametric dynamic model is related to the tail thickness of the distribution of α_{i1} and α_{i2} relative to the distribution of ε_{it}. The thinner the tail of $f_\varepsilon(\varepsilon)$ relative to the tails of $f_{\alpha_1}(\alpha_1)$ and $f_{\alpha_2}(\alpha_2)$ the smaller v will need to be to satisfy the inequality in Theorem 5 and the slower the identification rate will be. In this way the identification rate is slower the less strong the signal provided by ε_{it} relative to the individual effects. Here there is no γ present because both left and right tails matter, in order to bound the rate for the ATE, and not just for the ASF at a particular x.

For a specific example consider α_{i1} and α_{i2} as $N(0, \sigma_\alpha^2)$ and ε_{it} as $N(0, \sigma_\varepsilon^2)$ where $\sigma_\varepsilon^2 \leq \sigma_\alpha^2$. Then for constants C_1, C_2, and $v = \sigma_\varepsilon^2/\sigma_\alpha^2$ we have $f_{\alpha_j}(\varepsilon) = C_1[f_\varepsilon(\varepsilon)]^v$. Also, as is well known for the Gaussian distribution, $f_\varepsilon(\varepsilon) \geq C_2 F_\varepsilon(\varepsilon)[1 - F_\varepsilon(\varepsilon)]$, where $F_\varepsilon(\varepsilon)$ denotes the CDF of ε. It follows by $v \leq 1$ that

$$f_{\alpha_j}(\varepsilon) = C_1[f_\varepsilon(\varepsilon)]^{v-1} f_\varepsilon(\varepsilon) \leq C_1 C_2^{v-1} F_\varepsilon(\varepsilon)^{v-1}[1 - F_\varepsilon(\varepsilon)]^{v-1} f_\varepsilon(\varepsilon).$$

Thus equation (7) is satisfied with $v = \sigma_\varepsilon^2/\sigma_\alpha^2$ so that

$$\Delta_u - \Delta_\ell = O(T^{-\sigma_\varepsilon^2/\sigma_\alpha^2}).$$

Hence the width of the bounds shrinks at a rate no larger than T^{-1} and the rate is slower the smaller $\sigma_\varepsilon^2/\sigma_\alpha^2$ is. It can also be shown that convergence is faster than T^{-1} when $\sigma_\varepsilon^2 > \sigma_\alpha^2$ and increases with $\sigma_\varepsilon^2/\sigma_\alpha^2$. Thus we see that the stronger the signal provided by ε relative to that provided by α, in the sense that the higher σ_ε^2 is relative to σ_α^2, the faster will be the identification rate.

One can obtain analogous results in a static model. If $X_{it} = 1(\alpha_i \geq \eta_{it})$ is a binary regressor where η_{it} is i.i.d. over time then the identification rate will be T^{-v} when the inequality in Theorem 5 is satisfied with the pdf $f_\eta(\eta)$ of η_{it} replacing the pdf $f_\varepsilon(\varepsilon)$. If α_i and η_{it} are distributed as $N(0, \sigma_\alpha^2)$ and η_{it} as $N(0, \sigma_\eta^2)$ respectively with $\sigma_\eta^2 \leq \sigma_\alpha^2$, then the identified set shrinks at rate $T^{-\sigma_\eta^2/\sigma_\alpha^2}$. For brevity we omit the details.

7 Semiparametric Multinomial Choice Models

The nonparametric bounds are informative but may be quite wide for small T. They can be tightened by imposing additional structure on the model. One way to do this is to specify a parametric model for the conditional distribution of Y_i given values for (X_i, α_i). We focus here on multinomial choice models. In those models Y_i is one of a finite number of outcomes, denoted here by $\{Y^1, ..., Y^J\}$. The parametric part of the model are the known conditional probabilities $\mathcal{L}_j^k(\alpha, \beta)$ of $Y_i = Y^j$ given α_i and $X_i \in \mathcal{X}^k, (k = 1, ..., K)$, where β is a parameter vector with true value β^*, and $\mathcal{X}^k$ is the set of X_i values being conditioned on. Formulating the model in this way allows for X_i that are lagged dependent variables. The nonparametric part of the model will be the unknown CDF's $F_k^*(\alpha), (k = 1, ..., K)$ of α_i conditional on X_i in each $\mathcal{X}^k$. The model then satisfies

Assumption 8: $\Pr(Y_i = Y^j | X_i \in \mathcal{X}^k) = \int \mathcal{L}_j^k(\alpha, \beta^*) dF_k^*(\alpha), (j = 1, ..., J; k = 1, ..., K)$.

Some examples may be helpful. An important example is a binary choice model where $Y_{it} \in \{0, 1\}$, α is a scalar location individual effect, $\Pr(Y_{it} = 1 | X_i, \alpha_i, \beta^*) = H(X_{it}'\beta^* + \alpha_i)$ for a CDF $H(\varepsilon)$, and $Y_{i1}, ..., Y_{iT}$ are mutually independent conditional on X_i and α_i. In this case we would let $\mathcal{X}^k$ be a singleton given by the k^{th} value X^k in the finite support of X_i and

$$\mathcal{L}_j^k(\alpha, \beta) = \prod_{t=1}^{T} H(X_t^{k\prime}\beta + \alpha)^{Y_t^j}[1 - H(X_t^{k\prime}\beta + \alpha)]^{1-Y_t^j}. \tag{8}$$

Time effects can be included in this model by specifying that some components of X_t^k only depend on t. This model can also be generalized to allow for some slopes to vary across individuals by specifying that

$$\mathcal{L}_j^k(\alpha, \beta) = \prod_{t=1}^{T} H(z_t'\beta_1 + X_{t1}^{k\prime}\beta_2 + X_{t2}^{k\prime}\alpha)^{Y_t^j}[1 - H(z_t'\beta_1 + X_{t1}^{k\prime}\beta_2 + X_{t2}^{k\prime}\alpha)]^{1-Y_t^j}. \tag{9}$$

This model allows the coefficients of X_{t2}^k to vary with individuals, which will include a location effect when some element of X_{t2}^k does not vary with t or k.

This set up also allows for dynamic models. For example, consider a binary choice model with a lagged dependent variable where $\Pr(Y_{it} = 1 | Y_{i,t-1}, ..., Y_{i0}, \alpha_i, \beta^*) = H(Y_{i,t-1}\beta^* + \alpha_i)$. Here $X_i = (Y_{i,T-1}, ..., Y_{i0})$ and we take $K = 2$, with $\mathcal{X}^k = \{X_i : X_{i1} = Y_{i0} = k - 1\}$. The parametric part of the model is

$$\begin{aligned} \mathcal{L}_j^k(\alpha, \beta) &= \prod_{t=2}^{T} H(Y_{t-1}^j\beta + \alpha)^{Y_t^j}[1 - H(Y_{t-1}^j\beta + \alpha)]^{1-Y_t^j} \\ &\quad \times H((k-1)\beta + \alpha)^{Y_1^j}[1 - H((k-1)\beta + \alpha)]^{1-Y_1^j}. \end{aligned} \tag{10}$$

This model could be generalized to allow individual specific coefficients for the dynamic effect, time effects, and other covariates, including the model of Browning and Carro (2009). For brevity we omit this generalization.

The ATE and its bounds can be decomposed into a weighted average of conditional ATE and corresponding bounds, weighted by the identified $\Pr(X_i \in \mathcal{X}^k)$. The semiparametric model may restrict the conditional bounds so we focus first on them. We will assume that a conditional ATE takes the form

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

where $\Delta(\alpha, \beta)$ denotes a treatment effect conditional on α. For example, in the model of equation (8) we could take $\Delta(\alpha, \beta) = H(x^{a\prime}\beta + \alpha) - H(x^{b\prime}\beta + \alpha)$, in which case

$$\Delta^k = \int [H(x^{a\prime}\beta^* + \alpha) - H(x^{b\prime}\beta^* + \alpha)] dF_k^*(\alpha)$$

is the ATE conditional on $X_i = X^k$. One could also consider the ASF conditional on $X_i = X^k$, that would be $\int H(x'\beta^* + \alpha) dF_k^*(\alpha)$ in this example.

Neither Δ^k nor β^* need be identified. Instead, there may 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 with $\mathcal{P}_j^k = \Pr(Y_i = Y^j | X_i \in \mathcal{X}^k)$ and

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

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

$$\Delta_u^k = \sup_{\beta \in B, F_k \in \mathcal{F}_k(\beta, \mathcal{P})} \int \Delta(\alpha, \beta) dF_k(\alpha), \; \Delta_\ell^k = \inf_{\beta \in B, F_k \in \mathcal{F}_k(\beta, \mathcal{P})} \int \Delta(\alpha, \beta) dF_k(\alpha). \tag{11}$$

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.

For purposes of comparison with the nonparametric results we consider models without trends, where the semiparametric models in equations (8) and (10) are nested in the nonparametric static or dynamic model. In those models Δ^k will be identified if it is also identified in the nonparametric model. In the static case Δ^k is nonparametrically identified if X_t^k takes on the values x^b and x^a for some time periods. This follows similarly to the identification of the conditional effect δ in Section 3. Therefore, in static models obtaining a smaller identified set by imposing the restrictions of a semiparametric model is limited to those Δ^k where at least one of x^b or x^a does not appear in any time period. In what follows we focus on these Δ^k.

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 $Y_{it} = 1((\alpha_{i2}-\alpha_{i1})X_{it}+\alpha_{i1} > \varepsilon_{it})$, ε_{it} is independent of $(X_i, \alpha_{i2}, \alpha_{i1})$, and ε_{it} has known CDF $H(\varepsilon)$ that is strictly increasing on the entire real line. The joint distribution of $H(\alpha_{i1})$ and $H(\alpha_{i2})$ conditional on $X_i = X^k$ is entirely unrestricted. Therefore when $X^k = (0, ..., 0)'$ the fact that $E[H(\alpha_{i1})|X_i = X^k] = E[Y_{it}|X_i = X^k]$ for every every t, and so is identified gives no information about $E[H(\alpha_{i2})|X_i = X^k]$. Thus, $E[H(\alpha_{i2})|X_i = X^k]$ can be anything in the unit interval. Therefore, the width of the bound for $\Delta^k = E[H(\alpha_{i2}) - H(\alpha_{i1})|X_i = X^k]$ will be equal to the width in the nonparametric case, $\Delta^k_u - \Delta^k_\ell = 1$. More generally, in the panel binary choice model of equation (9), when there are no time effects, every coefficient of X_{it} varies across individuals, and X_{it} is fully saturated (e.g. is a complete set of dummies, one for every possible value of X_{it}), the semiparametric bounds will equal the nonparametric ones.

In the binary-regressor case the width of the overall bound on the ATE is given by

$$\Delta_u - \Delta_\ell = \bar{\mathcal{P}}(0)(\Delta^1_u - \Delta^1_\ell) + \bar{\mathcal{P}}(1)(\Delta^2_u - \Delta^2_\ell). \tag{12}$$

where we assume $X^1 = (0, ..., 0)'$ and $X^2 = (1, ..., 1)'$. The semiparametric bounds will be smaller than the nonparametric bounds if and only if $\Delta^1_u - \Delta^1_\ell$ or $\Delta^2_u - \Delta^2_\ell$ are smaller than the nonparametric values of 1. This decomposition also shows that the semiparametric identification rate will be determined by the nonparametric rate, which governs how fast $\bar{\mathcal{P}}(0)$ and $\bar{\mathcal{P}}(1)$ shrink, and the rate that the conditional bounds converge. When the slope does not vary across individuals it turns out that the conditional bounds can converge very rapidly. The following result shows this in static and dynamic, binary-choice logit models with binary regressors.

Theorem 6: *Suppose that* $H(v) = e^v/(1 + e^v)$, $\Delta(\beta, \alpha) = H(\beta + \alpha) - H(\alpha)$, *and either equation (8) is satisfied with,* $X_{it} \in \{0, 1\}$, *and* $X^1 = (0, ..., 0)'$ *and* $X^2 = (1, ..., 1)'$, *or equation (10) is satisfied with* $k \in \{1, 2\}$. *Then there are* $C > 0$ *and* $1 > \varepsilon > 0$ *such that*

$$\Delta^k_u - \Delta^k_\ell \leq C(1 - \varepsilon)^T, k = 1, 2.$$

This fast rate occurs because T conditional moments of a one-to-one transformation of α_i are identified from probabilities of various Y values, and these moments lead to a fast approximation of the conditional ATE. For example, $\Pr(Y_i = (1, ..., 1)'|X_i = X^1) = E[H(\alpha_i)^T|X_i = X^1]$, and other conditional moments of $H(\alpha_i)$ can be similarly identified. For the logit $H(\alpha)$, identification of these moments leads to fast approximation of $\Delta^1 = E[H(\beta^* + \alpha_i) - H(\alpha_i)|X_i = X^1]$ and hence to fast shrinkage of the conditional bound.

From equation (12) we see that the semiparametric identification rate in this example will be at least exponential, and may be even faster, depending on the nonparametric rate. This result illustrates how imposing a single, additive individual effect can speed up the identification rate. We expect that this type of improvement will extend beyond the logit model with binary regressors.

8 Computation of Semiparametric 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 the unknown parameters and the nonlinearity of the probabilities in those parameters. 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 7: *If Assumptions 5 and 8 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 $\int \mathcal{L}_j^k(\alpha, \beta)\, dF_k^J(\alpha) = \int \mathcal{L}_j^k(\alpha, \beta) dF_k(\alpha)$ $(j = 1, ..., J)$. 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 the 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 7 would imply 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. It is interesting that the model takes a discrete mixture form, although the finite-dimensional nature of the model is expected because the data have finite support.

Although the individual-effect distribution can be taken to be finite dimensional, the 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 a 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 $\mathcal{P}^k / \mathcal{P}_j^k$, for $\mathcal{P}^k = \Pr(X_i \in \mathcal{X}^k)$, 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 and ensures 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

$$B(M) = \{\beta : \tilde{T}_\lambda(\beta, M) \leq \epsilon_M\}, \epsilon_M > 0.$$

The use of ϵ_M here in allowing a range of values of the objective function is analogous to Manski and Tamer's (2002) estimation method. A positive ϵ_M ensures that the set sequence $(B(M))_{M=1}^{\infty}$ is lower hemi-continuous and that $B(M)$ need not be smaller than the identified set, even though the individual effect distributions are restricted by fixing their support points for each 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 $X \in \mathcal{X}^k$) 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 $\tilde{\pi}(\beta) \in \arg\min_{\pi \in \mathcal{S}_M^K} T_\lambda(\beta, \pi, M)$, $\tilde{S}^k(\beta) = \{\pi^k : P_j^k(\beta, \pi, M) = P_j^k(\beta, \tilde{\pi}(\beta), M),\ j = 1, ..., J\}$, 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 the Supplementary Material. The semiparametric bounds for the ATEs (SP-bounds) are tighter than the nonparametric bounds and shrink very fast with T. In the Supplementary Material we report similar results for the logit, including nonidentification of the ATEs, except that β^* is identified, as is well known. Honore and Tamer (2006) also found tight bounds for the coefficient of a dynamic model.

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 9: *(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* $(\alpha, \beta), (\tilde{\alpha}, \tilde{\beta}) \in \Upsilon \times \mathbb{B}$, $\left| \mathcal{L}_j^k\left(\tilde{\alpha}, \tilde{\beta}\right) - \mathcal{L}_j^k(\alpha, \beta) \right| \leq C[d(\tilde{\alpha}, \alpha) + \left\| \tilde{\beta} - \beta \right\|]$*; and v)* $\Delta(\alpha, \beta)$ *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 equation (8) 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 in the following result.

LEMMA 8: *If Assumptions 5 and 8 and equation (8) are satisfied, where $H(v)$ is strictly monotonic on $\Re$ with bounded continuous derivative, 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 9 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 9: *If Assumptions 5, 8, and 9 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 and Inference

Under Assumptions 5 and 8 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 $\Pi = (\mathcal{P}_X', \mathcal{P}')'$, where $\mathcal{P}_X = (\mathcal{P}^k, k=1,...,K)'$ and $\mathcal{P} = (\mathcal{P}_j^k, j=1,...,J, k=1,...,K)'$, and the empirical estimate is $\hat{\Pi} = (\hat{P}_X', \hat{P}')'$, where $\hat{P}_X = (\hat{P}^k, k=1,...,K)'$ and $\hat{P} = (\hat{P}_j^k, j=1,...,J, k=1,...,K)'$.

The estimation method is like the computational one in using linear-in-parameters approximations to the probabilities. Here we describe the estimation method and give a consistency result, and in the Supplementary Material we provide the implementation details. We follow the same steps as the computational one except that we use estimated weights $\hat{w}_j^k$ and estimated probabilities $\hat{P}_j^k$. Let $\hat{M}$ be 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

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

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

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

This approach to estimation (and computation) can be easily modified 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 a model 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 following is a consistency result.

Theorem 10: *If Assumptions 5, 8, and 9 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 9. The reason for this is that 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.

Turning now to the inference for the semiparametric models, we note that it is rather challenging. The estimators of 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. We also give an alternative inference method based on a modified projection in the Supplementary Material.

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

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

where P^* denotes the projection of P onto the model space $\Xi = \{P : \exists \beta \in \mathbb{B} \text{ with } \mathcal{F}_k(\beta, P) \neq \varnothing, \forall k = 1, ..., K\}$, i.e.

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

and $B^*(P)$ is the corresponding projection for the identified set of the parameter, i.e.

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

Alternatively, θ^* can be an upper (or lower) bound on a scalar functional $c'\beta^*$ of the parameter β^*. Then we define

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

In both cases 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; see the Supplementary Material on the modified projection method for further details. Under correct specification, our inference targets the parameters of interest in the true 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

$$CR_{1-\gamma}(\mathcal{P}) = \{P \in S_J^K : W(P, \hat{P}) \leq c_{1-\gamma}(\chi^2_{K(J-1)})\}, \tag{13}$$

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

$$\underline{G}_n^{-1}(\alpha, \mathcal{P}) / \overline{G}_n^{-1}(\alpha, \mathcal{P}) = \inf / \sup_{P \in CR_{1-\gamma}(\mathcal{P})} G_n^{-1}(\alpha, P), \tag{14}$$

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 ($\alpha_1 = \alpha_2 = \alpha/2$).

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 the data-generating process.

Assumption 10: $\Pi \in \mathbb{P} = \{(P_X, P) : P^k > \varepsilon, P_j^k > \varepsilon; j = 1, ..., J, k = 1, ..., K\}$ for some $\varepsilon > 0$.

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

Theorem 11: *If Assumptions 5, 8, and 9 are satisfied then for any sequence of data-generating process* $\Pi = \Pi_n$ *satisfying Assumption 10,*

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

In practice, we use the following simulation approach to compute the confidence intervals.

Algorithm: Perturbed Bootstrap

1. *Draw a potential DGP* $P_r = (P'_{r1}, ..., P'_{rK})$, *where* $P_{rk} \sim \mathcal{M}(n\hat{P}^k, (\hat{P}^k_1, ..., \hat{P}^k_J))/(n\hat{P}^k)$ *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 in equation (13), 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* $CR_{1-\alpha-\gamma}(\theta^*) = [\underline{\theta}, \overline{\theta}]$, *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* $\alpha_1 + \alpha_2 = \alpha$.

10 Empirical Examples

We illustrate the estimation and inference results with two empirical examples. One estimates identified effects and calculates 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.

10.1 Union Premium

We revisit the empirical question of how unions impact 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 differences in unobservables 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. Also, Angrist and Newey (1991) reject the hypothesis that the unobserved heterogeneity acts solely in an additive fashion, motivating the need to control for more general unobserved heterogeneity. Card (1996) found differences in the union and selection effect across skill levels. Here we account fully for differences across individuals in the union effect while allowing correlation of that effect with union status, thus accounting for selection. Recently Frandsen (2011) focused on quantile union effects using a regression-discontinuity design that estimates union effects for those near a union election discontinuity rather than for those whose union status changes. We find a flatter quantile profile than he does, consistent with his theoretical results that suggest a flatter profile away from the discontinuity.

We use data from the National Longitudinal Survey (Youth Sample). The sample consists of full-time, young, working males, 20 to 29 years 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 armed forces or were students any year, or who reported too high (more than $500 per hour) or too low (less than $1 per hour) wages. The final sample includes 2,065 men followed over 8 years. 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 by a collective bargaining agreement. 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.

We begin by imposing the stationarity condition that income with and without union membership has the same distribution in each time period but also will allow for location and scale time effects. It turns out that time effects are not important in this data. Some covariates are also allowed for since time-invariant covariates are absorbed in the individual effects. Insensitivity to time effects also suggests that time-varying covariates may not be important though a fuller exploration would be useful. For brevity we focus on the case without covariates.

In our analysis, we focus on estimating the union quantile effect for the subpopulations of workers that ever became unionized within the sample (47% of the sample) or that were unionized in the first year (20% of the sample). For these subpopulations, the union effect is not point-identified, since there are 13% of the ever-unionized workers that always stayed unionized between 1986 and 1993, and there are 32% of the workers unionized in 1986 that remained unionized until 1993. 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 their union status at least once within the sample. For this subpopulation, the effect is point-identified in the static model, 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 1).

All the results are reported in Table 1 and Figure 4. Table 1 assesses the plausibility of the time-homogeneity assumption by comparing moments and quantiles of the cross-sectional distributions of log-wages across years for workers that do not change union status. Under time homogeneity, these cross sectional distributions should remain time invariant in the static model. In the table we observe distributional changes across years, but most of the variation can be captured by additive location effects for both always-unionized and never-unionized workers.

Panels A and B of fig. 4 present the estimates of the union effect in the static model for the subpopulation of workers who change their union status at least once within the sample. In panel A we compare our panel data estimates of quantile effects that control for individual heterogeneity with pooled estimates that do not control for individual heterogeneity. In the pooled estimates, we see that the quantile effect of union membership 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 unobserved skills, we see that the quantile effects of union membership become very flat across the quantile indices. Thus, by controlling for individual heterogeneity, we have eliminated the selection effect. Panel B shows that the results are not sensitive to the inclusion of location and scale effects.

Panel C presents estimated bounds on the union effect for the subpopulation of workers that ever became unionized within the sample using the static model with time effects. The bounds are informative, and show that the effect is positive for most of the quantile indices. The panel also shows bounds obtained using the assumption of monotonic and positive union effect on earnings described in the Supplementary Material. These bounds are also informative, and in fact are substantially tighter than the bounds obtained without the monotonicity assumption. Panel D presents similar bounds on the union effect for the subpopulation of workers unionized in the first period using the dynamic model. The bounds in this case are not informative, even after imposing monotonicity.

All the panels include 90% uniform confidence bands for the quantile union effects constructed by bootstrap with 200 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.

10.2 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), 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 2. The labor force participation variable (LFP) is an indicator that takes the value one if the woman's 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 children younger than 3 years. We focus on very young, preschool children as most empirical studies find that their presences have the strongest impact on the mother's participation decision. LFP is stable across the years considered, whereas $kids$ is decreasing. 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

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

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 on the effects. For the semiparametric bounds, we use the method 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 $\hat{M} = 23$ support points at $\{-\infty, -4, -3.6, ..., 3.6, 4, \infty\}$ for the parameter β^*, and with $\hat{M} = 163$ 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.

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

Table 3 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 9 and the Supplementary Material. 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. For the modified projection method (95% MP), the confidence interval for $\mathcal{P}$ in the first stage is approximated by 5,000 DGPs drawn from the empirical multinomial distributions that pass the goodness-of-fit test. Together the modified projection and the perturbed bootstrap took several days to compute on a personal computer. We also include confidence intervals obtained by a canonical projection method (95% CP) less robust to model misspecification than the modified projection method, that intersects a 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 important efficiency losses relative to the fixed-effects estimates. Another salient feature of the results is that the misspecification problem of the canonical projection method clearly arises in this application. Thus, this procedure gives empty confidence regions for the panel with 3 periods. The perturbed bootstrap and modified projection methods produce similar (non-empty) confidence regions for the model parameters and ATEs.

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.

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] Angrist, J. D. and W.K. Newey (1991), "Over-Identification Tests in Earnings Functions with Fixed Effects," with J.A. Angrist, *Journal of Business and Economic Statistics* 9, 317-323.

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

[6] Bester, A.C., and C. Hansen (2008), "Flexible Correlated Random Effects Estimation in Panel Models with Unobserved Heterogeneity," working paper, GSB, University of Chicago.

[7] Bhargava A., and J.D. Sargan (1983), "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods," *Econometrica* 51, 1635—1660.

[8] 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.

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

[10] Browning, M. and J. Carro (2009), "Dynamic Binary Outcome Models with Maximal Heterogeneity," working paper, Oxford.

[11] Card, D. (1996), The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica* 64, 957-979.

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

[13] 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.

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

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

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

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

[18] Chamberlain, G. (1994), ” “Quantile Regression, Censoring, and the Structure of Wages,” in C. Sims, ed., *Advances in Econometrics: Sixth World Congress, Volume I*, Cambridge: Cambridge University Press.

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

[20] 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.

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

[22] Chernozhukov, V., J.Hahn, and W.K.Newey (2004), “Bound Analysis in Panel Models with Correlated Random Effects,” *unpublished manuscript*, http://econ-www.mit.edu/files/5239.

[23] Chernozhukov, V., Fernandez-Val, I., Hahn, J., and W.K.Newey (2007), “Identification and estimation of marginal effects in nonlinear panel models,” *unpublished manuscript*, MIT.

[24] Chernozhukov, V., Fernandez-Val, I., Hahn, J., and W.K.Newey (2012), “Supplemental Material for Average and Quantile Effects in Nonseparable Panel Models,” *unpublished manuscript*, MIT.

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

[26] 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.

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

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

[29] Fernandez-Val, I. and J. Lee (2010), "Panel Data Models with Nonadditive Unobserved Heterogeneity: Estimation and Inference," working paper, Boston University.

[30] Frandsen, B. (2011), "Why Unions Still Matter: The Effects of Unionization on the Distribution of Employee Earnings," working paper, MIT.

[31] Graham,B.W. J. Hahn, and J.L. Powell (2009), "A quantile correlated random coefficient panel data model" working paper, Berkeley.

[32] Graham, B.W. and J.L. Powell (2011), "Identification and Estimation of Average Partial Effects in 'Irregular' Correlated Random Coefficient Panel Data Models", working paper, Berkeley.

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

[34] 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.

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

[36] 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.

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

[38] 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.

[39] Hoderlein, S. and H. White (2011), "Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects," working paper, Boston College.

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

[41] Honore, B.E. and E. Tamer (2003), "Bounds on Parameters in Dynamic Discrete Choice Models," working paper.

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

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

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

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

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

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

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

[49] Romano, J. P., and M. Wolf (2000), "Finite Sample Nonparametric Inference and Large Sample Efficiency," *Annals of Statistics*, 28(3), 756–778.

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

[51] 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.

[52] 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.

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

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

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

Table 1: Moments and quantiles of log wages by year and union sequence

Year	Mean	Std. Dev.	Q 10%	Q 25%	Q 50%	Q 75%	Q 90%
			Never unionized (53 %)				
1986	1.67	0.47	1.11	1.33	1.65	1.99	2.32
1987	1.76	0.52	1.08	1.42	1.75	2.11	2.40
1988	1.83	0.51	1.20	1.50	1.82	2.17	2.46
1989	1.85	0.51	1.24	1.52	1.84	2.19	2.49
1990	1.90	0.53	1.22	1.53	1.89	2.23	2.52
1991	1.90	0.51	1.28	1.55	1.87	2.22	2.55
1992	1.92	0.52	1.28	1.55	1.89	2.28	2.58
1993	1.96	0.53	1.30	1.59	1.94	2.30	2.60
			Always unionized (6 %)				
1986	2.04	0.31	1.67	1.85	2.05	2.24	2.44
1987	2.10	0.33	1.68	1.91	2.12	2.30	2.48
1988	2.17	0.36	1.71	1.91	2.20	2.37	2.53
1989	2.18	0.30	1.79	1.97	2.20	2.36	2.56
1990	2.21	0.29	1.80	2.01	2.24	2.40	2.55
1991	2.21	0.29	1.85	2.00	2.22	2.37	2.55
1992	2.22	0.38	1.84	2.01	2.20	2.42	2.59
1993	2.25	0.29	1.85	2.03	2.25	2.49	2.58

Source: NLSY79 1986-1993, 2,065 men.

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

Variable	Mean	Changes (%)
LFP1990	0.75	
LFP1992	0.74	0.17
LFP1994	0.75	0.28
kids1990	0.38	
kids1992	0.35	0.31
kids1994	0.28	0.51

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

Table 3: 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	[-.412, -.408]	-.88	-.51	
(95% N)			(-1.11, -.46)	(-.67, -.05)	(-.70, -.08)		(-1.24, -.52)	(-.86, -.16)	
(95% CP)		(-.75, .02)				(-.85, .03)			
(95% MP+)		(-.84, .01)				(-.92, .04)			
(95% PB^)		(-.87, .12)				(-1.10, .09)			
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^)		(-.18, .02)				(-.21, .02)			
					T = 3				
β*		-.42	-.71	-.46	-.46	[-.461, -.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^)		(-.69, -.01)				(-.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^)		(-.12, -.00)				(-.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 5,000 DGPs. ^Based on 100 DGP's and 200 simulations for each DGP.

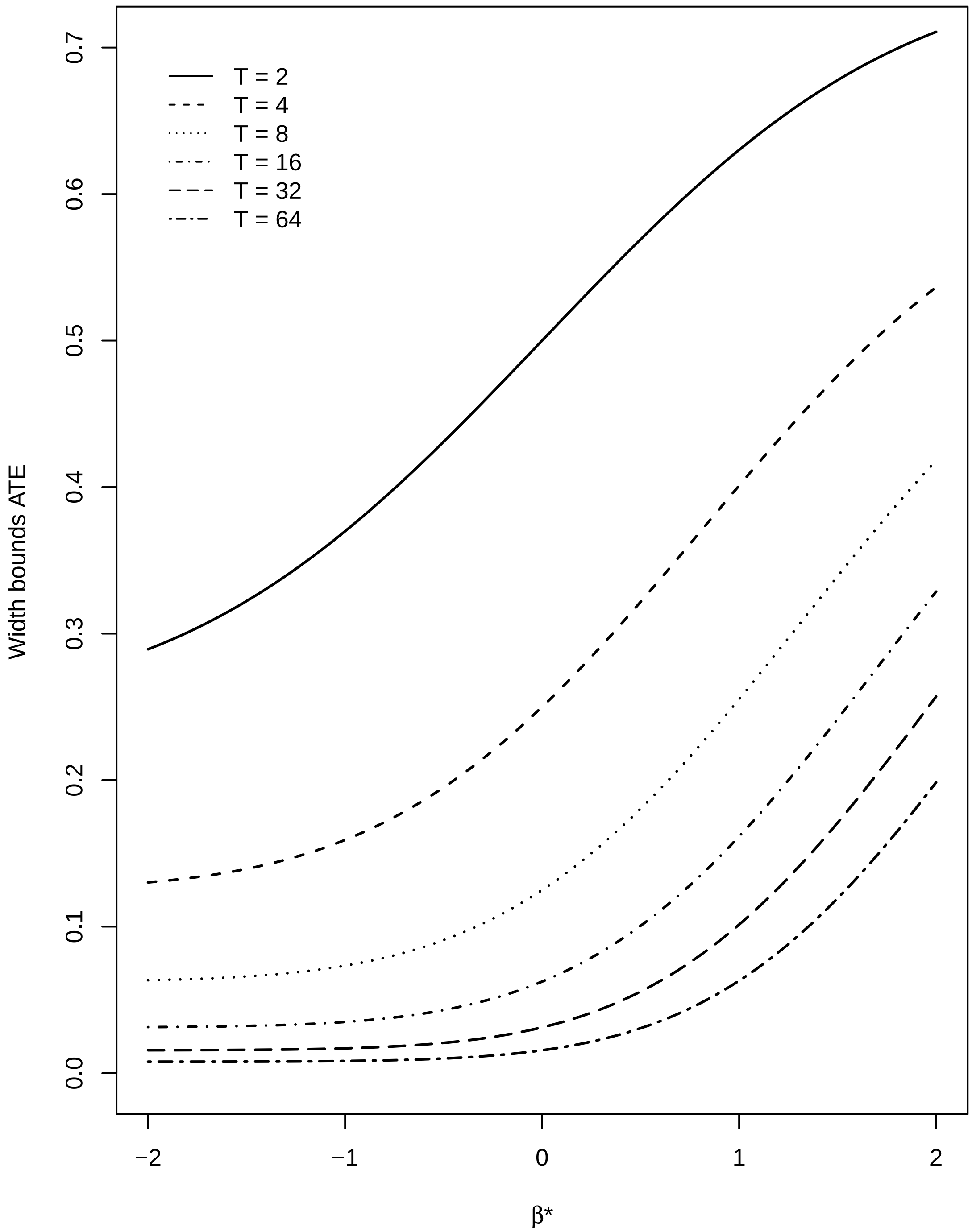

Figure 1: Width of nonparametric bounds for the ATE in dynamic binary choice probit models with $Y_{it} = 1(\beta^* Y_{i,t-1} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\Pr(Y_{i0} = 1) = .5$, $\beta^* \in [-2, 2]$, and $T \in \{2, 4, 8, 16, 32, 64\}$.

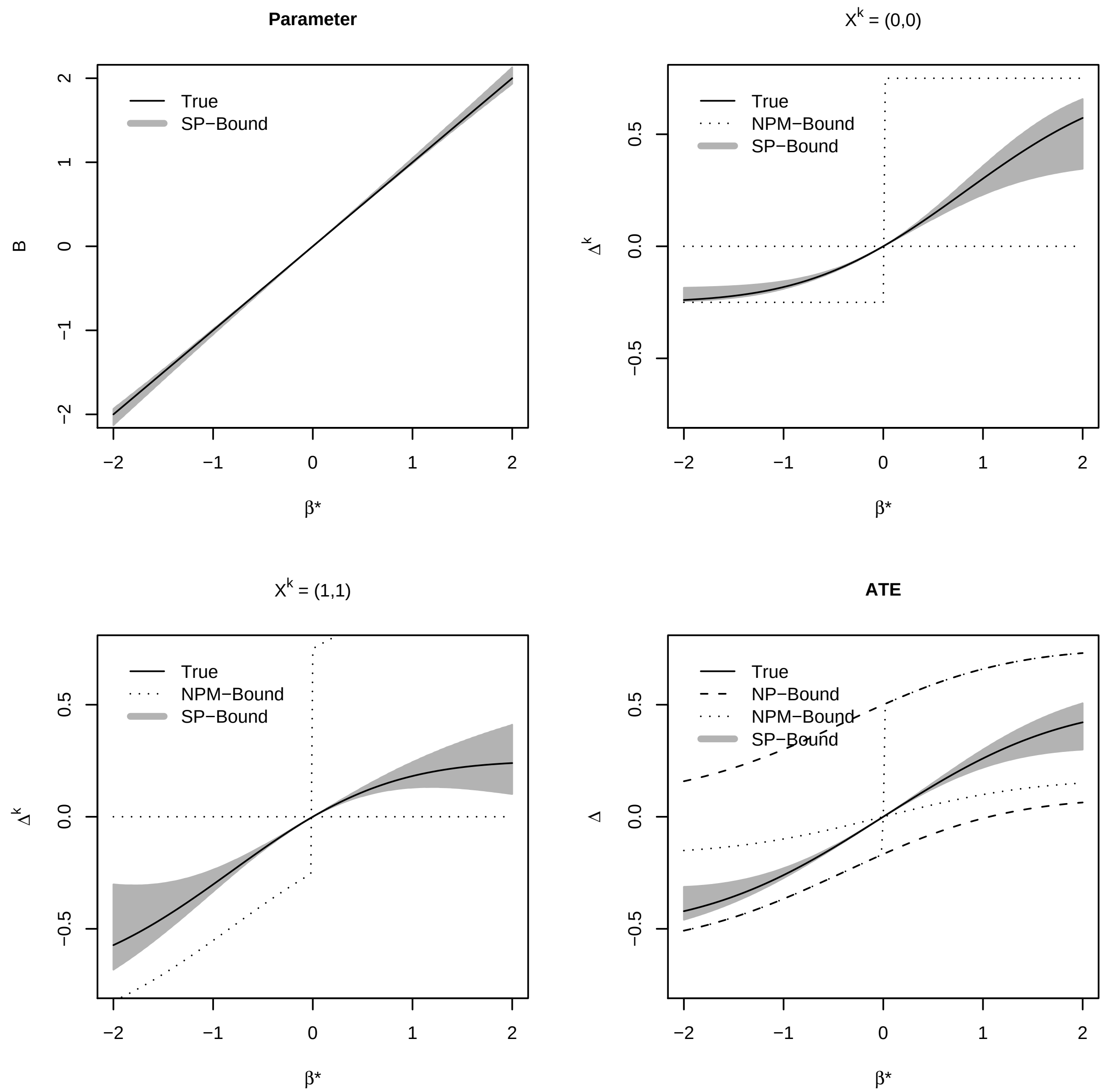

Figure 2: Identified set for parameter and ATEs in binary choice probit models with $Y_{it} = 1(\beta^* X_{it} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim N(0,1)$, $X_{it} = 1(\alpha_i \geq \eta_{it})$, $\eta_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\beta^* \in [-2,2]$, and $T = 2$.

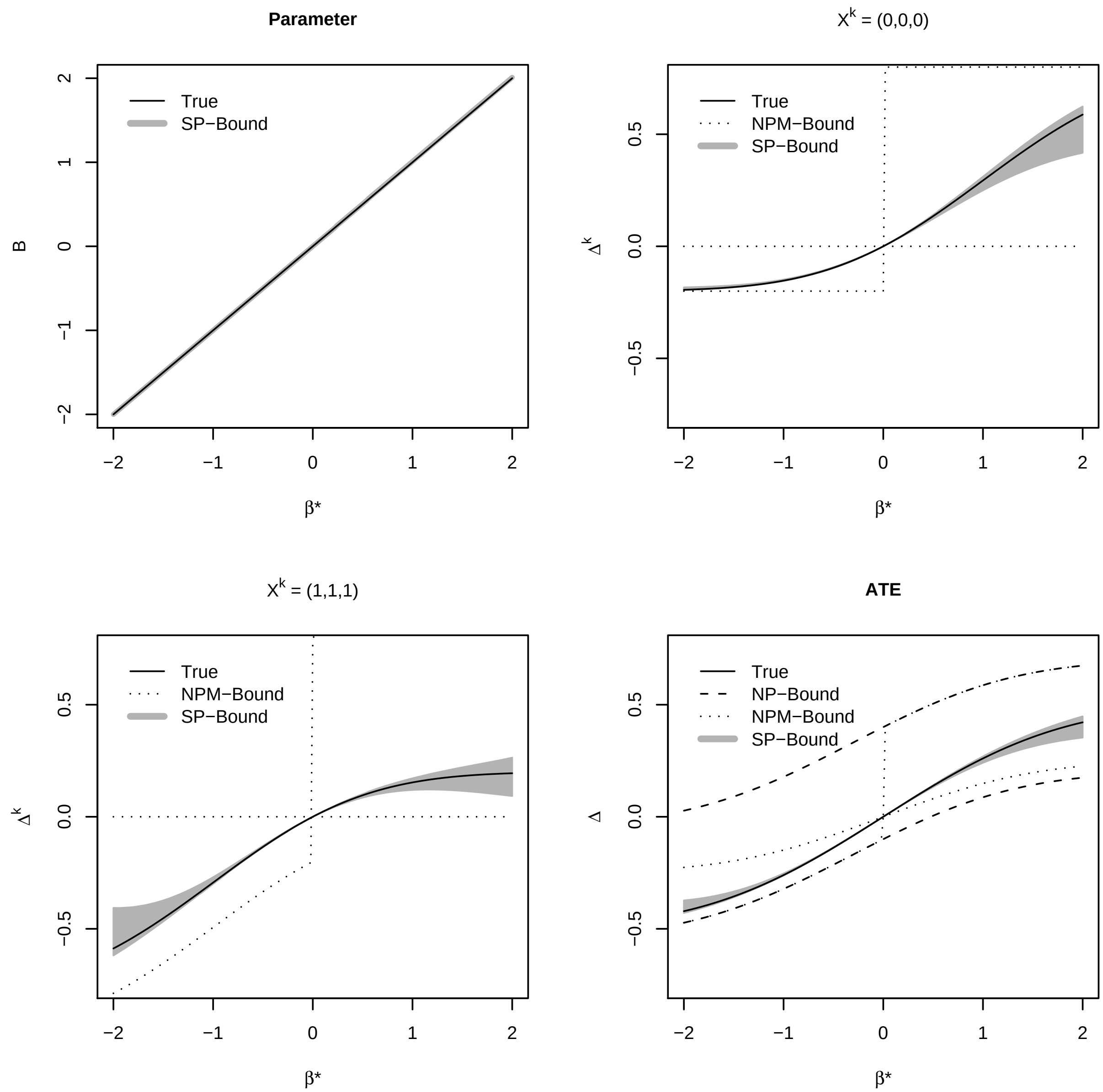

Figure 3: Identified set for parameter and ATEs in binary choice probit models with $Y_{it} = 1(\beta^* X_{it} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim N(0,1)$, $X_{it} = 1(\alpha_i \geq \eta_{it})$, $\eta_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\beta^* \in [-2,2]$, and $T = 3$.

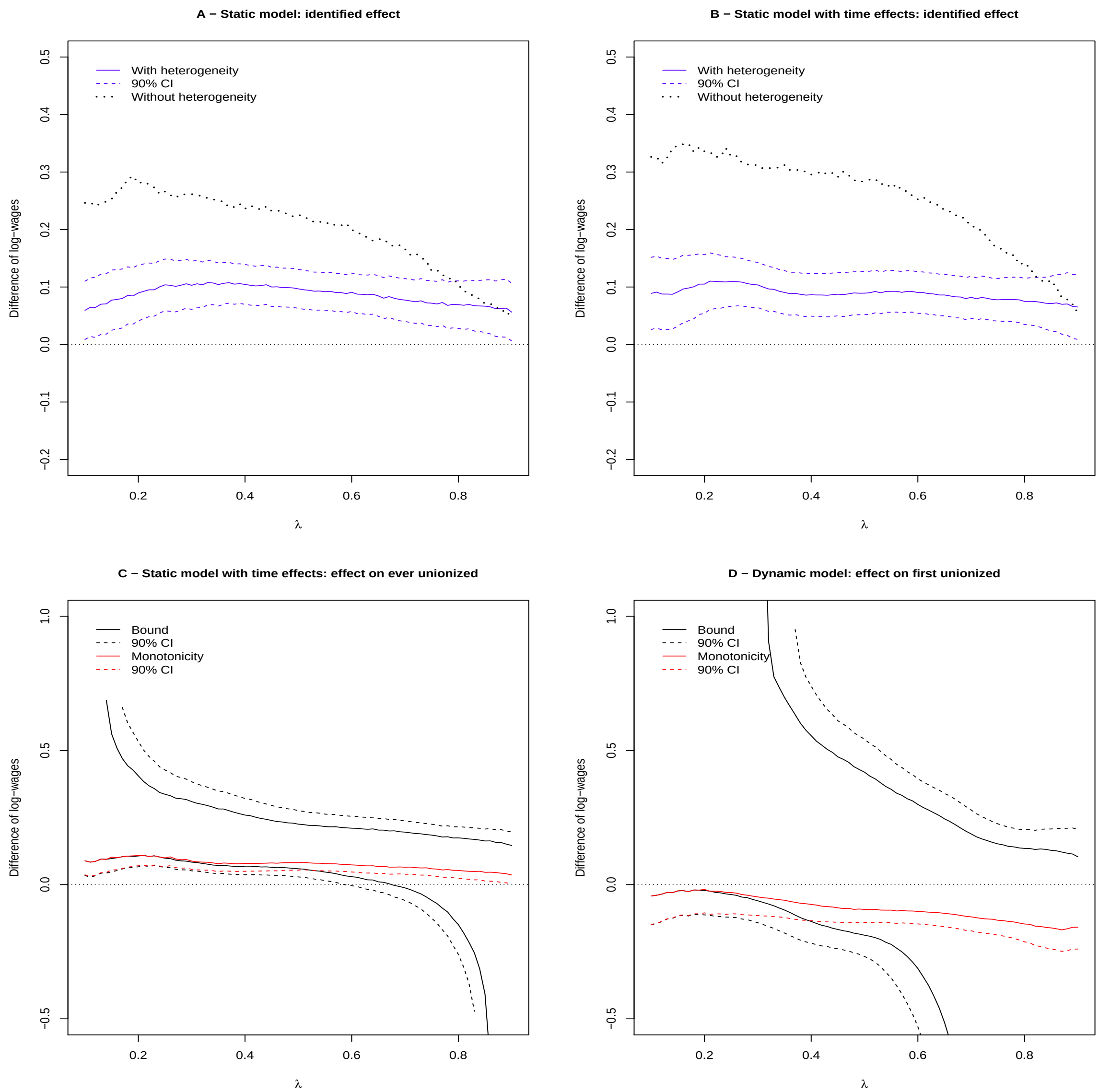

Figure 4: Quantile union effects for male workers. Panel A displays point and interval estimates of the identified quantile union effects in the static model with and without accounting for individual heterogeneity. Panel B displays point and interval estimates of the identified quantile union effects in the static model with location and scale time effects, averaged across time periods with and without accounting for individual heterogeneity. Panel C displays point and interval estimates of the bounds for the quantile effect on the ever unionized in the static model with time effects, with and without imposing monotonicity. Panel D displays point and interval estimates of the bounds for the quantile effect on the unionized in the first period in the dynamic model, with and without imposing monotonicity. Estimates based on NLSY79 for the years 1986–1993. 90% confidence intervals obtained by bootstrap with 200 repetitions.

Supplemental Material for Average and Quantile Effects in Nonseparable Panel Models

Victor Chernozhukov, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey

A1 Introduction

In this supplemental material we provide omitted discussions, results, and proofs by Section in the same order they are referred to in the paper. Let w.p.a.1 denote "with probability approaching one" and C denote a generic constant that may be different in different uses.

A2 Supplements to Section 2

We begin with the omitted discussion and results referred to in Section 2 of the paper. These concern the general, nonseparable model of Assumptions 1 - 3 and apply whether or not the regressors are discrete.

A2.1 Time homogeneity in the linear model

We will first show that Assumption 2 is a natural generalization of the following linear model:

$$Y_{it} = X_{it}'\beta_0 + \alpha_i + \varepsilon_{it}, E[X_{is}\varepsilon_{it}] = 0 \text{ for all } s \text{ and } t. \tag{15}$$

This is a standard linear model that leads to consistency of the within and other estimators. Let $\bar{E}(\cdot|X_i)$ denote the linear projection on $vec(X_i)$, as in Chamberlain (1982).

Theorem A1: *Suppose that Y_i and X_i have finite second moments. Then equation (15) is satisfied if and only if there is $\tilde{\varepsilon}_{it}$ with*

$$Y_{it} = X_{it}'\beta_0 + \tilde{\varepsilon}_{it},\ \bar{E}(\tilde{\varepsilon}_{it}|X_i) = \bar{E}(\tilde{\varepsilon}_{i1}|X_i), (t = 2, ..., T). \tag{16}$$

Proof: If eq. (*15*) is satisfied let $\tilde{\varepsilon}_{it} = \alpha_i + \varepsilon_{it}$. By orthogonality of ε_{it} with X_{is} for all s and t we have $\bar{E}(\varepsilon_{it}|X_i) = 0$ for all t, so that

$$\bar{E}(\tilde{\varepsilon}_{it}|X_i) = \bar{E}(\alpha_i|X_i) + \bar{E}(\varepsilon_{it}|X_i) = \bar{E}(\alpha_i|X_i) = \bar{E}(\alpha_i|X_i) + \bar{E}(\varepsilon_{i1}|X_i) = \bar{E}(\tilde{\varepsilon}_{i1}|X_i).$$

Now suppose eq. (16) is satisfied. Let $\alpha_i = \bar{E}[\tilde{\varepsilon}_{i1}|X_i]$ and $\varepsilon_{it} = \tilde{\varepsilon}_{it} - \alpha_i$. Then $Y_{it} = X_{it}'\beta_0 + \alpha_i + \varepsilon_{it}$ by construction and

$$E[X_{is}\varepsilon_{it}] = E[X_{is}(\tilde{\varepsilon}_{it} - \bar{E}[\tilde{\varepsilon}_{i1}|X_i])] = E[X_{is}(\tilde{\varepsilon}_{it} - \bar{E}[\tilde{\varepsilon}_{it}|X_i])] = 0,$$

where the second equality follows by $\bar{E}(\tilde{\varepsilon}_{it}|X_i) = \bar{E}(\tilde{\varepsilon}_{i1}|X_i)$ and the third quality by orthogonality of each element of X_i with the projection residual. Q.E.D.

This result shows that the standard linear model of equation (15) is equivalent to the model of equation (16). The second model is one that satisfies a time homogeneity condition analogous to Assumption 2. In equation (16) the linear projection of the disturbance on the elements of X_i is time invariant. What Assumption 2 does is strengthen this to time invariance of the conditional distribution. This strengthening seems like a natural thing to do when moving from a linear model to a nonlinear, nonseparable model.

A2.2 Relationship between static and dynamic models

We next show that the static model is nested within the dynamic model.

Theorem A2: *If Assumptions 1 and 2 are satisfied then Assumptions 1 and 3 are satisfied.*

Proof: Note that Assumptions 1 and 2 allow some flexibility in the definition of α_i, because Assumption 1 just specifies that there exists α_i with $Y_{it} = g_0(X_{it}, \alpha_i, \varepsilon_{it})$. This equation continues to hold if more variables are added to α_i. Furthermore, we can add any function of X_i to α_i without changing Assumption 2. Let $\tilde{\alpha}_i = (\alpha_i, X_i)$. Then Assumptions 1 and 2 are also satisfied for this $\tilde{\alpha}_i$. Furthermore, since $X_{it}, ..., X_{i1}$ are included in $\tilde{\alpha}$ and Assumption 2 for the original α_i implies that $\varepsilon_{it}|\tilde{\alpha}_i \overset{d}{=} \varepsilon_{i1}|\tilde{\alpha}_i$ we have

$$\varepsilon_{it}|X_{it}, ..., X_{i1}, \tilde{\alpha}_i \overset{d}{=} \varepsilon_{it}|\tilde{\alpha}_i \overset{d}{=} \varepsilon_{i1}|\tilde{\alpha}_i \overset{d}{=} \varepsilon_{i1}|X_{i1}, \tilde{\alpha}_i.$$

Thus we see that Assumptions 1 and 2 imply existence of $\alpha_i = \tilde{\alpha}_i$ such that Assumptions 1 and 3 are also satisfied. That is, Assumptions 1 and 2 imply Assumptions 1 and 3. Q.E.D.

A2.3 Relationship between nonseparable models and conditional mean models

Next we show that the nonseparable models given here imply conditional mean models where the ATE is also the conditional mean ATE.

THEOREM A3: *Suppose that Assumption 1 is satisfied and* $E[|g_0(x, \alpha_i, \varepsilon_{it})|] < \infty$ *for all* x. *If Assumption 2 is satisfied then for* $\tilde{\alpha}_i = X_i$ *and* $m_0(x, \tilde{\alpha}) = \int g_0(x, \alpha, \varepsilon)dF(\alpha, \varepsilon|\tilde{\alpha})$,

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

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

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

Proof: By Assumption 2, for $\tilde{\alpha} = X$ and $m_0(x, \tilde{\alpha}) = \int g_0(x, \alpha, \varepsilon)dF(\alpha, \varepsilon|X)$ we have

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

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

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

It may be helpful to explain this result and relate it to Chamberlain (1982). First, it should be noted that Assumptions 1 and 2 only assume the existence of some α_i such that the conditions are satisfied. Thus, we are free to choose α_i in whatever way is convenient. A convenient choice for Theorem A3 turns out to be $\tilde{\alpha}_i = X_i$, where we use the $\tilde{\alpha}_i$ notation to distinguish this time invariant effect from the one in Assumptions 1 and 2. Note then that the first conclusion implies that for $m_0(x, X) = \int g(x, \alpha, \varepsilon)dF(\alpha, \varepsilon|X)$,

$$E[Y_{it}|X_i] = m_0(X_{it}, X_i). \tag{17}$$

This statement has no content for any one time period, because the effect of X_{it} in the first argument of $m(X_{it}, X_i)$ is indistinguishable from the effect of X_{it} that appears in the second argument. However, for multiple time periods it does have content, because $m_0(x, X)$ is time invariant. Equation (17) implies that the effect of changing X_{it} on $E[Y_{it}|X_i]$ will be different than the effect on $E[Y_{is}|X_i]$ for $s \neq t$. Furthermore, this form leads directly to identification of conditional mean ATE conditioned on X_i. For any X_i where $X_{it} = x^b$ and $X_{is} = x^a$ for some t and s,

$$E[Y_{is} - Y_{it}|X_i] = m_0(x^a, X_i) - m_0(x^b, X_i),$$

that is a conditional mean ATE given X_i.

It may also help to think of $m(X_{it}, X_i)$ as a nonlinear version of Chamberlain's (1982) multivariate regression for panel data. In the linear model of equation (15), for $\bar{E}[\alpha_i|X_i] = \pi' vec(X_i)$ we have

$$\bar{E}[Y_{it}|X_i] = X_{it}'\beta_0 + \pi' vec(X_i) = \bar{m}(X_{it}, X_i),\ \bar{m}(x, X) = x'\beta_0 + \pi' vec(X).$$

For a single time period β_0 is indistinguishable from coefficients in π, but multiple time periods can be used to identify β_0 from these regressions. Equation (17) is like this except it is jointly nonlinear in its first and second arguments.

A3 Supplements to Section 3

A3.1 Auxiliary results

We turn now to identification and estimation with discrete regressors in the static case. Here we use the idea that "time is an instrument" or "time is randomly assigned." This allows us to vary the time period so as to match x with X_{it} and achieve identification.

The following Lemma applies this idea to obtain specific results. Let $g_{it}(x) = g_0(x, \alpha_i, \varepsilon_{it})$.

Lemma A4: *If Assumptions 1 and 2 are satisfied then*

$$E[\bar{G}_i(y, x)|X_i] = 1(T_i(x) > 0)E[\Phi(\frac{y - g_{i1}(x)}{h})|X_i].$$

If in addition $E[|g_0(x, \alpha_i, \varepsilon_{it})|] < \infty$ *for all* x *then*

$$E[\bar{Y}_i(x)|X_i] = 1(T_i(x) > 0)E[g_{i1}(x)|X_i].$$

.

Proof: By Assumptions 1 and 2,

$$\begin{aligned} E[1(X_{it} &= x)\Phi(\frac{y - Y_{it}}{h})|X_i] = E[1(X_{it} = x)\Phi(\frac{y - g_{it}(x)}{h})|X_i] \\ &= 1(X_{it} = x)E[\Phi(\frac{y - g_{it}(x)}{h})|X_i] = 1(X_{it} = x)E[\Phi(\frac{y - g_{i1}(x)}{h})|X_i]. \end{aligned}$$

Therefore, we have

$$\begin{aligned} E[\bar{G}_i(y, x)|X_i] &= 1(T_i(x) > 0)T_i(x)^{-1}\sum_{t=1}^{T} E[1(X_{it} = x)\Phi(\frac{y - Y_{it}}{h})|X_i] \\ &= 1(T_i(x) > 0)T_i(x)^{-1}\sum_{t=1}^{T} 1(X_{it} = x)E[\Phi(\frac{y - g_{i1}(x)}{h})|X_i] \\ &= 1(T_i(x) > 0)E[\Phi(\frac{y - g_{i1}(x)}{h})|X_i]. \end{aligned}$$

We also have

$$\begin{aligned}E[1(X_{it} &= x)Y_{it}|X_i] = E[1(X_{it}=x)g_{it}(x)|X_i] = 1(X_{it}=x)E[g_{it}(x)|X_i] \\ &= 1(X_{it}=x)E[g_{i1}(x)|X_i]\end{aligned}$$

so the second conclusion follows similarly to the first. Q.E.D.

We can use the previous result to show how δ is identified.

LEMMA A5: *If Assumptions 1 and 2 are satisfied,* $E[|g_0(x,\alpha_i,\varepsilon_{it})|] < \infty$ *for all* x, *and* $\Pr(D_i = 1) > 0$ *then* $\delta = E[D_i\{\bar{Y}_i(x^a) - \bar{Y}_i(x^b)\}]/E[D_i]$.

Proof: Note that $D_i = D_i 1(T_i(x^b) > 0) = D_i 1(T_i(x^a) > 0)$. Therefore, by Lemma A4

$$\begin{aligned}E[D_i\{\bar{Y}_i(x^a) - \bar{Y}_i(x^b)\}|X_i] &= D_i E[\bar{Y}_i(x^a)|X_i] - D_i E[\bar{Y}_i(x^b)|X_i] \\ &= D_i 1(T_i(x^a) > 0)E[g_{i1}(x^a)|X_i] - D_i 1(T_i(x^b) > 0)E[g_{i1}(x^b)|X_i] \\ &= D_i E[g_{i1}(x^a) - g_{i1}(x^b)|X_i] = E[D_i \left\{g_{i1}(x^a) - g_{i1}(x^b)\right\} |X_i]\end{aligned}$$

The conclusion then follows by iterated expectations. *Q.E.D.*

The asymptotic normality of $\hat{\delta}$ and consistency of the asymptotic variance estimator are simple applications of standard theory, as in the following result, that forms a prototype for the asymptotic normality of the nonparametric ATE bounds. Let $P = E[D_i]$.

THEOREM A6: *If Assumptions 1 and 2 are satisfied,* $E[|g_0(x,\alpha_i,\varepsilon_{it})|^2] < \infty$ *for all* x, *and* $\Pr(D_i = 1) > 0$, *then* $\sqrt{n}(\hat{\delta} - \delta) \xrightarrow{d} N(0, V)$ *and* $\sum_{i=1}^n \hat{\psi}_i^2/n \xrightarrow{p} V$, *where* $V = E[\psi_i^2]$ *and* $\psi_i = P^{-1}D_i\left[\bar{Y}_i(x^a) - \bar{Y}_i(x^b) - \delta\right]$.

Proof: Let $d_i = D_i\{\bar{Y}_i(x^a) - \bar{Y}_i(x^b)\}$ so that $\hat{\delta} = \bar{d}/\bar{D}$. By the central limit theorem (CLT), $\bar{d}$ and $\bar{D}$ are root-n consistent for $\mu_d = E[d_i]$ and P. Then by $P > 0$ and $\delta = \mu_d/P$,

$$\begin{aligned}\sqrt{n}(\hat{\delta} - \delta) &= \sqrt{n}(\frac{\bar{d}}{\bar{D}} - \frac{\mu_d}{P}) = \sqrt{n}\bar{D}^{-1}[\bar{d} - \mu_d - \delta(\bar{D} - P)] \\ &= \sqrt{n}P^{-1}[\bar{d} - \mu_d - \delta(\bar{D} - P)] + o_p(1) = \sum_{i=1}^n \psi_i/\sqrt{n} + o_p(1).\end{aligned}$$

The first conclusion then follows by the CLT. For the second conclusion note that

$$\sum_i(\hat{\psi}_i - \psi_i)^2/n \le C(\bar{D}^{-1} - P^{-1})^2\sum_i d_i^2/n + C(\bar{D}^{-1}\hat{\delta} - P^{-1}\delta)^2\sum_i D_i^2/n \xrightarrow{p} 0.$$

Therefore, the second conclusion follows by a standard argument. Q.E.D.

We now give an intermediate result that is useful for showing asymptotic normality for the estimator of the identified quantile treatment effect. This will also serve as a prototype for the proofs of Theorems 2 and 3 in the body of the paper. Let $\hat{G}_1(y,x) = \hat{G}(y,x|D_i = 1)$, $G_1(y,x) = G(y,x|D_i = 1)$, $G_i(y,x) = 1(T_i(x) > 0)T_i(x)^{-1}\sum_{t=1}^T 1(X_{it} = x)1(Y_{it} \le y)$, and $G_1'(y,x) = \partial G_1(y,x)/\partial y$.

Lemma A7: *If Assumption 7 is satisfied with $G_\ell(y,x)$ replaced by $G_1(y,x)$ then for any $0 < \lambda < 1$ and any x, there exists $\hat{q}_\lambda$ with $\hat{G}_1(\hat{q}_\lambda, x) = \lambda$ satisfying*

$$\sqrt{n}(\hat{q}_\lambda - q_\lambda) = -G_1'(q_\lambda, x)^{-1}\frac{1}{\sqrt{n}}P^{-1}\sum_i D_i\left[G_i(q_\lambda, x) - \lambda\right] + o_p(1).$$

Proof: Note that $\hat{G}_1(y,x)$ is strictly monotonic increasing in y and converges to 0 and 1 as y goes to $-\infty$ and ∞ respectively. Therefore there is a unique $\hat{q}_\lambda$ such that $\hat{G}_1(\hat{q}_\lambda, x) = \lambda$. Also, by $G_1(y,x)$ strictly monotonic in y there is a unique q_λ solving $G_1(q_\lambda, x) = \lambda$. By $G_1(y,x)$ strictly monotonic and continuous, it follows that for all $\varepsilon > 0$ small enough,

$$0 < G_1(q_\lambda - \varepsilon, x) < G_1(q_\lambda, x) = \lambda.$$

By $\hat{G}_1(q_\lambda - \varepsilon, x) \xrightarrow{p} G_1(q_\lambda - \varepsilon, x)$ it follows that w.p.a.1, for all $y \le q_\lambda - \varepsilon$

$$\hat{G}_1(y,x) \le \hat{G}_1(q_\lambda - \varepsilon, x) < G_1(q_\lambda, x) = \lambda.$$

Thus, it follows that $\hat{q}_\lambda \ge q_\lambda - \varepsilon$ w.p.a.1. Similarly it follows that $\hat{q}_\lambda \le q_\lambda + \varepsilon$ w.p.a.1. Since ε is arbitrary, we have $\hat{q}_\lambda \xrightarrow{p} q_\lambda$.

Next, note that $G_1(y,x)$ is differentiable in y by Assumption 7, so that $g_{i1}(x)$ is continuously distributed conditional on $D_i = 1$. Thus, $g_{it}(x)$ is also continuously distributed conditional on $D_i = 1$ by Assumption 2. It follows that as $h \longrightarrow 0$, $\Phi(\frac{y - g_{it}(x)}{h}) \longrightarrow 1(g_{it}(x) \le y)$ with probability one. By the dominated convergence theorem this convergence is also in mean-square. Recall that

$$G_i(y,x) = \begin{cases} T_i(x)^{-1}\sum_{t=1}^T 1(X_{it} = x)1(Y_{it} \le y), T_i(x) > 0, \\ 0, T_i(x) = 0. \end{cases}$$

We have $\bar{G}_i(y,x) \longrightarrow G_i(y,x)$ in mean square, so that

$$\sum_{i=1}^n [D_i\bar{G}_i(y,x) - D_iG_i(y,x)]/n \xrightarrow{p} 0,$$
$$\sum_{i=1}^n \{D_i\bar{G}_i(y,x) - E[D_i\bar{G}_i(y,x)] - D_iG_i(y,x) + E[D_iG_i(y,x)]\}/\sqrt{n} \xrightarrow{p} 0.$$

Let $W_i = g_0(x, \alpha_i, \varepsilon_{i1})$ and $f(w)$ and $F(w)$ denote the pdf and CDF of W_i conditional on $D_i = 1$ and $P = E[D_i]$. Note that $\Phi(\frac{y-w}{h})F(w)$ converges to zero as $w \longrightarrow \infty$ and as $w \longrightarrow -\infty$.

Therefore, integration by parts gives

$$\begin{aligned} E[\bar{G}_i(y,x)|D_i &= 1] = \int \Phi(\frac{y-w}{h})f(w)dw = h^{-1}\int \phi(\frac{y-w}{h})F(w)dw \\ &= \int \phi(u)F(y-hu)du = F(y) + (h^2/2)\int \phi(u)F''(y-\bar{h}u)u^2du \\ &= F(y) + o(h^2) = G_1(y,x) + o(h^2), \end{aligned}$$

where the fifth equality follows by an expansion

$$F(y-hu) = F(y) - F'(y)hu + F''(y-\bar{h}u)h^2u^2/2,$$

and $\bar{h}$ can depend on u. Therefore it follows by $E[D_iG_i(q_\lambda,x)] = PG_1(q_\lambda,x) = P\lambda$ that

$$\begin{aligned} \sum_{i=1}^n D_i[\bar{G}_i(q_\lambda,x)-\lambda]/\sqrt{n} &= \sum_{i=1}^n\{D_i\bar{G}_i(q_\lambda,x) - E[D_i\bar{G}_i(q_\lambda,x)]\}/\sqrt{n} \\ &\quad +\sqrt{n}\{E[D_i\bar{G}_i(q_\lambda,x)] - \lambda P\} - \lambda\sum_{i=1}^n(D_i-P)/\sqrt{n} \\ &= \sum_{i=1}^n\{D_iG_i(q_\lambda,x) - E[D_iG_i(q_\lambda,x)]\}/\sqrt{n} + o_p(1) \\ &\quad +O(\sqrt{n}h^2) - \lambda\sum_{i=1}^n(D_i-P)/\sqrt{n} \\ &= \sum_{i=1}^n D_i[G_i(q_\lambda,x)-\lambda]/\sqrt{n} + o_p(1) = O_p(1). \end{aligned}$$

Next, note that from standard uniform convergence of kernel density results, $\hat{G}_1'(y,x)$ converges uniformly in probability to $G_1'(y,x)$, where the "prime" superscript denotes the partial derivative with respect to y. Therefore, for $\bar{q}_\lambda \xrightarrow{p} q_\lambda$, $\hat{G}_1'(\bar{q}_\lambda,x) \xrightarrow{p} G_1'(q_\lambda,x) > 0$, and hence $\hat{G}_1'(\bar{q}_\lambda,x)^{-1} = O_p(1)$. An expansion then gives $\lambda = \hat{G}_1(\hat{q}_\lambda,x) = \hat{G}_1(q_\lambda,x) + \hat{G}_1'(\bar{q}_\lambda,x)(\hat{q}_\lambda - q_\lambda)$. Solving and inverting gives

$$\begin{aligned} \sqrt{n}(\hat{q}_\lambda - q_\lambda) &= -\hat{G}_1'(\bar{q}_\lambda,x)^{-1}\sqrt{n}[\hat{G}_1(q_\lambda,x)-\lambda] \\ &= -\hat{G}_1'(\bar{q}_\lambda,x)^{-1}\left(\sum_{i=1}^n D_i/n\right)^{-1}\sum_{i=1}^n D_i[\bar{G}_i(q_\lambda,x)-\lambda]/\sqrt{n} \\ &= -G_1'(q_\lambda,x)^{-1}P^{-1}\sum_{i=1}^n D_i[G_i(q_\lambda,x)-\lambda]/\sqrt{n} + o_p(1). Q.E.D. \end{aligned}$$

THEOREM A8: *If Assumptions 1, 2, and 7 are satisfied and* $E[D_i] > 0$, *then* $\sqrt{n}(\hat{\delta}_\lambda - \delta_\lambda) \xrightarrow{d} N(0,V_\lambda)$ *and* $\sum_{i=1}^n \hat{\psi}_{\lambda i}^2/n \xrightarrow{p} V_\lambda$, *where* $V_\lambda = E[\psi_{\lambda i}^2]$ *and*

$$\psi_{i\lambda} = -\frac{D_i}{P}\left\{\frac{G_i(q^a,x^a)-\lambda}{G_1'(q^a,x^a)} - \frac{G_i(q^b,x^b)-\lambda}{G_1'(q^b,x^b)}\right\}$$

Proof: By Lemma A7 we have

$$\sqrt{n}(\hat{\delta}_\lambda - \delta_\lambda) = \sum_{i=1}^{n} \psi_{i\lambda}/\sqrt{n} + o_p(1).$$

The CLT gives the first conclusion. Next, note that by $\Phi(v)$ having a bounded derivative,

$$\sum_{i=1}^{n} [\bar{G}_i(\hat{q}^a, x^a) - \bar{G}_i(q^a, x^a)]^2/n \leq Ch^{-1}(\hat{q}^a - q^a) = O_p((h\sqrt{n})^{-1}) \xrightarrow{p} 0.$$

Then by mean square convergence of $\bar{G}_i(q^a, x^a)$ to $G_i(q^a, x^a)$ and the triangle inequality we have $\sum_{i=1}^{n} [\bar{G}_i(\hat{q}^a, x^a) - G_i(q^a, x^a)]^2/n \xrightarrow{p} 0$. The second conclusion then follows similarly to the proof of Theorem A6. *Q.E.D.*

A3.2 Proof of Theorem 1

Note that $\sigma_i^2 > 0$ if and only if $D_i = 1$, so that

$$\sigma_i^2 = D_i \sigma_i^2, X_{it} - \bar{X}_i = D_i(X_{it} - \bar{X}_i).$$

Furthermore, since X_{it} is a dummy variable, the usual difference in means formula for the slope of a regression on a constant and dummy variable gives

$$D_i \frac{\sum_{t=1}^{T} (X_{it} - \bar{X}_i) Y_{it}}{\sum_{t=1}^{T} (X_{it} - \bar{X}_i)^2} = D_i \{\bar{Y}_i(1) - \bar{Y}_i(0)\}.$$

Also, by the Khintchine's weak law of large numbers (LLN),

$$n^{-1}(T-1)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} (X_{it} - \bar{X}_i)^2 = n^{-1} \sum_{i=1}^{n} \sigma_i^2 \xrightarrow{p} E[\sigma_i^2] = E[D_i \sigma_i^2].$$

Furthermore, by LLN

$$\begin{aligned} n^{-1}(T-1)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} (X_{it} - \bar{X}_i) Y_{it} &= n^{-1}(T-1)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} D_i (X_{it} - \bar{X}_i) Y_{it} \\ &= n^{-1} \sum_{i=1}^{n} D_i \sigma_i^2 \{\bar{Y}_i(1) - \bar{Y}_i(0)\} \\ &\xrightarrow{p} E[D_i \sigma_i^2 \{\bar{Y}_i(1) - \bar{Y}_i(0)\}]. \end{aligned}$$

The conclusion then follows by the continuous mapping theorem. Q.E.D.

A4 Supplements to Section 4

Here we include the proof of Theorem 2 as well as bounds that impose monotonicity.

A4.1 Proof of Theorem 2

Let

$$\begin{pmatrix} m_{\ell i} \\ m_{ui} \end{pmatrix} = \begin{pmatrix} \bar{Y}_i(x^a) - \bar{Y}_i(x^b) + B_\ell 1(T_i(x^a) = 0) - B_u 1(T_i(x^b) = 0) \\ \bar{Y}_i(x^a) - \bar{Y}_i(x^b) + B_u 1(T_i(x^a) = 0) - B_\ell 1(T_i(x^b) = 0) \end{pmatrix}.$$

Note that $\hat{\Delta}_\ell = \sum_{i=1}^n m_{\ell i}/n$ and $\hat{\Delta}_u = \sum_{i=1}^n m_{ui}/n$. Then for $\Sigma = Var((m_{\ell i}, m_{ui}))$, $\Delta_\ell = E[m_{\ell i}]$, and $\Delta_u = E[m_{ui}]$ the first and second conclusions follow by standard arguments for a vector of sample means.

Next, note that by Lemma A4 and iterated expectations

$$\begin{aligned} \Delta_\ell &= E[1(T_i(x^a) > 0) g_{i1}(x^a) + B_\ell 1(T_i(x^a) = 0)] \\ &\quad - E[1(T_i(x^b) > 0) g_{i1}(x^b) + B_u 1(T_i(x^b) = 0)] \leq E[g_{i1}(x^a)] - E[g_{i1}(x^b)] = \Delta. \end{aligned} \tag{18}$$

It follows similarly that $\Delta \leq \Delta_u$. To show sharpness, let $\tilde{\alpha}_i = (\alpha_i, X_i)$. Define

$$\begin{aligned} g(x, \tilde{\alpha}_i, \varepsilon_{it}, C_a, C_b) &= 1(T_i(x) > 0) g_0(x, \alpha_i, \varepsilon_{it}) \\ +1(T_i(x) &= 0)[C_a 1(x = x^a) + C_b 1(x = x^b)], \end{aligned}$$

where $B_\ell \leq C_a \leq B_u$ and $B_\ell \leq C_b \leq B_u$. Note that $T_i(X_{it}) > 0$ with probability one, so that $g(X_{it}, \tilde{\alpha}_i, \varepsilon_{it}, C_a, C_b) = g_0(X_{it}, \alpha_i, \varepsilon_{it}) = Y_{it}$. Hence the conditional distribution of $(Y_{i1}, ..., Y_{iT})'$ given X_i is the same for g 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 $(C_a, C_b) = (B_\ell, B_u)$ we have

$$\begin{aligned} \Delta &= E[g(x^a, \tilde{\alpha}_i, \varepsilon_{it}, B_\ell, B_u) - g(x^b, \tilde{\alpha}_i, \varepsilon_{it}, B_\ell, B_u)] \\ &= E[1(T_i(x^a) > 0) g_i(x^a) + 1(T_i(x^a) = 0) B_\ell] \\ &\quad - E[1(T_i(x^b) > 0) g_i(x^b) + 1(T_i(x^b) = 0) B_u] = \Delta_\ell, \end{aligned}$$

and the lower bound is attained. Similarly the upper bound is attained when $(C_a, C_b) = (B_u, B_\ell)$.

Turning now to the quantile bounds, it follows as in the proof of Lemma A7 applied to $\hat{G}_\ell(y, x^a)$ and to $\hat{G}_\ell(y, x^b) + \bar{P}(x^b)$ that

$$\hat{q}_u^d \xrightarrow{p} q_u^d, \hat{q}_\ell^d \xrightarrow{p} q_\ell^d, G_\ell(q_u^d, x^d) = \lambda, G_\ell(q_\ell^d, x^d) + \bar{\mathcal{P}}(x^d) = \lambda, d \in \{a, b\}.$$

It also follows as in eq. (18) that $G_\ell(y,x) \leq G(y,x) \leq G_\ell(y,x) + \bar{\mathcal{P}}(x)$, implying $\Delta_{\lambda\ell} \leq \Delta_\lambda \leq \Delta_{\lambda u}$. Next, it follows as in Lemma A7 that

$$\begin{aligned}
\sqrt{n}(\hat{q}_u^a - q_u^a) &= -G_\ell'(q_u^a, x^a)^{-1}\frac{1}{\sqrt{n}}\sum_i [G_i(q_u^a, x) - \lambda] + o_p(1), \\
\sqrt{n}(\hat{q}_\ell^b - q_\ell^b) &= -G_\ell'(q_\ell^b, x^b)^{-1}\frac{1}{\sqrt{n}}\sum_i \left[G_i(q_\ell^b, x^b) + 1(T_i(x^b) = 0) - \lambda\right] + o_p(1).
\end{aligned}$$

Differencing then gives

$$\sqrt{n}(\hat{\Delta}_u - \Delta_u) = -\sum_{i=1}^n \frac{\Psi_{\lambda i}^u}{\sqrt{n}} + o_p(1), \Psi_{\lambda i}^u = \frac{G_i(q_u^a, x^a) - \lambda}{G_\ell'(q_u^a, x^a)} - \frac{G_i(q_\ell^b, x^b) + 1(T_i(x^b) = 0) - \lambda}{G_\ell'(q_\ell^b, x^b)}.$$

It follows similarly that

$$\sqrt{n}(\hat{\Delta}_\ell - \Delta_\ell) = -\sum_{i=1}^n \frac{\Psi_{\lambda i}^\ell}{\sqrt{n}} + o_p(1), \Psi_{\lambda i}^\ell = \frac{G_i(q_\ell^a, x) + 1(T_i(x^a) = 0) - \lambda}{G_\ell'(q_\ell^a, x^a)} - \frac{G_i(q_u^b, x^b) - \lambda}{G_\ell'(q_u^b, x^b)}.$$

Then for $\Sigma_\lambda = Var(\Psi_{\lambda i}^\ell, \Psi_{\lambda i}^u)$ the next conclusion follows by the CLT. It also follows by similar arguments to the proof of Theorem A8 that $\sum_{i=1}^n \left(\hat{\Psi}_{\lambda i}^\ell - \Psi_{\lambda i}^\ell\right)^2/n \xrightarrow{p} 0$ and $\sum_{i=1}^n \left(\hat{\Psi}_{\lambda i}^u - \Psi_{\lambda i}^u\right)^2/n \xrightarrow{p} 0$. The consistency of $\hat{\Sigma}_\lambda$ then follows by standard methods.

To show sharpness of the QTE bounds, define $\tilde{\alpha}_i$ and $g(x, \tilde{\alpha}_i, \varepsilon_{it}, C_a, C_b)$ as in the proof of the ATE bounds, but now for any $C_a, C_b \in \mathbb{R}$. Let $G(y, x, C_a, C_b) = E[1(g(x, \tilde{\alpha}_i, \varepsilon_{it}, C_a, C_b) \leq y)]$. Note that for $d \in \{a, b\}$,

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

Let $q(\lambda, x, C_a, C_b)$ be the associated QSF. For $d \in \{a, b\}$,

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

For λ with $\bar{\mathcal{P}}(x^d) < \lambda < 1 - \bar{\mathcal{P}}(x^d)$ we have $q(\lambda, x^d, C_a, C_b) = q_\ell(\lambda, x^d)$ for C_d small enough that $G_\ell(C_d, x) + \bar{\mathcal{P}}(x^d) < \lambda$ and $q(\lambda, x^d, C_a, C_b) = q_u(\lambda, x^d)$ for C_d big enough. For $\lambda \leq \bar{\mathcal{P}}(x^d)$ we have $q(\lambda, x^d, C_a, C_b) = q_u(\lambda, x)$ for all C_d big enough (by $\lambda < 1 - \bar{\mathcal{P}}(x^d)$) and $\lim_{C_d \longrightarrow -\infty} q(\lambda, x^d, C_a, C_b) = -\infty = q_\ell(\lambda, x)$. For $\lambda \geq 1 - \bar{\mathcal{P}}(x^d)$ we have $q(\lambda, x^d, C_a, C_b) = q_\ell(\lambda, x^d)$ for all C_d small enough and $\lim_{C_d \longrightarrow \infty} q(\lambda, x^d, C_a, C_b) = +\infty = q_u(\lambda, x^d)$. Therefore, we have

$$\begin{aligned}
\lim_{C_a \longrightarrow -\infty, C_b \longrightarrow +\infty} [q(\lambda, x^a, C_a, C_b) - q(\lambda, x^b, C_a, C_b)] &= q_\ell(\lambda, x^a) - q_u(\lambda, x^b), \\
\lim_{C_a \longrightarrow +\infty, C_b \longrightarrow -\infty} [q(\lambda, x^a, C_a, C_b) - q(\lambda, x^b, C_a, C_b)] &= q_u(\lambda, x^a) - q_\ell(\lambda, x^b),
\end{aligned}$$

showing the bounds are sharp. *Q.E.D.*

A4.2 Bounds under monotonicity

We now turn to the bounds when g_0 is known to be monotonic, satisfying the following condition.

ASSUMPTION A1: *For some x^a and x^b, $g_0(x^a, \alpha_i, \varepsilon_{it}) \geq g_0(x^b, \alpha_i, \varepsilon_{it})$.*

This condition leads to tighter bounds for the ASF and QSF. Here we will give results showing estimable population bounds under monotonicity. We will also briefly describe how to estimate them but for brevity do not give the full asymptotic theory. Define $1_i^a = 1(T_i(x^a) > 0)$, $1_i^b = 1(T_i(x^b) > 0)$, $\bar{\mathcal{P}}(x^b, x^a) = \Pr(T_i(x^a) = T_i(x^b) = 0)$, and

$$\begin{aligned} G_u^*(y, x^a) &= E[G_i(y, x^a) + (1 - 1_i^a)G_i(y, x^b)] + \bar{\mathcal{P}}(x^b, x^a), \\ G_\ell^*(y, x^b) &= E[G_i(y, x^b) + (1 - 1_i^b)G_i(y, x^a)]. \end{aligned}$$

THEOREM A9: *Suppose that Assumptions 1, 2, 5, and A1 are satisfied. If $E[|g_0(x, \alpha_i, \varepsilon_{it})|] < \infty$ for $x \in \{x^a, x^b\}$ then $\Delta \geq P\delta$. Also, if $G_u^*(y, x^a)$ and $G_\ell^*(y, x^b)$ are continuous and strictly increasing on the interior of their range then $q(\lambda, x^a) \geq Q(\lambda, G_u^*(\cdot, x^a))$ and $q(\lambda, x^b) \leq Q(\lambda, G_\ell^*(\cdot, x^b))$, so that*

$$\Delta_\lambda \geq Q(\lambda, G_u^*(\cdot, x^a)) - Q(\lambda, G_\ell^*(\cdot, x^b)).$$

Proof: Note that $1 = 1_i^a + (1 - 1_i^a)1_i^b + (1 - 1_i^a)(1 - 1_i^b)$. By Lemma A4,

$$E[1_i^a g_{i1}(x^a)] = E[\bar{Y}_i(x^a)], E[1_i^b g_{i1}(x^b)] = E[\bar{Y}_i(x^b)].$$

Then by monotonicity

$$\begin{aligned} \mu(x^a) &= E[g_{i1}(x^a)] \geq E[\{1_i^a + (1 - 1_i^a)(1 - 1_i^b)\}g_{i1}(x^a)] + E[(1 - 1_i^a)1_i^b g_{i1}(x^b)] \\ &= E[1_i^a \bar{Y}_i(x^a) + (1 - 1_i^a)1_i^b \bar{Y}_i(x^b) + (1 - 1_i^a)(1 - 1_i^b)g_{i1}(x^a)]. \end{aligned}$$

Similarly

$$\mu(x^b) \leq E[1_i^b \bar{Y}_i(x^b) + (1 - 1_i^b)1_i^a \bar{Y}_i(x^a) + (1 - 1_i^a)(1 - 1_i^b)g_{i1}(x^b)].$$

Subtracting this inequality from the previous one, and noting that $1_i^a - (1 - 1_i^b)1_i^a = 1_i^b 1_i^a = D_i$ and $-1_i^b + (1 - 1_i^a)1_i^b = -D_i$,

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

giving the first conclusion.

Next, similarly to above,

$$\begin{aligned} G(y, x^a) &= E[\{1_i^a + (1 - 1_i^a)(1 - 1_i^b) + (1 - 1_i^a)1_i^b\}1(g_{i1}(x^a) \leq y)] \\ &\leq E[G_i(y, x^a)] + E[(1 - 1_i^a)G_i(y, x^b)] + \bar{\mathcal{P}}(x^b, x^a) = G_u^*(y, x^a). \\ G(y, x^b) &\geq G_\ell^*(y, x^b). \end{aligned}$$

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

Estimation of the bounds under monotonicity is straightforward. We can estimate the lower bound for the ATE by $(\sum_{i=1}^{n} D_i/n)\,\hat{\delta}$. We can estimate the quantile bounds by inverting

$$\begin{aligned}\hat{G}_u^*(y,x^a) &= \sum_{i=1}^{n}[\bar{G}_i(y,x^a)+(1-1_i^a)\bar{G}_i(y,x^b)+1(T_i(x^b)=T_i(x^a)=0)]/n,\\ \hat{G}_\ell^*(y,x^b) &= \sum_{i=1}^{n}[\bar{G}_i(y,x^b)+(1-1_i^b)\bar{G}_i(y,x^a)]/n.\end{aligned}$$

Asymptotic theory for these estimators of bounds under monotonicity is straightforward. We do not know if they are sharp.

A5 Supplements to Section 5

Here we give the proof of Theorem 3 as well as bounds that impose monotonicity.

A5.1 Proof of Theorem 3

We first prove the second part of Lemma A4 for the dynamic model. Let $d_{it}(x) = 1(X_i \in \mathcal{X}_t(x))$. By Assumption 3, $\sum_{t=1}^{T} d_{it}(x) = 1(T_i(x) > 0)$, and the fact that $d_{it}(x)$ depends only on $X_{it}, X_{i,t-1}, ..., X_{i1}$ we have

$$\begin{aligned}E[\hat{Y}_i(x)|X_{i1}] &= \sum_{t=1}^{T} E[d_{it}(x)Y_{it}|X_{i1}] = \sum_{t=1}^{T} E[d_{it}(x)E[g_{it}(x)|X_{it},...,X_{i1}]|X_{i1}]\\ &= \sum_{t=1}^{T} E[d_{it}(x)E[g_{i1}(x)|X_{i1}]|X_{i1}] = E[1(T_i(x)>0)|X_{i1}]E[g_{i1}(x)|X_{i1}].\end{aligned}$$

Let

$$\begin{pmatrix} m_{\ell i} \\ m_{ui} \end{pmatrix} = \begin{pmatrix} \hat{Y}_i(x^a) - \hat{Y}_i(x^b) + B_\ell 1(T_i(x^a)=0) - B_u 1(T_i(x^b)=0) \\ \hat{Y}_i(x^a) - \hat{Y}_i(x^b) + B_u 1(T_i(x^a)=0) - B_\ell 1(T_i(x^b)=0) \end{pmatrix}.$$

Note that $\hat{\Delta}_\ell = \sum_{i=1}^{n} m_{\ell i}/n$ and $\hat{\Delta}_u = \sum_{i=1}^{n} m_{ui}/n$.Then for $\Sigma = Var((m_{\ell i}, m_{ui}))$, $\Delta_\ell = E[m_{\ell i}]$, and $\Delta_u = E[m_{ui}]$ the first and second conclusions follow by standard arguments for a vector of sample means.

Next, note that $E[g_{i1}(x^a)|X_{i1}] \le B_u$ by Assumption 6, so that

$$E[B_u 1(T_i(x^a)=0)|X_{i1}] \ge E[1(T_i(x^a)=0)|X_{i1}]E[g_{i1}(x^a)|X_{i1}].$$

Then by iterated expectations and $T_i(x^a) \ge 0$,

$$\begin{aligned}E[\hat{Y}_i(x^a) + B_u 1(T_i(x^a)=0)|X_{i1}] &\ge E[1(T_i(x^a)>0)|X_{i1}]E[g_{i1}(x^a)|X_{i1}]\\ &\quad +E[1(T_i(x^a)=0)|X_{i1}]E[g_{i1}(x^a)|X_{i1}] = E[g_{i1}(x^a)|X_{i1}].\end{aligned}$$

Taking expectations of both sides of this inequality gives

$$E[\hat{Y}_i(x^a) + B_u 1(T_i(x^a) = 0)] \geq \mu(x^a).$$

Similarly we have $E[\hat{Y}_i(x^a) + B_\ell 1(T_i(x^a) = 0)] \leq \mu(x^a)$. Replacing x^a by x^b and differencing gives $\Delta_\ell \leq \Delta \leq \Delta_u$.

Turning to the quantile bounds, we next prove the first part of Lemma A4 for a dynamic model. Let $G_i(y,x)$ here, in the dynamic case, be given by

$$\begin{aligned} G_i(y,x) &= \sum_{t=1}^{T} d_{it}(x) 1(Y_{it} \leq y) = \sum_{t=1}^{T} d_{it}(x) 1(g_{it}(x) \leq y), \\ G_\ell(y,x) &= E[E[1(T_i(x) > 0)|X_{i1}]1(g_{i1}(x) \leq y)]. \end{aligned}$$

Note that since $\sum_{t=1}^{T} d_{it}(x) = 1(T_i(x) > 0)$ and $d_{it}(x)$ depends only on $X_{it}, X_{it-1}, ..., X_{i1}$, Assumption 3 implies

$$\begin{aligned} E[G_i(y,x)] &= E[\sum_{t=1}^{T} d_{it}(x) 1(g_{it}(x) \leq y)] = E[\sum_{t=1}^{T} d_{it}(x) E[1(g_{it}(x) \leq y)|X_{it}, ..., X_{i1}]] \\ &= E[\sum_{t=1}^{T} d_{it}(x) E[1(g_{i1}(x) \leq y)|X_{i1}]] = E[1(T_i(x) > 0) E[1(g_{i1}(x) \leq y)|X_{i1}]] \\ &= G_\ell(y,x). \end{aligned}$$

Also, since $d_{it}(x) d_{is}(x) = 0$ for any $s \neq t$ and $d_{it}(x)^2 = d_{it}(x)$, Assumption 3 implies that

$$\begin{aligned} E[\{\hat{G}_i(y,x) - G_i(y,x)\}^2] &= E[\sum_{t=1}^{T} d_{it}(x) \{\Phi(\frac{y - g_{it}(x)}{h}) - 1(g_{it}(x) \leq y)\}^2] \\ &\leq E[\sum_{t=1}^{T} d_{it}(x) E[\{\Phi(\frac{y - g_{it}(x)}{h}) - 1(g_{it}(x) \leq y)\}^2 | X_{it}, ..., X_{i1}]] \\ &= E[1(T_i(x) > 0) E[\{\Phi(\frac{y - g_{i1}(x)}{h}) - 1(g_{i1}(x) \leq y)\}^2 | X_{i1}]] \\ &= E[E[1(T_i(x) > 0)|X_{i1}] \{\Phi(\frac{y - g_{i1}(x)}{h}) - 1(g_{i1}(x) \leq y)\}^2] \end{aligned}$$

By Assumption 7 with X_{i1} replacing X_i it follows that $g_{i1}(x)$ is continuously distributed for the probability measure weighted by $E[1(T_i(x) > 0)|X_{i1}]$. Therefore it follows similarly to the proof of Lemma A7 that $E[\{\hat{G}_i(y,x) - G_i(y,x)\}^2] \longrightarrow 0$ as $h \longrightarrow 0$. It also follows similarly to the proof of Lemma A7

$$E[\hat{G}_i(y,x)] = E[G_i(y,x)] + O(h^2).$$

The conclusion now follows exactly like the proof of Theorem 2. *Q.E.D.*

A5.2 Bounds under monotonicity

We now turn to the bounds when g_0 is known to be monotonic, satisfying Assumption A1, in the dynamic model. This condition leads to tighter bounds for the ASF and QSF. Here we will give results showing estimable population bounds under monotonicity. We will also briefly describe how to estimate them but for brevity do not give the full asymptotic theory. For $d \in \{a, b\}$, define $1_{it}^d = 1(X_i \in \mathcal{X}_t(x^d))$, $t = 1, ..., T$, $\bar{1}_i^d = 1(X_i \in \bar{\mathcal{X}}(x^d))$, and $\tilde{1}_{iT}^d = 1(X_{iT} = x^d)$. Let

$$\begin{aligned} G_u^*(y, x^a) &= E[G_i(y, x^a) + \bar{1}_i^a\{\tilde{1}_{iT}^b 1(Y_{iT} \leq y) + (1 - \tilde{1}_{iT}^b)\}], \\ G_\ell^*(y, x^b) &= E[G_i(y, x^b) + \bar{1}_i^b \tilde{1}_{iT}^a 1(Y_{iT} \leq y)]. \end{aligned}$$

THEOREM A10: *Suppose that Assumptions 1, 3, 5, and A1 are satisfied. If* $E[|g_0(x, \alpha_i, \varepsilon_{it})|] < \infty$ *for* $x \in \{x^a, x^b\}$ *then*

$$\Delta \geq E[\hat{Y}_i(x^a) - \hat{Y}_i(x^b)] + E[\bar{1}_i^a(\tilde{1}_{iT}^b Y_{iT} + (1 - \tilde{1}_{iT}^b)B_\ell)] - E[\bar{1}_i^b(\tilde{1}_{iT}^a Y_{iT} + (1 - \tilde{1}_{iT}^a)B_u)].$$

Also, if $G_u^*(y, x^a)$ *and* $G_\ell^*(y, x^b)$ *are continuous and strictly increasing on the interior of their range then* $q(\lambda, x^a) \geq Q(\lambda, G_u^*(\cdot, x^a))$ *and* $q(\lambda, x^b) \leq Q(\lambda, G_\ell^*(\cdot, x^b))$, *so that*

$$\Delta_\lambda \geq Q(\lambda, G_u^*(\cdot, x^a)) - Q(\lambda, G_\ell^*(\cdot, x^b)).$$

Proof: Note that $1 = \sum_{t=1}^T 1_{it}^a + \bar{1}_i^a \tilde{1}_{iT}^b + \bar{1}_i^a(1 - \tilde{1}_{iT}^b)$. By Lemma A4,

$$\sum_{t=1}^T E[1_{it}^a g_{it}(x^a)] = E[\hat{Y}_i(x^a)], \sum_{t=1}^T E[1_{it}^b g_{it}(x^b)] = E[\hat{Y}_i(x^b)].$$

Then by Assumption 3, monotonicity and $g_{iT}(x^a) \geq B_\ell$,

$$\begin{aligned} \mu(x^a) &= \sum_{t=1}^T E[1_{it}^a g_{it}(x^a)] + E[\bar{1}_i^a g_{iT}(x^a)] \\ &\geq \sum_{t=1}^T E[1_{it}^a g_{it}(x^a)] + E[\bar{1}_i^a \tilde{1}_{iT}^b g_{iT}(x^b)] + E[\bar{1}_i^a(1 - \tilde{1}_{iT}^b)]B_\ell \\ &= E[\hat{Y}_i(x^a)] + E[\bar{1}_i^a \tilde{1}_{iT}^b Y_{iT}] + E[\bar{1}_i^a(1 - \tilde{1}_{iT}^b)]B_\ell. \end{aligned}$$

Similarly we have

$$\mu(x^b) \leq E[\hat{Y}_i(x^b)] + E[\bar{1}_i^b \tilde{1}_{iT}^a Y_{iT}] + E[\bar{1}_i^b(1 - \tilde{1}_{iT}^a)]B_u.$$

Subtracting this inequality from the previous one gives the first conclusion.

Next, similarly to above,

$$\begin{aligned} G(y, x^a) &= \sum_{t=1}^T E[1_{it}^a 1(g_{it}(x^a) \leq y)] + E[\bar{1}_i^a 1(g_{iT}(x^a) \leq y)] \\ &\leq E[G_i(y, x^a)] + E[\bar{1}_i^a \tilde{1}_{iT}^b 1(Y_{iT} \leq y)] + E[\bar{1}_i^a(1 - \tilde{1}_{iT}^b)] = G_u^*(y, x^a). \\ G(y, x^b) &\geq G_\ell^*(y, x^b). \end{aligned}$$

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

If $X_{it} \in \{0,1\}$, $x^b = 0$, and $x^a = 1$, then $\bar{1}_i^b(1-\tilde{1}_{iT}^a) = \bar{1}_i^a(1-\tilde{1}_{iT}^b) = 0$ and the lower bound for Δ does not depend on B_ℓ and B_u.

Estimation of the bounds under monotonicity is straightforward. We can estimate the lower bound for the ATE by

$$\sum_{i=1}^{n}[\hat{Y}_i(x^a) - \hat{Y}_i(x^b) + \bar{1}_i^a(\tilde{1}_{iT}^b Y_{iT} + (1-\tilde{1}_{iT}^b)B_\ell) - \bar{1}_i^b(\tilde{1}_{iT}^a Y_{iT} + (1-\tilde{1}_{iT}^a)B_u)]/n.$$

We can estimate the quantile bounds by inverting

$$\begin{aligned}\hat{G}_u^*(y,x^a) &= \sum_{i=1}^{n}[\hat{G}_i(y,x^a) + \bar{1}_i^a\{\tilde{1}_{iT}^b 1(Y_{iT} \le y) + (1-\tilde{1}_{iT}^a)\}]/n, \\ \hat{G}_\ell^*(y,x^b) &= \sum_{i=1}^{n}[\hat{G}_i(y,x^b) + \bar{1}_i^b \tilde{1}_{iT}^a 1(Y_{iT} \le y)]/n.\end{aligned}$$

Asymptotic theory for these estimators of bounds under monotonicity is straightforward. We do not know if they are sharp.

A6 Supplements to Section 6

In addition to the proofs of the rate results of Section 6, we here give necessary and sufficient conditions for identification as $T \longrightarrow \infty$ and extend the identification and rate results to the QTE.

A6.1 Identification as $T \to \infty$

We begin with the identification result. The necessary and sufficient condition for identification of Δ as T grows is

Assumption A2: $\Pr(\Pr(X_{it} = x|\alpha_i) > 0) = 1$ *for* $x \in \{x^a, x^b\}$ *and some* $t \in \{1,\dots,T\}$.

If this condition does not hold for both x^b and x^a then some individuals, as represented by α_i, will never reach either x^b or x^a, so we cannot nonparametrically identify the treatment effect for those individuals, and hence the overall treatment effect is not identified. A related condition was formulated in Chamberlain (1982, p. 17) but was used for a different purpose, as a sufficient condition for a least squares estimate for a single individual to converge to that individual's coefficient.

The following result shows the key role of Assumption A2 in achieving identification as $T \longrightarrow \infty$.

THEOREM A11: *Suppose that Assumptions 1 and 5 are satisfied. If Assumption A2 is not satisfied then* $\bar{\mathcal{P}}(x)$ *is bounded away from zero uniformly in* T *for* $x = x^a$ *or* $x = x^b$*, so that if Assumption 6 is satisfied,* $\Delta_u - \Delta_\ell$ *does not converge to zero as* T *grows. Suppose also that* $(X_{i1}, X_{i2}, ...)$ *is stationary and ergodic conditional on* α_i*. If Assumptions 2 and A2 are satisfied and* $E[|g_0(x, \alpha_i, \varepsilon_{i1})|] < \infty$ *for* $x = x^a$ *and* $x = x^b$*, then* $\delta \longrightarrow \Delta$ *as* $T \longrightarrow \infty$*. If Assumptions 3, 6, and A2 are satisfied then* $\Delta_u - \Delta_\ell \longrightarrow 0$ *as* $T \longrightarrow \infty$*.*

Proof: First, note that if Assumption A2 is not satisfied then for some $x^d \in \{x^a, x^b\}$ there is a set $\mathcal{A}$ with $\Pr(\mathcal{A}) > 0$ such that $\Pr(X_{it} = x^d|\alpha_i) = 0$ for all t and $\alpha_i \in \mathcal{A}$. Then

$$E[T_i(x^d)|\alpha_i \in \mathcal{A}] = \sum_{t=1}^{T} E[1(X_{it} = x^d)|\alpha_i \in \mathcal{A}] = 0.$$

Since $T_i(x^d)$ is a nonnegative random variable, this implies that $\Pr(T_i(x^d) = 0|\alpha_i) = 1$ for all T and $\alpha_i \in \mathcal{A}$. Therefore

$$\bar{\mathcal{P}}(x^d) = E[\Pr(T_i(x^d) = 0|\alpha_i)] \geq E[1(\mathcal{A}) \Pr(T_i(x^d) = 0|\alpha_i)] = \Pr(\mathcal{A}) > 0.$$

Thus $\bar{\mathcal{P}}(x^d)$ is bounded away from zero for all T, and hence under Assumption 6, $(B_u - B_\ell)[\bar{\mathcal{P}}(x^a) + \bar{\mathcal{P}}(x^b)] \geq (B_u - B_\ell)\bar{\mathcal{P}}(x^d)$ does not converge to zero.

Next suppose that Assumptions 2 and A2 are satisfied, $(X_{i1}, X_{i2}, ...)$ is stationary and ergodic conditional on α_i, and that $x \in \{x^a, x^b\}$. Recall that $T_i(x) = \sum_{t=1}^{T} 1(X_{it} = x)$. By the ergodic theorem, there is a set of α_i having probability one such that

$$T_i(x)/T \xrightarrow{a.s.} E[1(X_{it} = x)|\alpha_i] = \Pr(X_{it} = x \mid \alpha_i).$$

Under Assumption A2 $\Pr(X_{it} = x \mid \alpha_i) > 0$ on a set of α_i with probability one (a.s. α_i henceforth). Therefore $1(T_i(x) > 0) \xrightarrow{a.s.} 1$ a.s. α_i. Since this holds for both x^a and x^b, it follows that

$$D_i = 1(T_i(x^a) > 0)1(T_i(x^b) > 0) \xrightarrow{a.s.} 1$$

a.s. α_i. Let $\Delta_i = g_{i1}(x^a) - g_{i1}(x^b)$. Note that $|D_i\Delta_i| \leq |\Delta_i|$ and $E[|\Delta_i||\alpha_i] < \infty$ a.s. α_i. Then by the dominated convergence theorem (DCT henceforth),

$$E[D_i\Delta_i|\alpha_i] \longrightarrow E[\Delta_i|\alpha_i], E[D_i|\alpha_i] \longrightarrow 1 \text{ a.s. } \alpha_i.$$

Then by the applying the DCT again,

$$E[D_i\Delta_i] \longrightarrow E[\Delta_i] = \Delta, E[D_i] \longrightarrow 1,$$

giving the first conclusion.

Suppose next that Assumptions 3 and 6 are satisfied, and $(X_{i1}, X_{i2}, ...)$ is stationary and ergodic conditional on α_i. Recall that $\Delta_u - \Delta_\ell = (B_u - B_\ell)[\bar{\mathcal{P}}(x^a) + \bar{\mathcal{P}}(x^b)]$. If Assumption A2 is satisfied then since $1(T_i(x^a) > 0) \geq D_i$ we have

$$\bar{\mathcal{P}}(x^a) = E[1(T_i(x^a) = 0)] \leq 1 - E[D_i] \longrightarrow 0$$

Similarly we have $\bar{\mathcal{P}}(x^b) \longrightarrow 0$ so the second conclusion holds. Q.E.D.

A6.2 Proof of Theorem 4

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 $T > J$,

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

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

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

giving the first conclusion.

For the second conclusion, note that the conditional i.i.d. assumption and the bound implies that for $P_i = \Pr(X_{it} \neq x | \alpha_i)$ we have $\bar{\mathcal{P}}(x) = E[P_i^T]$ being no greater than a constant times the T^{th} raw moment of a Beta distribution with parameters γ and v. Also, it is well known that $T^v \Gamma(T+\gamma)/\Gamma(T+\gamma+v) \longrightarrow 1$ as $T \longrightarrow \infty$. Therefore, we have

$$\begin{aligned} E[P_i^T] &\leq C[\Gamma(\gamma+v)/\Gamma(\gamma)\Gamma(v)] \int_0^1 p^{T+\gamma-1}(1-p)^{v-1} dp \\ &\leq C[\Gamma(\gamma+v)/\Gamma(\gamma)\Gamma(v)][\Gamma(T+\gamma)\Gamma(v)/\Gamma(T+\gamma+v)] \\ &= C\Gamma(T+\gamma)/\Gamma(T+\gamma+v) \leq CT^{-v}. \quad Q.E.D. \end{aligned}$$

A6.3 Proof of Theorem 5

Note that $\Pr(Y_{it} = 0 | Y_{i,t-1} = 0, \alpha_i) = 1 - H(\alpha_{1i})$

$$\begin{aligned} \bar{\mathcal{P}}(1) &= E[\Pr(Y_{i,T-1} = Y_{i,T-2} = ... = Y_{i0} = 0 | \alpha_i)] \\ &= E[\Pi_{t=1}^{T-1} \Pr(Y_{it} = 0 | Y_{i,t-1} = 0, \alpha_i) \Pr(Y_{i0} = 0 | \alpha_i)] \\ &\leq E[\{1 - H(\alpha_{i1})\}^{T-1}]. \end{aligned}$$

By a change of variables we find that the pdf $f(p)$ of $1 - H(\alpha_{i1})$ is

$$f(p) = f_1(H^{-1}(1-p))/f_\varepsilon(H^{-1}(1-p)) \leq C(1-p)^{v-1}p^{v-1}.$$

Thus, the pdf of $1 - H(\alpha_{i1})$ is bounded above by a Beta pdf with parameters v, v. It then follows as in the proof of Theorem 4 that $\bar{\mathcal{P}}(1) \leq C(T-1)^{-v} \leq CT^{-v}$. It follows similarly that $\bar{\mathcal{P}}(0) \leq CT^{-v}$. *Q.E.D.*

A6.4 Identification rates for QTE

Finally, we show that the nonparametric rates and nonidentification results apply to the QTE. We do this by giving Lemmas for quantile bounds that apply to both static and dynamic models. The first Lemma shows that the identification rate is at least as fast as the rate at which $\bar{\mathcal{P}}(x)$ decreases.

LEMMA A12: *Suppose that $G(y)$ is a CDF that is strictly increasing and continuously differentiable on $\{y : 0 < G(y) < 1\}$ and that $G_T(y)$ is a continuous function and $\bar{\mathcal{P}}_T$ a nonnegative constant satisfying*

$$G_T(y) \leq G(y) \leq G_T(y) + \bar{\mathcal{P}}_T, G_T(-\infty) = 0, G_T(\infty) + \bar{\mathcal{P}}_T = 1.$$

If $\bar{\mathcal{P}}_T \longrightarrow 0$ as $T \longrightarrow \infty$ then for $0 < \lambda < 1$ and large enough T there are $q_{\ell T} \leq q \leq q_{uT}$ satisfying

$$\lambda = G_T(q_{uT}) = G(q) = G_T(q_{\ell T}) + \bar{\mathcal{P}}_T.$$

Also, any such q_{uT} and $q_{\ell T}$ satisfy: $q_{uT} - q_{\ell T} = O(\bar{\mathcal{P}}_T)$.

Proof: Choose T large enough that $\bar{\mathcal{P}}_T < \min(\lambda, 1-\lambda)$. Then $G_T(\infty) = 1 - \bar{\mathcal{P}}_T > \lambda$ and $G_T(-\infty) + \bar{\mathcal{P}}_T = \bar{\mathcal{P}}_T < \lambda$. Therefore by continuity of $G_T(y)$ there exist q_{uT} such that $\lambda = G_T(q_{uT})$ and $q_{\ell T}$ such that $\lambda = G_T(q_{\ell T}) + \bar{\mathcal{P}}_T$. Also, by $G(y)$ being a strictly increasing CDF there is a unique q with $\lambda = G(q)$. Note $G(q) = G_T(q_{uT}) \leq G(q_{uT})$ so that $q_{uT} \geq q$ by $G(q)$ strictly monotonic. It follow similarly that $q_{\ell T} \leq q$. Also, for any $\varepsilon > 0$ we have $G(q-\varepsilon) < G(q)$, so that for large enough T it follow

$$G(q-\varepsilon) < G(q) - \bar{\mathcal{P}}_T = G_T(q_{\ell T}) \leq G(q_{\ell T}).$$

By strict monotonicity of $G(q)$ it follows that $q_{\ell T} > q - \varepsilon$ for large enough T. Since ε is arbitrary we have $q_{\ell T} \longrightarrow q$. It follow similarly that $q_{uT} \longrightarrow q$.

Next, choose ε small enough that $\partial G(\tilde{q})/\partial q \geq C > 0$ for $\tilde{q} \in \mathcal{I} =[q-\varepsilon, q+\varepsilon]$. Note that for T large enough, $q_{\ell T}, q_{uT} \in \mathcal{I}$. Also we have

$$G(q_{\ell T}) + 2\bar{\mathcal{P}}_T \geq G_T(q_{\ell T}) + 2\bar{\mathcal{P}}_T = G(q) + \bar{\mathcal{P}}_T = G_T(q_{uT}) + \bar{\mathcal{P}}_T \geq G(q_{uT}).$$

Subtracting $G(q_{\ell T})$ from both sides and expanding gives

$$2\bar{\mathcal{P}}_T \geq G(q_{uT}) - G(q_{\ell T}) = \frac{\partial G(\bar{q}_T)}{\partial q}(q_{uT} - q_{\ell T}) \geq C(q_{uT} - q_{\ell T}).$$

Dividing through by C gives $q_{uT} - q_{\ell T} \leq C\bar{\mathcal{P}}_T$, implying the conclusion. *Q.E.D.*

The next result gives conditions under which the identification rate is no faster than the rate at which $\bar{\mathcal{P}}(x)$ decreases. This result will also show that quantile effects are not identified as $T \longrightarrow \infty$ if $\bar{\mathcal{P}}(x)$ does not go to zero.

LEMMA A13: *If the conditions of Lemma A12 are satisfied and $G_T(y)$ is continuously differentiable with $|\partial G_T(y)/\partial y| \leq C$ for all y and T then there is C* such that for $\bar{\mathcal{P}}_T > 0$,

$$q_{uT} - q_{\ell T} \geq C\bar{\mathcal{P}}_T.$$

Proof: As in the proof of Lemma A12 we have $G_T(q_{uT}) = G_T(q_{\ell T}) + \bar{\mathcal{P}}_T$. By the intermediate value theorem it follows that for some $q_{\ell T} \leq \bar{q} \leq q_{uT}$

$$\frac{\partial G_T(\bar{q})}{\partial q}(q_{uT} - q_{\ell T}) = \bar{\mathcal{P}}_T.$$

For $\bar{\mathcal{P}}_T > 0$ we must have $\partial G_T(\bar{q})/\partial q \neq 0$, so that

$$q_{uT} - q_{\ell T} = \left[\frac{\partial G_T(\bar{q})}{\partial q}\right]^{-1} \bar{\mathcal{P}}_T \geq C^{-1}\bar{\mathcal{P}}_T. \quad Q.E.D.$$

Taken together these two results show that the identification rate for the QTE is the same as the rate at which $\bar{\mathcal{P}}(x)$ decreases. Together they also show that if $\bar{\mathcal{P}}(x)$ does not go to zero the bounds do not shrink to a point. It is straightforward to check that the conditions of these results are satisfied.

A7 Supplements to Section 7

We now turn to the results of Section 7 and to one additional result on the consistency of non-linear fixed effects estimators for the identified ATE.

A7.1 Proof of Theorem 6

Consider first the static case where $X_{it} \in \{0, 1\}$. We show the result for $X^k = (0, ..., 0)'$. The result for $X^k = (1, ..., 1)'$ will follow similarly. Note that β^* is identified for logit so $B = \{\beta^*\}$. Let $Z = H(\alpha)$ and let $G(z)$ be the CDF of Z when $F \in \mathcal{F}_k = \mathcal{F}_k(\beta^*, \mathcal{P})$ is the CDF of α. By $(Y_{i1}, ..., Y_{iT})$ mutually independent conditional on α we have

$$M_t = \Pr(Y_{it} = 1, ..., Y_{i1} = 1 | X_i \in X^k) = \int H(\alpha)^t dF(\alpha) = \int Z^t dG(Z),$$

so that M_t is identified for $t = 1, ..., T$. Now consider a T^{th} order polynomial $P(z,T) = b_0 + b_1 z + ... + b_T z^T$ in z. Note that

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

does not depend on $F \in \mathcal{F}_k$. As a special case, $\int Z dG(Z) = M_1$ also does not depend on $F \in \mathcal{F}_k$. Define the function $h(z) = H(\beta^* + H^{-1}(z)) = ze^{\beta^*}/1 - (1 - e^{\beta^*})z$. Note $\Delta^k = \int [h(Z) - Z]dG(Z)$ for all $F \in \mathcal{F}_k$. For any polynomial $P(z,t)$ let $R(z,t) = h(z) - P(z,t)$ be the remainder. Then we have

$$\begin{aligned} \Delta_u^k - \Delta_\ell^k &= \sup_{F \in \mathcal{F}_k} \int [h(Z) - Z]dG(Z) - \inf_{F \in \mathcal{F}_k} \int [h(Z) - Z]dG(Z) \qquad (19) \\ &= \sup_{F \in \mathcal{F}_k} \int [P(Z,T) + R(Z,T)]dG(Z) - \inf_{F \in \mathcal{F}_k} \int [P(Z,T) + R(Z,T)]dG(Z) \\ &= \sup_{F \in \mathcal{F}_k} \int R(Z,T)dG(Z) - \inf_{F \in \mathcal{F}_k} \int R(Z,T)dG(Z) \le 2 \sup_{0 \le z \le 1} |R(z,T)|. \end{aligned}$$

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

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

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. (19).

Next consider the dynamic binary logit model where $X_{it} = Y_{i,t-1}$. It is known from Cox (1958) and Chamberlain (1985) that β^* identified for T large enough. We show the result for Δ^1 where $\mathcal{X}^1 = \{X_i : X_{i1} = 0\}$. The result for the ATE conditional on $X_{i1} = 1$ will follow analogously. Then

$$\Pr(Y_{it} = 0, ..., Y_{i1} = 0 | X_{i1} = 0) = \int [1 - H(\alpha)]^t dF(\alpha)$$

is identified for $t = 1, ..., T$. It follows by a standard argument that $M_t = \int H(\alpha)^t dF(\alpha)$ is identified for $t = 1, ..., T$. The proof then proceeds exactly as for the static case. *Q.E.D.*

A7.2 Consistency of fixed effects for identified ATE

We now consider the fixed effects estimator in a binary choice model with a binary regressor and $T = 2$. In some models fixed effect (FE) estimators of the ATE appear to have small biases; e.g. see Hahn and Newey (2004) and Fernández-Val (2009). Here we show consistency of FE for δ. To describe this result, note that the FE estimator of the ASF conditional on $X_i = X^k$ is

$$\begin{aligned}\hat{\mu}_k^{FE}(x) &= \sum_{i=1}^{n} 1(X_i = X^k) H(x\hat{\beta}_{FE} + \hat{\alpha}_i) / \sum_{i=1}^{n} 1(X_i = X^k), \\ \hat{\beta}_{FE}, \hat{\alpha}_1, ..., \hat{\alpha}_n &= \arg\max_{\beta, \alpha_1, ..., \alpha_n} \sum_{i,t} \ln\{H(X_{it}\beta + \alpha_i)^{Y_{it}} [1 - H(X_{it}\beta + \alpha_i)]^{1-Y_{it}}\}.\end{aligned}$$

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

$$\begin{aligned}\beta_T &= \text{argmax}_{\beta} \sum_{k=1}^{K} \mathcal{P}^k \sum_{j=1}^{J} \mathcal{P}_j^k \log \mathcal{L}_j^k \left(\alpha_j^k(\beta), \beta\right), \\ \alpha_j^k(\beta) &= \text{argmax}_{\alpha} \mathcal{L}_j^k (\alpha, \beta), (j = 1, ..., J; k = 1, ..., K),\end{aligned} \tag{20}$$

where $\mathcal{P}^k = E[1(X_i = X^k)]$. The corresponding limit of $\hat{\mu}_k^{FE}(x)$ is then given by

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

Note that 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 A14: *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

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

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

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

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.

A8 Supplements to Section 8

Here we give the proofs of Section 8 and additional numerical results for the logit model.

A8.1 Proof of Lemma 7

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), $\breve{\Gamma}_k(\beta) = \{\int \mathcal{L}^k(\alpha, \beta)\, dF(\alpha) : F$ is a CDF on $\Upsilon\}$. Therefore, $\int \mathcal{L}^k(\alpha, \beta)\, dF_k(\alpha) \in \breve{\Gamma}_k(\beta)$. 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

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

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

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

Similarly to the previous paragraph, let $\Gamma_k^\Delta(\beta)\equiv\left\{(\mathcal{L}^k\left(\alpha,\beta\right)',\Delta(\alpha,\beta))':\alpha\in\Upsilon\right\}$ and $\breve{\Gamma}_k^\Delta(\beta)$ be the convex hull of $\Gamma_k^\Delta(\beta)$. Then $(\mathcal{P}_1^k,...,\mathcal{P}_J^k,\bar{\Delta}(\beta))'\in\breve{\Gamma}_k^\Delta\left(\beta\right)$, 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 $\int\Delta(\alpha,\beta)dF_{k\beta}^{J+1}\left(\alpha\right)=\bar{\Delta}(\beta)$.

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

$$\max_{\left(\pi_1^k,...,\pi_{J+1}^k\right)}\sum_{m=1}^{J+1}\Delta(\alpha_m^k,\beta)\pi_m^k,\ s.t.$$

$$\sum_{m=1}^{J+1}\pi_m^k\mathcal{L}_j^k\left(\alpha_m^k,\beta\right)\ =\ \mathcal{P}_j^k,\sum_{m=1}^{J+1}\pi_m^k=1,\pi_m^k\geq 0,(m=1,...,J+1).$$

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 $rank\left(A\right)+1$ can be equality constraints. This means that at least $J+1-rank([A',1]')$ of active constraints are of the form $\pi_m^k=0$, see, e.g., Theorem 2.3 and Definition 2.9 (ii) in Bertsimas and Tsitsiklis (1997). Since each column of A sums to 1, $rank([A',1]')\leq J$ and a basic solution to this linear programming problem will have at least one zero. Thus, there are at most J strictly positive π_m^k's.[2] Therefore, 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

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

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

[2]Note that $rank([A',1]')\leq J$, since $\sum_{j=1}^J\mathcal{L}_j^k\left(\alpha,\beta\right)=1$. The exact rank of $[A',1]'$ depends on the sequence X^k, the parameter β, the form of $\mathcal{L}_j^k\left(\alpha,\beta\right)$, and T. For example in the model of equation (8) of the main text 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.

A8.2 Numerical results for logit model

We carry out some additional numerical calculations for the logit model where

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

where $L(0,1)$ denotes the standard logistic distribution normalized to have zero mean and unit variance. We consider different DGPs indexed by $\beta^* \in [-2,2]$ and $T \in \{2,3\}$. Figures 1 and 2 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 in Section 8 of the paper 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. As is well-known, we find that β^* is identified for $T \geq 2$. The nonparametric bounds for the ATEs (NP-bounds) can be very wide, even when we impose monotonicity (NPM-bounds). The semiparametric bounds for the ATEs (SP-bounds) are tighter than the nonparametric bounds and shrink exponentially fast with T, as shown in Theorem 6.

A8.3 Proof of Lemma 8

Consider the set $\bar{\Re} = (-\infty, +\infty) \cup \{-\infty, +\infty\}$. By assumption $H(v)$ 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 $\bar{v} = \max_{X^k \in \{X^1,\dots,X^K\}, \beta \in B} |X_t^{k\prime}\beta|$ and define the function

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

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

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

where the inequality holds by $\bar{v}+v \geq 0$ (implied by $v \geq -\bar{v}$) 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 $u, \tilde{u} \in [0,1]$,

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

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

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

A8.4 Proof of Theorem 9

This proof is omitted because it is very similar (but easier) than the proof of Theorem 10 to follow.

A9 Supplements to Section 9

Here we describe the estimation algorithm, give the proofs of Theorems 10 and 11, and present an alternative inference method based on projection.

A9.1 Estimation: Implementation Details

To implement the estimation method, we also start from 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.

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 $\beta \in \hat{B}$.

A9.2 Proof of Theorem 10

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 Θ_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 9 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 9,

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

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 $Q(\theta)$, $\inf\limits_{\Theta \backslash \Theta_I^\delta} Q(\theta) = \rho > 0$. It follows by uniform convergence that $\inf\limits_{\Theta \backslash \Theta_I^\delta} \hat{Q}(\theta) > \frac{\rho}{2}$ 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.*

A9.3 Proof of Theorem 11

We have that for $S_n(\mathcal{P}) = \hat{\theta} - \theta^* = \hat{\theta} - \theta^*(\mathcal{P})$

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

Thus if $\limsup_{n\to\infty} \Pr_\Pi\{\mathcal{P} \notin \mathrm{CR}_{1-\gamma}(\mathcal{P})\} \leq \gamma$, we obtain that $\lim_n \Pr_\Pi\{\theta^* \notin [\underline{\theta}, \overline{\theta}]\} \leq \alpha + \gamma$, 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,

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

Q.E.D.

A9.4 Modified Projection Method

The following method projects a confidence region for conditional choice probabilities onto a simultaneous confidence region for all possible ATEs and other structural parameters. In general, this method is more conservative than the perturbed bootstrap method when a single ATE or structural parameter is of interest. We include a more detailed comparison between the two methods at the end of this section.

It is convenient to describe the modified projection method in two stages.

Stage 1. The probabilities $\mathcal{P}_j^k$ belong to the product $\mathcal{S}_J^K$ 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 $P = (P_1^1, ..., P_J^1, ..., P_J^K)' \in \mathcal{S}_J^K$ that pass a goodness-of-fit test:

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

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:

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

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 is compatible with the model. We obtain this projection $P^*(P)$ by solving the minimum distance problem:

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

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

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

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

$$\Delta_\ell^k(P) = \min_{\beta \in B^*(P), F_k \in \mathcal{F}_k(\beta, P^*(P))} \int \Delta(\alpha,\beta) dF_k(\alpha),$$
$$\Delta_u^k(P) = \max_{\beta \in B^*(P), F_k \in \mathcal{F}_k(\beta, P^*(P))} \int \Delta(\alpha,\beta) dF_k(\alpha).$$

Denote the resulting confidence regions as

$$CR_{1-\alpha}[\Delta_\ell^{k*}, \Delta_u^{k*}] = \{[\Delta_\ell^k(P), \Delta_u^k(P)] : P \in CR_{1-\alpha}(\mathcal{P})\}.$$

These sets are confidence regions for the sets $[\Delta_\ell^{k*}, \Delta_u^{k*}]$, where Δ_ℓ^{k*} and Δ_u^{k*} 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 $[\Delta_\ell^k, \Delta_u^k]$ induced by any true model in $(B, \mathcal{P})$.

In a canonical projection method we would implement the second stage by simply intersecting $CR_{1-\alpha}(\mathcal{P})$ with Ξ, but this may give an empty intersection either in finite samples or under misspecification. We avoid this problem by using the projection step instead of the intersection, and also by re-targeting our confidence regions onto the best approximating model.

THEOREM A15: *If Assumptions 5, 8, and 9 are satisfied then for any sequence of data-generating process* $\Pi = \Pi_n$ *satisfying Assumption 10*,

$$\lim_{n\to\infty} \Pr_\Pi \Big[\{\mathcal{P} \in CR_{1-\alpha}(\mathcal{P})\} \cap \{B^* \in CR_{1-\alpha}(B^*)\} \cap \{[\Delta_\ell^{k*}, \Delta_u^{k*}] \in CR_{1-\alpha}[\Delta_\ell^{k*}, \Delta_u^{k*}], \forall k\} \Big] = 1-\alpha.$$

Proof: 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 the 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.

We conclude giving a comparison of the modified projection and perturbed bootstrap methods. 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, the 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.

References

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

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

[3] Chamberlain, G. (1985), "Heterogeneity, Omitted Variables Bias, and Duration Dependence," in J. Heckman and B. Singer eds *Longitudinal Analysis of Labor Market Data.* Cambridge University Press.

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

[5] Cox, D. R. (1958), "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society*, Series B, 20, 215–232.

[6] 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.

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

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

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

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

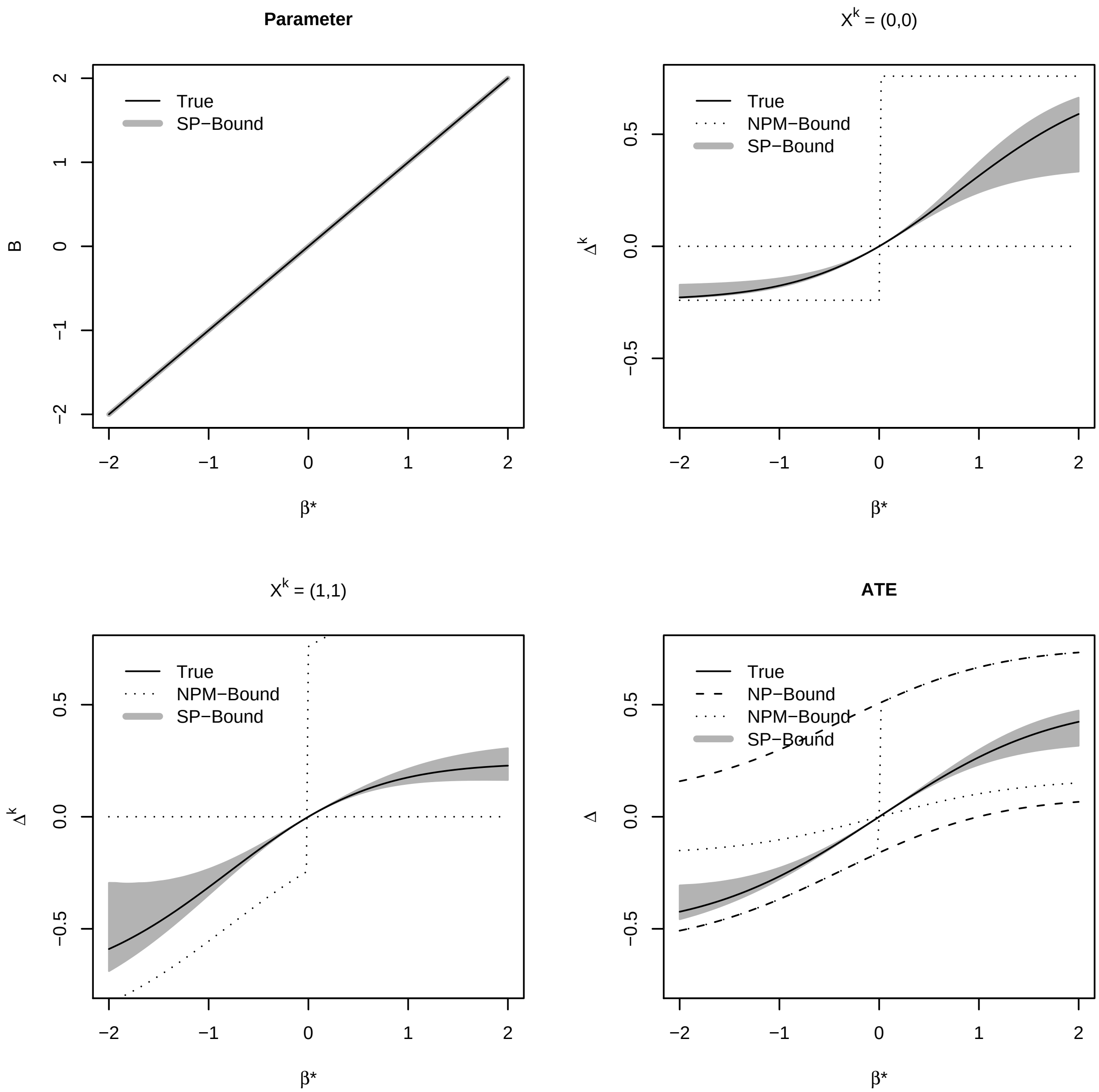

Figure 5: Identified set for parameter and ATEs in binary choice logit models with $Y_{it} = 1(\beta^* X_{it} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim L(0,1)$, $X_{it} = 1(\alpha_i \geq \eta_{it})$, $\eta_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\beta^* \in [-2,2]$, and $T = 2$.

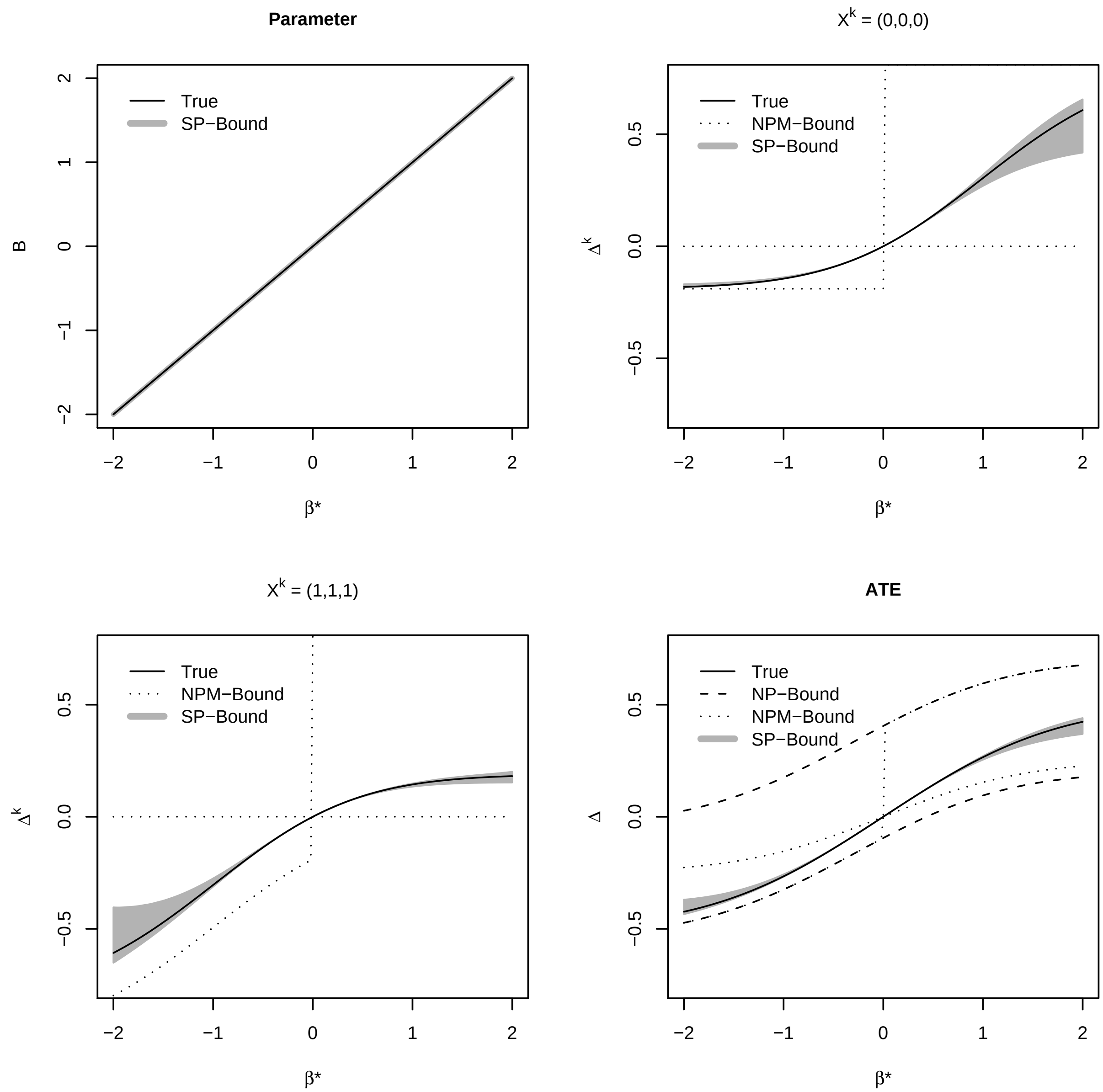

Figure 6: Identified set for parameter and ATEs in binary choice logit models with $Y_{it} = 1(\beta^* X_{it} + \alpha_i \geq \varepsilon_{it})$, $\varepsilon_{it} \sim L(0,1)$, $X_{it} = 1(\alpha_i \geq \eta_{it})$, $\eta_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\beta^* \in [-2,2]$, and $T = 3$.