Partial Identification and Inference in Binary Choice and Duration Panel Data Models*

Jason R. Blevins

Duke University

January 10, 2010

Abstract. Many semiparametric fixed effects panel data models, such as binary choice models and duration models, are known to be point identified when at least one regressor has full support on the real line. It is common in practice, however, to have only discrete or continuous but possibly bounded regressors. This paper addresses identification, estimation, and inference for the identified set in such cases, when the parameters of interest may only be partially identified. We develop a set of general results for criterion-function-based estimation and inference in partially identified models which can be applied to both regular and irregular models. We apply our general results first to a fixed effects binary choice panel data model. We obtain a sharp characterization of the identified set, propose a consistent set estimator, and establish its rate of convergence under different conditions. Rates arbitrarily close to $n^{-1/3}$ are possible when a continuous, but possibly bounded, regressor is present. On the other hand, when all regressors are discrete the estimates converge arbitrarily fast to the identified set. We also propose a subsampling-based procedure for constructing confidence regions and show that it is valid in the models we consider. Finally, we carry out a series of Monte Carlo experiments to illustrate and evaluate the proposed procedures. We also consider extensions to other fixed effects panel data models such as binary choice models with lagged dependent variables and duration models.

Keywords: partial identification, set estimation, panel data, fixed effects, binary choice, duration, discrete regressors, subsampling.

JEL Classification: C13, C14, C25, C41.

1. Introduction

Many economic variables of interest are qualitative in nature and therefore discrete response models have become a standard tool in applied econometrics and their properties

*I am indebted to the members of my dissertation committee, Han Hong, Shakeeb Khan, Paul Ellickson, and Andrew Sweeting, for their invaluable guidance and support. I thank Arie Beresteanu, Federico Bugni, Joseph Hotz and Xiaoxia Shi for many helpful comments and suggestions. This paper also benefited from the feedback of seminar participants at Duke University and the 2009 Triangle Econometrics Conference.

have been studied thoroughly in the econometrics literature (McFadden, 1974; Maddala, 1983; Amemiya, 1985). Semiparametric methods such as maximum score have emerged to estimate such models without tenuous parametric assumptions, however, these methods typically assume the existence of an exogenous explanatory variable with full support (Manski, 1975, 1985; Horowitz, 1992). Similar rank conditions have been successful in estimating more general regression models but the known conditions for point identification still include a full support condition (Han, 1987; Abrevaya, 2000). In practice, however, it is not uncommon to encounter datasets with genuinely discrete or bounded variables. In general, without a regressor with full support on the real line, under semiparametric assumptions the models we consider are only partially identified (Horowitz, 1998).

This paper develops estimators, consistency results, and an inference procedure for a general class of partially identified models with limited support regressors. While the previous literature has focused on partially identified regular models which give rise to set estimators that are essentially $\sqrt{n}$-consistent,[1] this paper provides conditions under which irregular rates of convergence may also arise. Our analysis is motivated by several semiparametric fixed effects panel data models including binary choice and duration models. We apply our general results to these models and show that depending on the assumptions made on the support of the regressors, estimators may achieve nearly cube-root convergence or they may converge arbitrarily fast.

In a broad sense, this paper concerns econometric models characterized by a finite vector of parameters θ which lie in some parameter space Θ. Our particular focus is on semiparametric models which also have unknown infinite-dimensional components, such as the distribution of disturbances, which are not specified a priori. However, to address the concepts of partial identification it suffices to consider a standard parametric model. Suppose that the data generating process, the distribution of observables, is induced by a true parameter $\theta_0 \in \Theta$ which is unknown by the researcher and is the primary object of interest. The model is *point identified* if θ_0 is the only element of Θ such that the model would be consistent with the population distribution P_{θ_0}, assuming for a moment that it were perfectly observable. On the other hand, the model is *partially identified* if there are multiple elements $\theta \in \Theta$ that are observationally equivalent to θ_0, that is, such that $P_\theta = P_{\theta_0}$. The set of all such θ is the *identified set* and is denoted Θ_I. See Manski (2003) and Tamer (2009) for surveys of partial identification in econometric models.

This paper contributes to both the emerging literature on partial identification and the broad literature on nonlinear panel data models. First, it presents general estimation and inference results for two new classes of models: models with continuous but potentially bounded regressors which may have non-standard rates of convergence and models

[1] That is, they can achieve rates arbitrarily close to $1/\sqrt{n}$.

with discrete regressors which are characterized by a discontinuity in the population objective function at the boundary of the identified set. Our results parallel those of Chernozhukov, Hong, and Tamer (2007) in that we propose criterion-function-based set estimators, derive their rates of convergence, and propose an inference procedure based on subsampling. We obtain these results under new conditions which are applicable to the specific models we consider: binary choice panel data models and panel data duration models with discrete or continuous but potentially bounded regressors. Thus, this paper also contributes to the subset of the partial identification literature which is concerned with semiparametric estimation of models with limited support regressors, as well as to the nonlinear semiparametric panel data literature. We provide sharp characterizations of the identified sets of the fixed effects models we consider which are then used to motivate estimators. The consistency and rates of convergence of these estimators are established, as is the validity of subsampling for constructing confidence regions in these models.

This paper is organized as follows. First, Section 2 provides a brief review of the related literature. Then, in Section 3, we formally describe the specific models and assumptions that motivate our analysis. Subsequent sections first introduce general definitions or theorems and then apply them to the panel data binary choice models we consider. In particular, Section 4 focuses on identification, Section 5 discusses consistent estimation and rates of convergence, and Section 6 proposes a subsampling-based algorithm for performing inference in a class of discrete models. We discuss extensions to a class of panel data duration models in Section 7. Several Monte Carlo experiments are described in Section 8 and Section 9 concludes.

2. Related Literature

This paper is related to several topics in the econometrics literature. First, it contributes to a series of papers on criterion-function-based estimation and inference in partially identified models beginning with Manski and Tamer (2002), who consider regression models with interval data. They derive the sharp identified set in a semiparametric binary response model with an interval-valued regressor under a conditional quantile restriction and propose a set estimator which is defined as an appropriately-chosen contour set of a modified maximum score objective function. This estimator is shown to be consistent. They also consider nonparametric estimation as well as modified minimum distance and maximum likelihood estimation of parametric models. Chernozhukov et al. (2007) develop a general framework for criterion-function-based estimation of partially identified models, obtain rates of convergence, and develop an inference procedure based on subsampling (Politis, Romano, and Wolf, 1999). They then apply their general results to models characterized by moment equalities and inequalities. Romano and Shaikh (2008, 2009) further explore subsampling-based inference in partially identified models. Bugni (2008), on the other hand, introduces a bootstrap procedure for performing inference. He also works within the criterion function framework and considers models characterized by a finite number of moment equalities and inequalities.

Another, fundamentally different method for constructing confidence regions in partially identified models is based on set expansion. Expanding the identified set requires a better understanding of its boundary, which is easy to characterize, for instance, when the identified set is an interval on the real line. See Horowitz and Manski (2000) and Imbens and Manski (2004) for examples of the use of set expansion. Beresteanu and Molinari (2008) extend this method to more general settings and develop inference procedures based on the theory of random sets for partially identified models where the identified set can be expressed as the Aumann expectation of a set valued random variable.

There is also a distinction made in the literature, pointed out by Imbens and Manski (2004), between two possible objects of interest: the identified set itself, which is the focus of the present paper, and individual points within the identified set, including the true parameter θ_0. Stoye (2009) observes that the conditions of Imbens and Manski (2004) implicitly assume the existence of a superefficient estimator of the width of the identified interval. He revisits the problem under assumptions that both weaken and remove this condition. Note that although some of the estimators proposed in this paper are superefficient, this occurs as a result of the inherent properties of the model, not due to an implicit assumption.

There are numerous other areas where partially identified econometric models have arisen including, but not limited to games with multiple equilibria (Tamer, 2003; Andrews, Berry, and Jia, 2004; Pakes, Porter, Ho, and Ishii, 2006; Aradillas-Lopez and Tamer, 2008; Ciliberto and Tamer, 2009; Beresteanu, Molchanov, and Molinari, 2009), and models characterized by conditional moment inequalities (Khan and Tamer, 2009; Kim, 2009; Andrews and Shi, 2009).

Of particular relevance to the present paper is a growing literature on semiparametric binary response models with limited support regressors, typically involving either discrete or interval-valued regressors. In terms of cross-sectional models, Bierens and Hartog (1988) show that there are infinitely many single-index representations of the mean regression of a dependent variable when all covariates are discrete. Horowitz (1998) discusses the non-identification of single-index and binary response models with only discrete regressors. Generic non-identification results such as these serve to motivate our analysis.

Manski and Tamer (2002) and Magnac and Maurin (2008) consider partial identification and estimation of binary choice models with an interval-valued regressor. This is a related, but different source of partial identification than those that we consider. Honoré and

Lleras-Muney (2006) estimate partially identified competing risk models with interval outcome data and discrete explanatory variables. Komarova (2008) considers partial identification in static binary response models with discrete regressors. Despite using a different methodology, part of the present paper is similar to her work in that we consider a fixed effects panel extension of the static binary choice model with discrete regressors. However, our analysis differs substantially in that we also consider models with continuous regressors and analyze other unrelated models. Even similarities in the binary choice case are limited since, for example, sharpness of the identified set does not follow directly from the cross-sectional case since we must account for the distribution of the fixed effect in the panel case. Finally, the importance of the theoretical topics addressed in this paper are highlighted by empirical work such as that of Bajari, Fox, and Ryan (2008), who use a maximum-score procedure for estimating a partially identified model of demand for mobile phone service.

Previous papers have considered partial identification in panel data models, with different points of departure and quantities of interest. They highlight the importance of studying the identifying power of various assumptions and provide practitioners with methods to assess the robustness of their results. In particular, Honoré and Tamer (2006) analyze dynamic random effects panel data models and discus how to calculate the identified set using minimum distance, maximum likelihood, and linear programming methods. Chernozhukov, Fernández-Val, Hahn, and Newey (2009) derive bounds on marginal effects in nonlinear panel models with discrete regressors. Rosen (2009) considers partial identification in fixed effects panel data models under conditional quantile restrictions.

This paper is also related to the point-identified fixed effects panel data literature, especially the semiparametric analysis of Manski (1987) for the basic fixed effects model and Honoré and Kyriazidou (2000) for dynamic models with lagged dependent variables. Our characterizations of the identified sets in the models we consider are based in part on known necessary conditions for point identification established in these papers, however, establishing sharpness in partially identified models requires additional work.

3. Models and Assumptions

We consider panel data models where T observations are available at times $t = 0, 1, \ldots, T - 1$ for each individual. An observation is a random vector $(y_0, x_0, u_0, \ldots, y_{T-1}, x_{T-1}, u_{T-1}, c)$ where y_t is a binary response variable in period t, x_t is a vector of k observed explanatory variables, u_t is an unobserved disturbance in period t, and c is a time invariant individual-specific unobserved effect. Let $y = (y_0, \ldots, y_{T-1})$ and define x and u similarly. Let F denote the joint distribution of (y, x, u, c) and let P denote the underlying probability measure generating F. In this case, F_{yx} is the joint distribution of the observed variables.

Our first objective is to combine our knowledge of F_{yx} and a set of weak semiparametric assumptions on F to determine the identified set of parameters of interest. We let θ denote the finite vector of parameters of interest and we will denote the set of possible values of θ by Θ. We assume F is induced by some true unknown parameter θ_0.

In the models we consider, the distribution of the available regressors may not be rich enough to point identify θ_0 without additional assumptions. Therefore, we focus instead on the identified set Θ_I which contains θ_0 itself, as well as all other parameter vectors which cannot be distinguished from θ_0. We address these issues in depth in Section 4.

Our goal is to combine data and prior knowledge about the joint distribution F to learn about θ. First, note that we can always write F as the product of conditional distributions $F = F_{y|xcu}F_{u|xc}F_{c|x}F_x$. In principle, F_x is observable and therefore any restrictions on it should be determined by the data. Much of the literature assumes assumes the presence of at least one component of x, say x_1, which has full support conditional on the remaining components $x_2, \ldots, x_k$. Instead, we consider what can be learned about θ without this assumption in order to develop methods which are appropriate to datasets with only discrete regressors, regressors with compact support, or which otherwise fail to satisfy a full support condition. The present paper focuses on models for which $F_{y|xcu}$ will be fully specified by an underlying latent variable model. For example, in panel data discrete choice models, $F_{y|xcu}$ is determined by a latent variable model. Following the fixed effects literature, $F_{c|x}$ will not be restricted in any way. We will, however, restrict $F_{u|xc}$ with a standard stationarity assumption used in the literature.

3.1. Basic Fixed Effects Panel Data Model

We begin with the fundamental restriction on $F_{y|xcu}$ which defines the basic linear fixed effects binary response model.

Model 1 (Fixed Effects Model)**.** *For all t,*

(1) $$y_t = 1\{x_t'\beta + c + u_t \geq 0\}$$

where x_t is a random variable with support $\mathcal{X} \subset \mathbb{R}^k$, c is a real-valued random variable, and $\theta = \beta$ is the parameter of interest, a member of some parameter space $\Theta \subseteq \mathbb{R}^k$. In addition, for all x and c, $F_{u_t|xc}$ satisfies the following:

a. $F_{u_t|xc} = F_{u_0|xc}$ for all t.

b. The support of u_t is $\mathbb{R}$.

Here, $1\{\cdot\}$ denotes the indicator function, equal to one when the event $\{\cdot\}$ is true and zero otherwise. Condition a above is a substantive restriction, necessary for the estimation methods we introduce below. It requires u_t to be is stationary conditional on the identity of the panel member—that is, conditional on (x, c). Note, however, that it does not restrict the form of serial dependence of u_t in any way. Condition b is a regularity condition which serves to ensure that for any c, the event $y_1 \neq y_0$ occurs with positive probability. Otherwise, the model provides no information about θ.

3.2. Limited Support Regressors

Now, turning to F_x, we begin by reviewing existing conditions for point identification. In the cross-sectional model with a conditional median restriction, analogous to the fixed effects model above, Manski (1985) showed that a full rank, full support condition on x was sufficient to point identify β up to scale. That is, he assumes that x is not contained in a proper linear subspace of $\mathbb{R}^k$ and that the first component of x has positive density everywhere on $\mathbb{R}$ for almost every value of the remaining components. The same conditions were invoked by Han (1987) for the maximum rank correlation estimator and Horowitz (1992) for the smoothed maximum score estimator. The panel version of this assumption (for $T = 2$) was used by Manski (1987) to establish point identification of β up to scale in a semiparametric fixed effects panel data model of the kind considered in the present paper.

Thus, modulo assumptions on the disturbances, point identification of β hinges on the assumptions one is willing to make on the underlying data generating process. The validity of a full support assumption depends critically on the particular explanatory variables available for and relevant to a particular application. It is therefore up to the researcher to determine whether it holds. Many common variables such as age, number of children, years of education, and gender are inherently discrete and so in many cases the decision will be clear. Similarly, many variables such as income have only partial support on the real line (e.g., $\mathbb{R}^+ \subset \mathbb{R}$). The estimators proposed in this paper do not distinguish between the point identified and partially identified cases. They exploit additional information available from regressors with full support if available, but do not require it.

We consider two alternatives to the full support condition. The first applies when x_t is a discrete random variable with finite support. The second applies when at least one component of $x_t - x_{t-1}$ is continuous but may fail to have full support on $\mathbb{R}$.

Assumption 1 (Discrete Regressors)**.** x_t is a discrete random vector with finite support. That is, $|\mathcal{X}| < \infty$, where $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$.

This assumption applies to models which include only genuinely discrete explanatory variables, including indicator variables.

Assumption 2 (Continuous Regressor)**.** The first component of $x_1 - x_0$ has positive density everywhere on a set $\mathcal{W}_1 \subseteq \mathbb{R}$ for almost every value of the remaining components.

Note that this assumption does not rule out the possibility that $\mathcal{W}_1 = \mathbb{R}$ but it also includes cases when the support $x_1 - x_0$ is bounded in some sense. Therefore, this condition includes variables with one-sided support such as income, which is non-negative. As we discuss in detail below, the implications of these two assumptions for estimation are very different.

3.3. Lagged Dependent Variable Model

We also consider a lagged dependent variable model, an extension to the basic fixed effect model which allows for state dependence. Since we do not observe y_t in periods prior to the sample, the model is left unspecified in the first period.

Model 2 (Lagged Dependent Variable Model)**.** *The choice probabilities in the first period are* $P(y_0 = 0 \mid x, c) = p_0(x, c)$, *where* p_0 *is unknown and* $0 < p_0(x, c) < 1$ *for all* x *and* c. *In subsequent periods* $t = 1, \ldots, T$,

$$y_t = 1\{x_t'\beta + \gamma y_{t-1} + c + u_t \geq 0\} \tag{2}$$

where x_t *is a random vector with support* $\mathcal{X} \subset \mathbb{R}^k$, c *is a real-valued random variable, and* $\theta = (\beta, \gamma)$ *are the parameters of interest which lie in some parameter space* $\Theta \subseteq \mathbb{R}^{k+1}$. *In addition, the unobservables* u_t *are serially independent, identically distributed with cdf* $F_{u_t|xc} = F_{u_0|xc}$ *for all* t, *and have full support on* $\mathbb{R}$.

Note that in this model, as opposed to the basic fixed effects model, we do not allow serial correlation in the disturbances. The full support assumption on $u_t \mid x, c$ is a regularity condition which guarantees that certain events used for estimation occur with positive probability.

3.4. Panel Data Duration Models

We also consider estimation of fixed effects panel data versions of a general class of transformation models.

Model 3 (Panel Data Transformation Model)**.** *For all* t,

$$\Lambda(y_t) = x_t'\beta + c + u_t \tag{3}$$

where Λ *is a strictly monotonic function,* x_t *is a random vector with support* $\mathcal{X} \subseteq \mathbb{R}^k$, c *is a real-valued random variable, and* $\theta = \beta$ *is the parameter of interest which lies in some parameter space* $\Theta \subseteq \mathbb{R}^k$. *The disturbances* u_t *are serially independent with identical distribution* $F_{u_0|xc}$ *and independent of* x.

Here, t denotes a single spell. The covariates x_t remain constant within a spell, but vary may across spells. Again, c is a time-invariant individual-specific unobserved variable.

This model is quite general and contains many common duration models in their panel data forms with individual specific time invariant unobserved heterogeneity. For example, the generalized accelerated failure time (GAFT) model of Ridder (1990) is of this form. The mixed proportional hazards model arises when u_t has the minus extreme value distribution with $F_{u_0|xc}(u) = 1 - \exp(-\mathrm{e}^u)$ and Λ is the log integrated baseline hazard function

4. Identification

We begin our identification analysis by developing a broad definition of the identified set in a generic regression model which can later be applied to the specific models we consider. Let F_{yx} denote the joint distribution of (y, x), the observable variables, and v, a vector of unobservables. In Model 1, for example, we have $v = (c, u)$. Let θ be a vector of parameters of interest and let Θ be the parameter space, the set of all feasible values of θ. Assume that we observe the marginal distributions $F_{y|x}$ and F_x, but not F_v. The unknown primitives of the model are thus θ and F_v. Let $\pi(\cdot \mid \theta, F_v, x)$ denote the distribution of $y \mid x$ implied by the model under θ and F_v. The set of primitives that are observationally equivalent to $F_{y|x}$ is thus

$$\Psi(F_{yx}) = \{(\theta, F_v) : \pi(y \mid \theta, F_v, x) = F_{y|x}(y \mid x) \ F_x - \text{a.s.}, y - \text{a.e.}\}.$$

Definition. The *identified set* for θ given $F_{y|x}$ is

$$\Theta_I(F_{yx}) = \left\{\theta \in \Theta : \exists F_v \text{ such that } (\theta, F_v) \in \Psi(F_{yx})\right\}. \tag{4}$$

This set is *sharp* by definition in the sense that each $\theta \in \Theta_I(F_{yx})$ is consistent with F_{yx} and cannot be rejected given the maintained assumptions of the model. Henceforth, we simply write Θ_I, with the dependence on F_{yx} understood.

We also assume throughout that the model is correctly specified: $\Theta_I \neq \emptyset$. See Komarova (2008) for a discussion of misspecification in terms of the closely-related static binary choice model.

Note that we do not rule out cases where point identification obtains. If the model is actually point identified, then our estimates will converge to a point. In practice, models with richer regressor support will have a smaller identified set. Our Θ_I characterizes this set, but when $x_1 - x_0$ has richer support, Θ_I naturally becomes smaller. For example, in the fixed effects model with discrete regressors, considered below, the number of equality constraints defining Θ_I increases with the cardinality of the support of $x_1 - x_0$. Intuitively speaking, when a component is continuous, but perhaps bounded, the number of equalities becomes infinite. When a component has full support, Θ_I collapses to a singleton. This may happen in other situations as well.

4.1. Fixed Effects Model

In Model 1, the primitives of the model are β, $F_{u_0|xc}$, and $F_{c|x}$. We now provide a characterization of Θ_I in terms of observables and show that it is equivalent to the identified set defined above. Since c is unobserved, in order to estimate β we must find implications of the model that are independent of c.

Our identification analysis follows that of Manski (1987). Although our characterization of the identified set is based on a previously known necessary condition for point identification, our characterization of the identified set and the conclusion that it is sharp in this setting are new. The following theorem provides a tractable representation of the identified set, Θ_I, in terms of observables: $P(y_0 = 1 \mid x)$, $P(y_1 = 1 \mid x)$, and F_x.

Theorem 1. *In Model 1,*

$$\Theta_I = \left\{\theta \in \Theta : \operatorname{sgn}\left(P(y_1 = 1 \mid x) - P(y_0 = 1 \mid x)\right) = \operatorname{sgn}\left((x_1 - x_0)'\beta\right) \;\; F_x - a.s.\right\}. \tag{5}$$

Proof. See Appendix D. ■

Henceforth, in discussions of Model 1, we use (5) to characterize the identified set rather than the general definition given in (4).

4.2. Lagged Dependent Variable Model

In this section we turn to the identification of Model 2. Our analysis follows along the lines of Chamberlain (1985) and Honoré and Kyriazidou (2000) and we focus on the case where $T = 4$. Again, although we build on a previously established necessary condition for identification in point identified models, the result that the corresponding set of θ equals the identified set (and is thus sharp) in the present partially identified model is new.

The identification of the model is based on comparing observations for which we observe the same outcome in periods 0 and 3 but different outcomes in periods 1 and 2. Consider the following events for given values of $d_0, d_3 \in \{0, 1\}$:

$$A = \{y_0 = d_0, y_1 = 0, y_2 = 1, y_3 = d_3\},$$
$$B = \{y_0 = d_0, y_1 = 1, y_2 = 0, y_3 = d_3\}.$$

Letting G denote $F_{u_0|xc}$ for simplicity, the corresponding choice probabilities are:

$$\begin{aligned}P(A \mid x, c, x_2 = x_3) &= p_0(x,c)^{1-d_0}(1-p_0(x,c))^{d_0} G(-x_1'\beta - \gamma d_0 - c) \\ &\quad \times \left[1 - G(-x_2'\beta - c)\right] G(-x_2'\beta - \gamma - c)^{1-d_3} \\ &\quad \times \left[1 - G(-x_2'\beta - \gamma - c)\right]^{d_3}, \\ P(B \mid x, c, x_2 = x_3) &= p_0(x,c)^{1-d_0}(1-p_0(x,c))^{d_0} \left[1 - G(-x_1'\beta - \gamma d_0 - c)\right] \\ &\quad \times G(-x_2'\beta - \gamma - c) G(-x_2'\beta - c)^{1-d_3} \\ &\quad \times \left[1 - G(-x_2'\beta - c)\right]^{d_3}.\end{aligned}$$

Note that the latter probability is nonzero since u_t has full support on $\mathbb{R}$ for all t and since $p_0(x,c) > 0$. Dividing, we have

$$\begin{aligned}\frac{P(A \mid x, c, x_2 = x_3)}{P(B \mid x, c, x_2 = x_3)} &= \frac{G(-x_1'\beta - \gamma d_0 - c)}{G(-x_2'\beta - \gamma - c)} \times \frac{1 - G(-x_2'\beta - c)}{1 - G(-x_1'\beta - \gamma d_0 - c)} \\ &\quad \times \left[\frac{G(-x_2'\beta - \gamma - c)}{G(-x_2'\beta - c)}\right]^{1-d_3} \times \left[\frac{1 - G(-x_2'\beta - \gamma - c)}{1 - G(-x_2'\beta - c)}\right]^{d_3}.\end{aligned}$$

When $d_3 = 0$,

$$\frac{P(A \mid x, c, x_2 = x_3)}{P(B \mid x, c, x_2 = x_3)} = \frac{G(-x_1'\beta - \gamma d_0 - c)}{G(-x_2'\beta - \gamma d_3 - c)} \times \frac{1 - G(-x_2'\beta - \gamma d_3 - c)}{1 - G(-x_1'\beta - \gamma d_0 - c)},$$

and when $d_3 = 1$,

$$\frac{P(A \mid x, c, x_2 = x_3)}{P(B \mid x, c, x_2 = x_3)} = \frac{G(-x_1'\beta - \gamma d_0 - c)}{G(-x_2'\beta - \gamma d_3 - c)} \times \frac{1 - G(-x_2'\beta - \gamma d_3 - c)}{1 - G(-x_1'\beta - \gamma d_0 - c)}.$$

We have used the fact that when $d_3 = 0$, $\gamma d_3 = 0$, and when $d_3 = 1$, $\gamma d_3 = \gamma$. In both cases, by the monotonicity of G,

$$P(A \mid x, c, x_2 = x_3) \geq P(B \mid x, c, x_2 = x_3) \iff -x_1'\beta - \gamma d_0 - c \geq -x_2'\beta - \gamma d_3 - c,$$

or equivalently, since this event is independent of c,

$$\operatorname{sgn}\left(P(A \mid x, x_2 = x_3) - P(B \mid x, x_2 = x_3)\right) = \operatorname{sgn}\left((x_2 - x_1)'\beta + \gamma(d_3 - d_0)\right).$$

It turns out that, as in Model 1, this condition is a sharp characterization of the identified set. For any given population distribution of observable data, no θ which satisfies this condition can be rejected. That is, for any such θ, we can choose functions $F_{u_0|xc}$ and $F_{c|x}$ so that the model matches the observed distributions. This is formalized in the following theorem.

Theorem 2. *In Model 2,*

$$\begin{aligned}(6) \quad \Theta_I = \big\{\theta \in \Theta : &\operatorname{sgn}\left(P(A \mid x, x_2 = x_3) - P(B \mid x, x_2 = x_3)\right) \\ &= \operatorname{sgn}\left((x_1 - x_2)'\beta + \gamma(d_3 - d_0)\right) \ F_x - a.s.\ \forall d_0, d_3 \in \{0,1\}\big\}.\end{aligned}$$

Proof. See Appendix E. ■

5. Consistent Estimation

In the remainder of the paper we focus on criterion-function-based estimation and inference. In this section, we first propose consistent estimators for the identified set in a class of models that satisfy a set of general conditions. We also provide rates of convergence for models with objective functions that are either step functions in the limit (e.g., Model 1 under with discrete regressors) or that are bounded by a polynomial in $d(\theta, \Theta_I)$ on regions away from the identified set (e.g., Model 1 with a continuous regressor). In both cases our conditions are new. In the latter case we provide new conditions which allow analysis of irregular models with non-standard rates of convergence. We then verify the conditions of the general theorems in the specific models we consider.

First, we assume that an iid sample is available for use in estimation.

Assumption 3 (Sampling)**.** We observe a iid sample $\{(x_{i,0}, \ldots, x_{i,T-1}, y_{i,0}, \ldots, y_{i,T-1})\}_{i=1}^{n}$ drawn from the joint distribution F_{yx}.

Furthermore, we assume the existence of a population criterion function Q and a finite sample objective function Q_n. These functions must satisfy certain conditions which are stated formally below. A requirement of the population criterion function Q is that the set of parameters at which it attains its maximum must equal the identified set. The analogy principle then suggests estimating Θ_I using the set of maximizers of the sample objective function Q_n. However, in general, taking only the set of maximizers may result in an inconsistent estimator. Instead, we define the estimator $\hat{\Theta}_n(\tau_n)$ to be a contour set of Q_n for some non-negative sequence τ_n:

$$(7) \quad \hat{\Theta}_n(\tau_n) \equiv \left\{ \theta \in \Theta : Q_n(\theta) \geq \sup_{\Theta} Q_n - \tau_n \right\}.$$

The “slackness” sequence τ_n was introduced by Manski and Tamer (2002) and has been used by Chernozhukov et al. (2007), Bugni (2008), Kim (2009), and others. Below, we determine the properties of the sequence τ_n such that $\hat{\Theta}_n$ is a consistent estimator of Θ_I.

To discuss consistency and convergence, we must be precise about which metric space we are working in. We define convergence in terms of the Hausdorff distance, a generalization of Euclidean distance for sets, on the space of all subsets of Θ. Let (Θ, d) be a metric space where d is the standard Euclidean distance. For a pair of subsets $A, B \subset \Theta$, the *Hausdorff distance* between A and B is

$$(8) \quad d_H(A, B) = \max \left\{ \sup_{\theta \in B} d(\theta, A), \sup_{\theta \in A} d(\theta, B) \right\},$$

where, in a slight abuse of notation, $d(\theta, A) \equiv \inf_{\theta' \in A} d(\theta, \theta')$ is the distance between a point θ and a set A. This is illustrated in Figure 1.

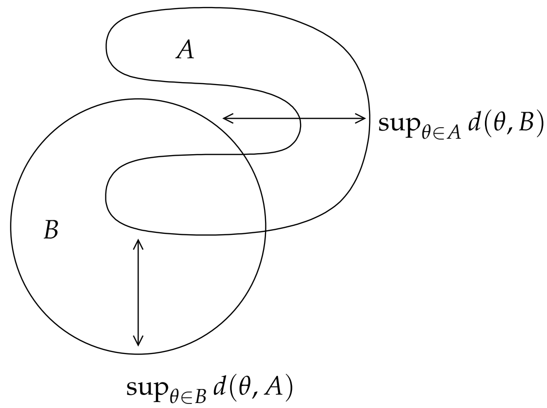

Figure 1: Hausdorff Distance.

5.1. Consistency in General Models

This section develops generic consistency results and rates of convergence. In the following sections, the conditions of these theorems will be verified in the context of the specific models discussed above. We first assume the existence of a population objective function $Q(\theta)$ that fully and exactly characterizes the identified set Θ_I. Using the analogy principle, we then use the finite sample objective function $Q_n(\theta)$ to obtain a set estimator $\hat{\Theta}_n$. Finally, we shall prove that the sequence of set estimates $\hat{\Theta}_n$ converges in probability to the identified set Θ_I in the Hausdorff metric and obtain rates of convergence under different assumptions on the curvature of the objective function.

Assumption 4. Suppose the following conditions are satisfied:

a. Θ is a nonempty subset of $\mathbb{R}^k$ and is compact with respect to the Euclidean metric.

b. There exists a function $Q : \Theta \to \mathbb{R}$ such that $\arg\max_\Theta Q = \Theta_I$.

c. Q has a well-separated maximum in that for all $\varepsilon > 0$ there exists a $\delta_\epsilon > 0$ such that $\sup_{\Theta\setminus\Theta_I^\varepsilon} Q \leq \sup_\Theta Q - \delta_\varepsilon$.

d. There exists a function $Q_n : \Theta \times \mathcal{X}^T \times \mathcal{Y}^T \to \mathbb{R}$, denoted $Q_n(\theta)$, which converges uniformly in probability to Q at the $1/b_n$ rate. That is, $\sup_\Theta |Q_n - Q| = O_p(1/b_n)$ for some sequence $b_n \to \infty$.

Part c is a regularity condition which rules out pathological cases that can arise without a continuity assumption. It is satisfied in the models we consider, for example, when Q is continuous or when Q is a step function.

Theorem 3 (Consistency in General Models)**.** *Suppose Assumption 4 holds.*

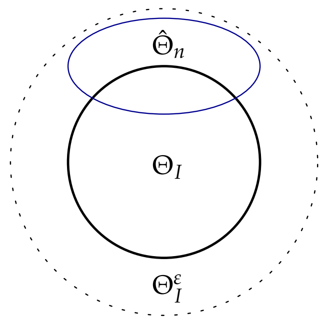

Figure 2: One-sided consistency without slackness.

1. *If* $\tau_n \xrightarrow{p} 0$*, then* $\sup_{\theta \in \hat{\Theta}_n} d(\theta, \Theta_I) \xrightarrow{p} 0$.

2. *If* $\tau_n \xrightarrow{p} 0$ *and* $\tau_n b_n \xrightarrow{p} \infty$*, then* $\lim_{n \to \infty} P(\Theta_I \subseteq \hat{\Theta}_n) = 1$.

If both conditions hold then $d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$.

Proof. See Appendix B. ■

Note that the first conclusion of Theorem 3 actually holds in general without slackness (i.e., with $\tau_n = 0$). This is formalized in the following corollary.

Corollary (One-Sided Consistency Without Slackness)**.** *Suppose Assumption 4 holds. If* $\tau_n = 0$*, then* $\sup_{\theta \in \hat{\Theta}_n} d(\theta, \Theta_I) \xrightarrow{p} 0$.

This corollary guarantees that asymptotically, without slackness, $\hat{\Theta}_n$ is close to Θ_I. The converse need not be true in general since $\sup_{\theta \in \Theta_I} d(\theta, \hat{\Theta}_n)$ may be large, as illustrated by Figure 2.

5.2. Rates of Convergence in General Models

The rate of convergence of the Hausdorff distance $d_H(\hat{\Theta}_n, \Theta_I)$ is the slowest rate at which the component distances $\sup_{\theta \in \Theta_I} d(\theta, \hat{\Theta}_n)$ and $\sup_{\theta \in \hat{\Theta}_n} d(\theta, \Theta_I)$ converge to zero. The second part of Theorem 3 establishes that with only Assumption 4, the first distance converges arbitrarily fast to zero in probability because with probability approaching one, $\Theta_I \subseteq \hat{\Theta}_n$.

The rate of convergence of the second component depends on the shape of the objective function. In the specific models we consider this shape depends in turn on the support of x_t. In this section, however, we prove general results by making assumptions about Q and Q_n. In later sections we provide conditions on the support of x_t that imply the required properties of these functions.

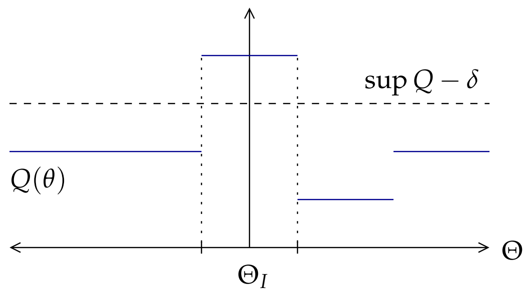

FIGURE 3: Infinite curvature of $Q(\theta)$.

In particular, we show that when Q has a discrete jump at the boundary of Θ_I, then $\hat{\Theta}_n$ converges arbitrarily fast in probability to Θ_I. That is, for *any* sequence r_n, including powers of n and exponential forms, $r_n d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$. This result also implies that $\hat{\Theta}_n = \Theta_I$ with probability approaching one.

On the other hand, when $Q_n(\theta)$ is bounded from above by a polynomial in the distance between θ and Θ_I, we show that the rate of convergence of $\sup_{\theta \in \hat{\Theta}_n} d(\theta, \Theta_I)$ depends on both the curvature of the bounding polynomial and the rate at which τ_n converges to zero.

We begin with models that satisfy the following assumption, where Q exhibits a discrete jump at Θ_I:

Assumption 5 (Existence of a Constant Majorant)**.** There exists a positive constant δ

$$Q(\theta) \leq \sup_{\Theta} Q - \delta$$

for all $\theta \in \Theta \setminus \Theta_I$.

When the above condition holds, $\hat{\Theta}_n$ converges arbitrarily fast to Θ_I. This result is due to the discrete jump in Q at the boundary of Θ_I. As we will see later, this can happen when the regressors in a binary response model have discrete support. We present the theorem first, followed by a discussion of the intuition.

Theorem 4. *Suppose Assumptions 4 and 5 hold. If $\tau_n \xrightarrow{p} 0$ and $\tau_n b_n \xrightarrow{p} \infty$, then for any sequence r_n, $r_n d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$.*

Proof. See Appendix B ∎

Figure 4 illustrates the notion that, due to the discrete nature of $Q_n(\theta)$, there are only a finite, though potentially very large, number of possible estimates $\hat{\Theta}_n$. For the realization of Q_n in the figure, the contour sets of Q_n determine a partition of Θ into four disjoint sets: $\Theta = \Theta_1 \cup \Theta_2 \cup \Theta_3 \cup \Theta_4$. In the present framework, where $\hat{\Theta}_n$ is defined by a threshold τ_n,

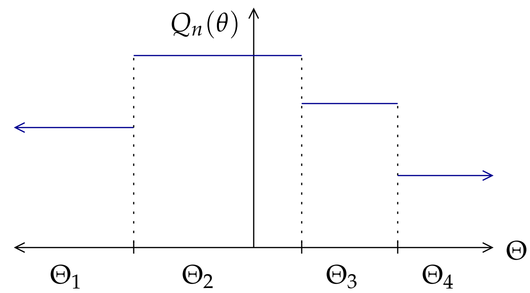

FIGURE 4: A realization of Q_n and the partition of Θ it generates.

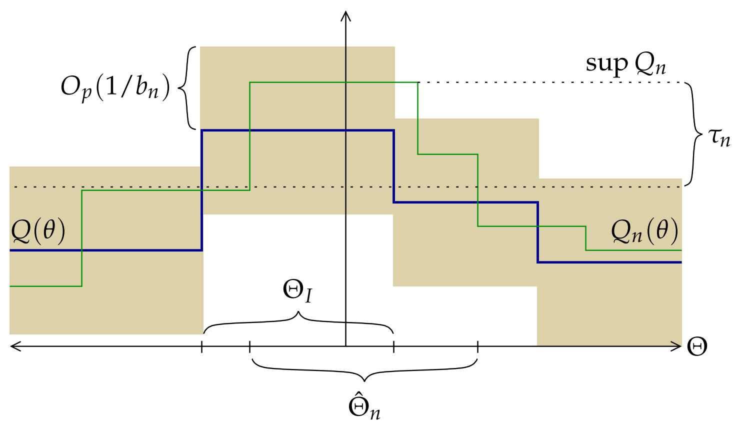

FIGURE 5: Convergence of Q_n to Q while $1/b_n \to 0$, $\tau_n \xrightarrow{p} 0$, and $b_n \tau_n \xrightarrow{p} \infty$.

so that it includes all values of θ for which $Q_n(\theta) \geq \sup Q_n - \tau_n$, there are four possible estimates: Θ_2, $\Theta_2 \cup \Theta_3$, $\Theta_2 \cup \Theta_3 \cup \Theta_1$, and $\Theta_2 \cup \Theta_3 \cup \Theta_1 \cup \Theta_4$. In higher dimensions, and for large sample sizes, the combinatorics of the problem dictate that the number of possibilities becomes large very quickly. On the other hand, as $n \to \infty$, the contour sets of Q_n approach those of Q, and the set of possible estimates contains a set equal to Θ_I with probability approaching one. Intuitively, as we obtain more data, we are able to detect which values of θ belong to Θ_I with increasing accuracy since there is a discrete jump in $Q(\theta)$ for all θ not in Θ_I. Furthermore, since τ_n converges to zero in probability slower than Q_n converges uniformly to Q, $\hat{\Theta}_n$ converges to Θ_I. This is illustrated in Figure 5.

We now consider models for which Q and Q_n may be smooth, but which satisfy a curvature condition such that, outside of a shrinking neighborhood of Θ_I, Q_n is bounded by a polynomial in the distance from the identified set. This condition is analogous to conditions used to obtain rates of convergence in point identified models.

Assumption 6 (Existence of a Polynomial Majorant)**.** There exist positive constants $(\delta, \kappa, \gamma_1, \gamma_2)$ with $\gamma_1 \geq \gamma_2$ such that for any $\varepsilon \in (0, 1)$ there are $(\kappa_\varepsilon, n_\varepsilon)$ such that for all $n \geq n_\varepsilon$,

$$Q_n(\theta) \leq \sup_{\Theta} Q_n - \kappa \cdot (d(\theta, \Theta_I) \wedge \delta)^{\gamma_1}$$

uniformly on $\{\theta \in \Theta : d(\theta, \Theta_I) \geq (\kappa_\varepsilon / b_n)^{1/\gamma_2}\}$ with probability at least $1 - \varepsilon$.

Theorem 5. *Suppose Assumptions 4 and 6 hold. If $\tau_n \xrightarrow{p} 0$ and $\tau_n b_n \xrightarrow{p} \infty$, then $d_H(\hat{\Theta}_n, \Theta_I) = O_p(\tau_n^{1/\gamma_2})$.*

Proof. See Appendix B ∎

5.3. Fixed Effects Models

In this section we focus on consistent estimation of Model 1. We first propose population and finite sample criterion functions and show that the population criterion function characterizes the identified set exactly. Then, we verify the conditions of Theorem 3, making use of empirical process techniques, to show that the estimator is consistent. Finally, we obtain the rate of convergence in two cases: models with only discrete regressors, under Assumption 1, and models with a continuous regressors, under assumption Assumption 2. In these cases we verify, respectively, the assumptions for Theorem 4 and Theorem 5.

5.3.1. Objective Function

The population objective function we propose for use in estimating Model 1 is the maximum score objective function of Manski (1987), a panel data analog of the cross-sectional maximum score objective function of Manski (1975, 1985):

$$Q(\theta) = \mathrm{E}\left[(y_1 - y_0)\,\mathrm{sgn}\left((x_1 - x_0)\beta\right)\right].$$

The corresponding finite sample analog objective function is

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - y_{i0})\,\mathrm{sgn}\left((x_{i1} - x_{i0})\beta\right).$$

Note that although the same objective function is for maximum score estimation in the point identified case, the set estimators proposed here are fundamentally different since they are contour sets of this function. Also, note that $Q(\theta)$ and $Q_n(\theta)$ effectively condition on the event $y_1 \neq y_0$. This does not result in a loss of efficiency since, as established by Theorem 1, the event $y_1 = y_0$ is not informative about θ.

Lemma 1 below establishes the equivalence between the identified set Θ_I and the set of maximizers of the population objective function.

Lemma 1. *Under the maintained assumptions of Model 1,*

$$\arg\max_{\theta \in \Theta} Q(\theta) = \Theta_I.$$

Proof. See Appendix D. ■

5.3.2. Consistency

We verify each of the conditions of Assumption 4 in order to use the general consistency result of Theorem 3. In doing so, we will make use of empirical process concepts such as the subgraph of a function, Vapnik-Chervonenkis (VC) classes of sets, and Euclidean classes of functions. We refer the reader to Section 2 of Pakes and Pollard (1989) for definitions and further details. Essentially, we construct a class of functions $\mathcal{F}$, indexed by Θ, such that $Q(\theta) = Pf_\theta$ and $Q_n(\theta) = P_n f_\theta$ for $f_\theta \in \mathcal{F}$. We begin by defining $\mathcal{F}$ and establishing that it is Euclidean.

Lemma 2. *Let $f(z, w, \theta) = z \cdot (2 \cdot 1\{w'\beta \geq 0\} - 1)$. Then, the class $\mathcal{F} = \{f(\cdot, \cdot, \theta) : \theta \in \Theta\}$ is Euclidean for the constant envelope $F = 1$.*

Proof. See Appendix D. ■

Now that we have established that the objective function is generated by an underlying Euclidean class of functions, we can use tools from empirical process theory to establish the uniform convergence required for consistency. In particular, we make use of a result from Kim and Pollard (1990) to establish uniform convergence of Q_n to Q at the rate $1/b_n$ with $b_n = n^{-1/2}$.

Lemma 3 (Uniform Convergence of Q_n to Q)**.** *Under Assumption 3,*

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| = O_p(n^{-1/2}).$$

Proof of Lemma 3. $\mathcal{F}$ is Euclidean, so it is also manageable in the sense of Pollard (1989) (cf. Pakes and Pollard, 1989, p. 1033). Since $\int F^2 \, dP = 1 < \infty$, the result follows from Corollary 3.2 of Kim and Pollard (1990). ■

Finally, combining the above results, we can apply Theorem 3 to establish consistency of $\hat{\Theta}_n$ for Model 1.

Theorem 6. *Suppose Assumption 3 holds in Model 1. If $\tau_n \xrightarrow{p} 0$, and $\tau_n n^{1/2} \xrightarrow{p} \infty$, then $d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$.*

Proof. See Appendix D. ■

5.3.3. Rates of Convergence

The rate of convergence of $\hat{\Theta}_n$ to Θ_I in Model 1 depends on the support of x_t. We obtain the rate under both Assumption 1 and Assumption 2. We show that when the support of x_t is finite, $\hat{\Theta}_n$ converges arbitrarily fast in probability to Θ_I. On the other hand, when at least one component of $x_2 - x_1$ is continuous, the estimator can achieve rates arbitrarily close to $n^{-1/3}$. The rate depends on τ_n and, although the exact rate $n^{-1/3}$ is not achievable, in practice, one can achieve convergence close to $n^{-1/3}$ by choosing, for example, $\tau_n \propto \sqrt{\ln n / n}$.

Discrete Regressors Here, we verify Assumption 5, the constant majorant condition, in the context of Model 1. We can then apply Theorem 4 to show that in this case, $\hat{\Theta}_n$ converges arbitrarily fast to Θ_I.

When the support of (x_0, x_1) is a finite set, henceforth $\mathcal{X}$, the objective function $Q(\theta)$ can be rewritten as follows:

$$
\begin{aligned}
Q(\theta) &= \mathrm{E}_x \, \mathrm{E}_{y|x} \left[(y_1 - y_0) \operatorname{sgn}\left((x_1 - x_0)'\beta \right) \right] \\
&= \sum_{x \in \mathcal{X}} P(x) \left[P(y_1 = 1 \mid x) - P(y_0 = 1 \mid x) \right] \operatorname{sgn}\left((x_1 - x_0)'\beta \right).
\end{aligned}
$$

Therefore, $Q(\theta)$ is a step function and there exists a real number $\delta > 0$ such that for all $\theta \in \Theta \setminus \Theta_I$, $Q(\theta) \leq \sup_\Theta Q - \delta$. In particular, δ is bounded below by the smallest nonzero value of $P(x) \left[P(y_1 = 0 \mid x) - P(y_0 = 1 \mid x) \right]$ for any $x \in \mathcal{X}$. Thus, applying Theorem 4, we have the following result.

Theorem 7. *Suppose that Assumption 1 holds in Model 1. For any sequence τ_n such that $\tau_n \xrightarrow{p} 0$ and $n^{1/2}\tau_n \xrightarrow{p} \infty$, then $\hat{\Theta}_n$ converges to Θ_I arbitrarily fast in probability in the Hausdorff metric. That is, for any sequence r_n, $r_n d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$.*

Continuous Regressors The properties of the maximum score objective function with continuous covariates have been studied carefully by Kim and Pollard (1990), Abrevaya and Huang (2005), and others. We follow Abrevaya and Huang (2005) in restricting the coefficient on one component of x, henceforth x_d, to be either 1 or -1 and consider β to be a vector in $\mathbb{R}^{k-1}$. Let $\tilde{x}$ denote the remaining components of x.[2]

Kim and Pollard's heuristic for cube root convergence translates almost directly to the set identified case. Let $\Gamma(\theta) \equiv Q(\theta) - Q(\theta_0)$ and $\Gamma_n(\theta) \equiv Q_n(\theta) - Q_n(\theta_0)$. We can decompose $\Gamma_n(\theta)$ into two components, a trend and a stochastic component: $\Gamma_n(\theta) =$

[2] Alternatively, Kim and Pollard (1990) work with parameters in unit sphere $\mathbb{S}^{k-1} \equiv \{x \in \mathbb{R}^k : \|x\| = 1\}$ in $\mathbb{R}^k$ and assume that the angular component of x has continuous, bounded density with respect to the surface measure on $\mathbb{S}^{k-1}$.

$\Gamma(\theta) + [\Gamma_n(\theta) - \Gamma(\theta)]$. The limiting objective function is approximately quadratic near the identified set: $\Gamma(\theta) = O(d^2(\theta, \Theta_I))$. The variance of the empirical process component is $O_p(d(\theta, \Theta_I)/n)$. When the trend overtakes the noise, Γ_n very likely to be below the maximum. Thus, the maximum is likely to occur when the standard deviation of the random component is of the same magnitude or larger than the trend. That is, when $\sqrt{d(\theta, \Theta_I)/n} > d^2(\theta, \Theta_I)$, or, $d(\theta, \Theta_I) < n^{-1/3}$. Therefore, $\hat{\Theta}_n$ the set of near maximizers of Γ_n, should be within an $n^{-1/3}$ neighborhood of Θ_I. In the set identified case, this is only one component of the distance. The other component, however, was shown to converge arbitrarily fast and therefore does not hinder the rate of convergence.

In terms of Theorem 5, the above argument corresponds to the case where $\gamma_1 = 2$ and $\gamma_2 = 3/2$. Since τ_n can be chosen arbitrarily close to $n^{-1/2}$, the rate of convergence can be made arbitrarily close to $(n^{-1/2})^{1/\gamma_2} = n^{-1/3}$. The following theorem formalizes this result. We also need several assumptions on the distribution of x, which are intentionally close to those made by Abrevaya and Huang (2005) in analyzing the cross-sectional model in the point identified case.

Let $w \equiv x_1 - x_0$ and $v \equiv u_1 - u_0$. Let F and f denote cdf and density of v and let G and g denote the cdf and density of w. Finally, let w_1 denote the first component of w and let $\tilde{w}$ denote the remaining $k-1$ components.

Theorem 8. *Suppose that Assumptions 2 and 3 hold in Model 1. In addition, suppose the following:*

a. The components of $\tilde{w}$ and $\tilde{w}\tilde{w}'$ have finite first absolute moments.

b. The function $g'(w_1 \mid \tilde{w})$ exists and, for some $M > 0$, $|g'(w_1 \mid \tilde{w})| < M$ and $|g(w_1 \mid \tilde{w})| < M$ for all w_1 and almost every $\tilde{w}$.

c. For all v in a neighborhood of 0, all w_1 in a neighborhood around $-\tilde{w}'\beta_0$, almost every $\tilde{w}$, and some $M > 0$, the function $f(v \mid \tilde{w}, w_1)$ exists and $f(v \mid \tilde{w}, w_1) < M$.

d. For all v in a neighborhood of 0, all w_1 in a neighborhood of $-\tilde{w}'\beta_0$, almost every $\tilde{w}$, and some $M > 0$, the function $\partial F(v \mid \tilde{w}, w_1)/\partial w_1$ exists and $|\partial F(v \mid \tilde{w}, w_1)/\partial w_1| < M$.

e. Θ_I is contained in the interior of Θ.

f. The matrix $V(\theta) \equiv \mathrm{E}\left[2f(0 \mid \tilde{w}, -\tilde{w}'\beta)g(-\tilde{w}'\beta \mid \tilde{w})\tilde{w}\tilde{w}'\right]$ is positive semidefinite for all $\theta \in \mathrm{bd}(\Theta_I)$.

Then for any sequence τ_n such that $\tau_n \xrightarrow{p} 0$ and $n^{1/2}\tau_n \xrightarrow{p} \infty$, $d_H(\hat{\Theta}_n, \Theta_I) = O_p(\tau_n^{2/3})$.

Proof. See Appendix D ■

5.4. Lagged Dependent Variable Model

In this section, we propose a consistent estimator for Model 2. The proofs of the results in this section largely parallel those for the fixed effects model and therefore all proofs are reserved for Appendix E. For simplicity we only consider the lagged dependent variable model under Assumption 1 (discrete regressors). An extension to Assumption 2 (continuous regressor) would involve the use of a kernel as in Honoré and Kyriazidou (2000), along with the associated assumptions. The kernel is used to condition on the event $x_3 = x_2$ and $x_3 - x_2$ is assumed to support in a neighborhood of zero. In the case of discrete regressors, this conditioning is accomplished with a simple indicator function.

We use the population objective function

$$Q(\theta) = \mathrm{E}\left[1\{x_2 = x_3\} \cdot (y_2 - y_1) \cdot \mathrm{sgn}((x_2 - x_1)'\beta + \gamma(y_3 - y_0))\right].$$

This function was used by Honoré and Kyriazidou (2000) for estimation in point identified models. The finite sample objective function is

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} 1\{x_{i2} = x_{i3}\} \cdot (y_{i2} - y_{i1}) \cdot \mathrm{sgn}((x_{i2} - x_{i1})'\beta + \gamma(y_{i3} - y_{i0})).$$

The set of maximizers of Q is indeed a sharp characterization of the identified set, as established by the following Lemma.

Lemma 4 (Objective Function Representation of Θ_I)**.** *Under the maintained assumptions of Model 2*, $\arg\max_{\theta\in\Theta} Q(\theta) = \Theta_I$.

Proof. See Appendix E. ■

Next, we verify each of the conditions of Assumption 4 in order to use Theorem 3 to establish consistency of the estimator $\hat{\Theta}_n$. As in the fixed effects model, we begin by establishing that the objective function belongs to a Euclidean class of functions indexed by θ so that we can leverage results from empirical process theory.

Lemma 5 (Euclidean Property)**.** *The class of functions $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, where $f_\theta(x, y) = 1\{x_2 = x_3\}(y_2 - y_1)\left[2 \cdot 1\{(x_2 - x_1)'\beta + \gamma(y_3 - y_0) \geq 0\} - 1\right]$, is Euclidean for the constant envelope $F = 1$.*

Proof. See Appendix E. ■

As before, the Euclidean property allows us to immediately establish uniform convergence and the P-Donsker property which we will in turn use to show consistency and, later, the conditions required by our inference procedure.

Theorem 9. *Suppose Assumption 3 holds in Model 2. If* $\tau_n \xrightarrow{p} 0$ *and* $\tau_n n^{1/2} \xrightarrow{p} \infty$*, then* $d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$.

Proof. See Appendix E. ∎

Additionally, when Assumption 1 is satisfied, Q is again a step function. The argument is analogous to that for the basic fixed effects model and is reserved for the proof. Thus, applying Theorem 4, we again find that $\hat{\Theta}_n$ converges arbitrarily fast to Θ_I in probability.

Theorem 10. *Suppose that Assumptions 1 and 3 hold in Model 2. If* $\tau_n \xrightarrow{p} 0$ *and* $n^{1/2}\tau_n \xrightarrow{p} \infty$*, then for any sequence* r_n*,* $r_n d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$.

Proof. See Appendix E. ∎

6. Confidence Regions

Confidence regions for Θ_I can be formed using contour sets of Q_n in much the same way as we defined the estimator $\hat{\Theta}_n$ in (7). Let $C_n(\kappa_n)$ denote the set

$$C_n(\kappa_n) = \{\theta \in \Theta : b_n Q_n(\theta) \geq \sup_{\Theta} b_n Q_n - \kappa_n\}. \tag{9}$$

Inference is based on the statistic

$$\mathcal{Q}_n \equiv \sup_{\Theta} b_n Q_n - \inf_{\Theta_I} b_n Q_n$$

and the following equivalence:

$$\{\Theta_I \subseteq C_n(\kappa_n)\} \iff \{\mathcal{Q}_n \leq \kappa_n\}.$$

The sets $C_n(\kappa_n)$ defined in (9) have the same form as (7), except that the objective function is now normalized by b_n, the rate of uniform convergence. We apply this normalization in order to use subsampling to approximate quantiles of $\mathcal{Q}_n$. As a result, the sequence κ_n is analogous to $b_n \tau_n$. Thus, while in Theorem 3 we required $\tau_n \xrightarrow{p} 0$ and $b_n \tau_n \to \infty$ for consistent estimation using (7), we could obtain consistent estimates with (9) if $\kappa_n \xrightarrow{p} \infty$ and $\kappa_n / b_n \xrightarrow{p} 0$. That is, κ_n approaches infinity at a rate slower than that of b_n.

For smooth models, where $\hat{\Theta}_n$ converges at a polynomial rate and where the limiting distribution of $\mathcal{Q}_n$ is continuous, Chernozhukov et al. (2007) provide methods of constructing confidence regions which cover Θ_I asymptotically with probability $1 - \alpha$ using subsampling. Their results are not applicable to the models we consider with discrete regressors due to the discrete nature of $\mathcal{Q}_n$. Instead, in the following sections, we provide conditions under which one can obtain *conservative* asymptotic confidence regions with coverage probability *at least* $1 - \alpha$.

Our confidence regions are based on estimates of quantiles of $\mathcal{Q}$. To understand why the confidence regions we propose are conservative, consider the cdf and quantile functions of a generic discrete random variable X depicted in Figure 6. There, for example, the 0.50 and 0.75 quantiles are equal. If we use the x_2, the 0.50 quantile in an attempt to form a 50% confidence region, the coverage will actually be over 75%.

6.1. Confidence Regions in General Discrete Models

For now, we assume the availability of a consistent estimate $\hat{c}_n$ of the corresponding $1-\alpha$ quantile of $\mathcal{Q}$, the limiting distribution of $\mathcal{Q}_n$. In the following section, we describe an algorithm to construct such a sequence. Large sample inference with discrete regressors is based on the following lemma.[3] We require only that $\mathcal{Q}_n$ has a nondegenerate limiting distribution.

Assumption 7 (Convergence of $\mathcal{Q}_n$)**.** Suppose that $P\{\mathcal{Q}_n \leq c\} \to P\{\mathcal{Q} \leq c\}$ for each $c \in \mathbb{R}$, where $\mathcal{Q}$ has a nondegenerate distribution function on $\mathbb{R}$.

Lemma 6. *Suppose Assumption 7 holds. Then, for any sequence $\hat{c}_n$ such that $\hat{c}_n \xrightarrow{\mathrm{p}} c(1-\alpha) \equiv \inf\{c : P\{\mathcal{Q} \leq c\} \geq 1-\alpha\}$ for some $\alpha \in (0,1)$,*

$$P\{\Theta_I \subseteq C_n(\hat{c}_n)\} \geq (1-\alpha) + o_p(1).$$

Proof. See Appendix C. ■

An appropriate sequence $\hat{c}_n$, and corresponding conservative confidence regions $C_n(\hat{c}_n)$ with asymptotic coverage probability of at least $1-\alpha$, can be constructed using the following algorithm.

Algorithm 1.

1. Choose a subsample size $m < n$ such that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$. Let M_n denote the number of subsets of size m and let κ_n be any sequence such that such that $C_n(\kappa_n)$ is a consistent estimator of Θ_I (e.g., $\kappa_n \propto \sqrt{\ln n}$).
2. Compute $\hat{c}_n$ as the $1-\alpha$ quantile of the values $\{\hat{\mathcal{Q}}_{n,m,j}\}_{j=1}^{M_n}$ where

$$\hat{\mathcal{Q}}_{n,m,j} \equiv \sup_{\theta\in\Theta} b_m Q_{n,m,j}(\theta) - \inf_{\theta\in C_n(\kappa_n)} b_m Q_{n,m,j}(\theta)$$

and $Q_{n,m,j}$ denotes the sample objective function constructed using the j-th subsample of size m.

[3] Lemma 6 is the discrete-distribution analog of Lemma 3.1 of Chernozhukov et al. (2007). The fundamental difference is that here, the distribution of $\mathcal{Q}$ may not be continuous. As a result, our confidence regions are conservative since we cannot place an upper bound on the coverage probability.

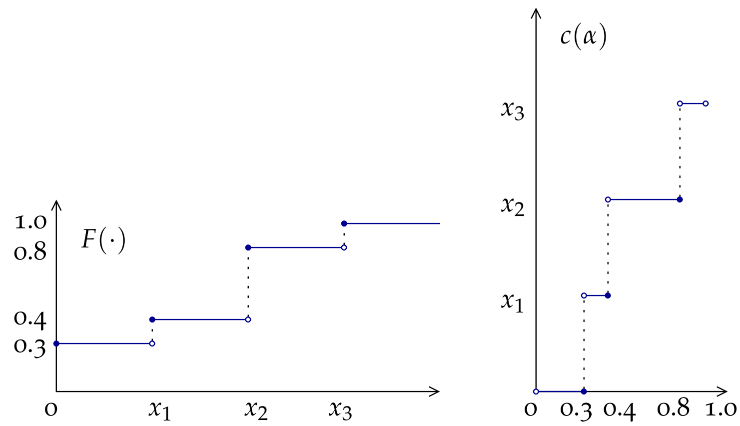

FIGURE 6: Cumulative distribution function and quantile function for a discrete distribution.

3. Report $C_n(\kappa_n)$ as a consistent estimate of Θ_I and $C_n(\hat{c}_n)$ as a conservative confidence region.

The following theorem addresses the validity of this algorithm for obtaining the desired sequence $\hat{c}_n$. Let $a_n \downarrow a$ denote a sequence which eventually equals a, or in other words, a sequence which converges arbitrarily fast to a.[4]

Assumption 8 (Approximability of $\mathcal{Q}_n$)**.** Let Θ_n be a sequence of subsets of Θ such that $d_H(\Theta_n, \Theta_I) \downarrow 0$ in probability and let $\mathcal{Q}'_n = \sup_{\Theta} b_n Q_n - \inf_{\Theta_n} b_n Q_n$. Then $P(\mathcal{Q}'_n \leq c) \to P(\mathcal{Q} \leq c)$ for each $c \in \mathbb{R}$.

Theorem 11. *Suppose that Assumptions 3, 4, 5, 7 and 8 hold and that $m \to \infty$, and $m/n \to 0$ as $n \to \infty$. Let $1 - \alpha$ denote the desired coverage level, where the distribution of $\mathcal{Q}$ is continuous at $c(1-\alpha)$. Then, $\hat{c}_n \xrightarrow{p} c(1-\alpha)$.*

Proof. See Appendix C ■

6.2. Fixed Effects Model

In this section we verify the conditions required for constructing confidence regions in the context of Model 1 under Assumption 1 (discrete regressors). The following lemma verifies both the convergence of $\mathcal{Q}_n$ required by Assumption 7 and the approximability of $\mathcal{Q}_n$ based on a sequence of estimates $\hat{\Theta}_n$, required by Assumption 8. Thus, this result establishes the validity of Algorithm 1 for constructing conservative confidence regions.

Lemma 7. *In Model 1 under Assumptions 1, and 3, both Assumptions 7 and 8 are satisfied.*

[4]See Appendix A.1 for a precise definition of $a_n \downarrow a$, both deterministically and in probability.

Proof. See Appendix D ■

6.3. Lagged Dependent Variable Model

For the case of Model 2 with discrete regressors, the arguments to establish the validity of the subsampling procedure of Algorithm 1 are identical to those of the previous section for Model 1. This follows since both objective functions are of the same form in the underlying functions f_θ and both functions satisfy Assumption 5. That is, in both cases, for the appropriate class of functions $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, we have $Q(\theta) = Pf_\theta$ and $Q_n(\theta) = P_n f_\theta$. Since both classes of functions $\mathcal{F}$ are Euclidean, it follows that Lemma 7 also applies to Model 2 under Assumption 1.

7. Panel Data Duration Models

This section considers Model 3 (defined on page 8), the fixed effects panel data duration model. Identification of this model and similar ones has been considered by a number of authors under a wide variety of conditions. For example, Ridder (1990) considers the nonparametric identification of the generalized accelerated failure time (GAFT) model, which contains both the mixed proportional hazards (MPH) model and the accelerated failure time (AFT). He shows that GAFT models are nonparametrically identified (up to an obvious normalization) with continuous duration data (and continuous covariates). Furthermore, it is identified even with discrete duration data with an additional parametric assumption on the regression function. We consider a similar model, when the observed durations are continuous but the covariates are discrete. Han (1987), Chen (2002), Abrevaya (2000) and others have considered point identification and estimation of various components of generalized regression models, which contain models of this type, but such studies are based on a full-support condition which we relax. Honoré and Lleras-Muney (2006) consider partial identification of a related competing risks model.

In many ways, this model is very similar to Model 1 and so many of the results will be familiar. When the disturbances are independent, we can carry out a similar ranking procedure relating the ordering of y_1 and y_0 to that of $x_1'\beta$ and $x_0'\beta$:

$$\begin{aligned}
&P(y_1 \geq y_0 \mid x, c) \geq P(y_0 \geq y_1 \mid x, c) \\
&\iff P(x_1'\beta + u_1 \geq x_0'\beta + u_0 \mid x, c) \geq P(x_0'\beta + u_0 \geq x_1'\beta + u_1 \mid x, c) \\
&\iff P(u_0 - u_1 \leq (x_1 - x_0)'\beta \mid x, c) \geq P(u_1 - u_0 \leq (x_0 - x_1)'\beta \mid x, c) \\
&\iff P(u_0 - u_1 \leq (x_1 - x_0)'\beta \mid x, c) \geq P(u_0 - u_1 \leq (x_0 - x_1)'\beta \mid x, c) \\
&\iff (x_1 - x_0)'\beta \geq 0
\end{aligned}$$

Note that we are able to exchange u_1 and u_0 due to the independence assumption.

Here we consider estimating the set suggested by the rank condition above:

$$\tilde{\Theta}_I = \left\{ \operatorname{sgn}\left(P(y_1 \geq y_0 \mid x, c) - P(y_1 \geq y_0 \mid x, c)\right) = \operatorname{sgn}\left((x_1 - x_0)'\beta\right) \right\}.$$

This set is guaranteed to contain Θ_I and establishing its relative sharpness is left for future research. The intuition underlying this set is that, due to the structure of the model, whenever $x_1'\beta \geq x_0'\beta$ it is likely also the case that $y_1 \geq y_0$.

Consider the following population objective function and sample analog:

$$\begin{aligned} Q(\theta) &= \mathrm{E}\left[\operatorname{sgn}(y_1 - y_0) \cdot \operatorname{sgn}\left((x_1 - x_0)'\beta\right)\right] \\ Q_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(y_1 - y_0) \cdot \operatorname{sgn}\left((x_1 - x_0)'\beta\right) \end{aligned}$$

Due to the close similarity of the objective functions, it follows from the proof of Lemma 1 for Model 1 that Q is maximized exactly on $\tilde{\Theta}_I$.

As before, we can write $Q(\theta) = Pf_\theta$ and $Q_n(\theta) = P_n f_\theta$ where

$$f_\theta(x, y) = 1\{y_1 > y_0\}\, 1\{x_1'\beta \geq x_0'\beta\} - 1\{y_1 < y_0\}\, 1\{x_1'\beta < x_0'\beta\}.$$

It should also be apparent from the arguments underlying Lemma 2 and Lemma 5 that the class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is Euclidean for the constant envelope $F = 1$. Therefore, the conditions of Theorem 3 are satisfied with $b_n = n^{1/2}$.

7.1. Bounding $\Lambda(y)$

In this model, in addition to β, one might be interested in estimating the transformation function Λ. This section discusses estimating bounds for $\Lambda(\bar{y})$ at particular values of $\bar{y}$. Since we can only identify $\Lambda(\bar{y})$ up to differences with respect to $\Lambda(\bar{y}_0)$ at some value $\bar{y}_0$, we normalize $\Lambda(\bar{y}_0) = 0$.

Suppose first that θ_0 is known. Then, again following the maximum score principle, we could estimate bounds for $\Lambda(\bar{y})$ by collecting all values of λ which maximize

$$\frac{1}{n}\sum_{i=1}^{n} \left(1\{y_1 > \bar{y}\} - 1\{y_0 > \bar{y}_0\}\right) 1\{(x_0 - x_1)'\beta_0 \leq \lambda\}.$$

Estimating the set above is infeasible because θ_0 is unknown. However, given an estimated set $\hat{\Theta}_n$, the above method can be applied for each $\theta \in \hat{\Theta}_n$. This suggests using the function

$$\Gamma_n(\bar{y}, \lambda, \theta) = \frac{1}{n}\sum_{i=1}^{n} \left(1\{y_1 > \bar{y}\} - 1\{y_0 > \bar{y}_0\}\right) 1\{(x_0 - x_1)'\beta \leq \lambda\}$$

and forming a set estimate $\hat{\Lambda}_n(\bar{y})$ of $\Lambda(\bar{y})$ which consists of all values of λ which maximize Γ_n for some $\theta \in \hat{\Theta}_n$. That is,

$$\hat{\Lambda}_n(\bar{y}) = \left\{\lambda : \lambda \in \arg\max \Gamma_n(\bar{y}, \lambda, \hat{\theta}) \text{ for some } \hat{\theta} \in \hat{\Theta}_n\right\}.$$

Establishing the asymptotic properties of two-stage estimators such as $\hat{\Lambda}_n(\bar{y})$, which depend on first-stage set estimators such as $\hat{\Theta}_n$, is a promising area for future work which we intend to pursue.

8. Monte Carlo Experiments

In this section we summarize the results of a series of Monte Carlo experiments intended to shed light on the finite sample properties of the proposed estimators defined in Section 5 and the inference procedures defined in Section 6.[5] First, we consider the estimator for Model 1 by replicating the following model:

$$y_{it} = 1\{x_{i1t} + \beta x_{i2t} + c_i + u_{it} \geq 0\}$$

where x_{i1t} and x_{i2t} are uniformly distributed for each t with $x_{i1t} \in \{-2, -1, 0, 1, 2\}$ and $x_{i2t} \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The individual effect is generated as $c_i = (x_{i11} + x_{i12} + x_{i21} + x_{i22})/4$ and the disturbances are iid standard Normal draws. The population parameter used in the experiments is $\theta_0 = \beta_0 = -0.15$ which yields the identified set $\Theta_I = [-0.163, -0.148]$.

Figure 7 displays one realization of $Q_n(\theta)$ for this model, with $n = 500$, along with the population objective function $Q(\theta)$. We compare the estimates for several sample sizes in Table 1, which lists the mean estimated set over 1000 replications for each sample size with $\kappa_n = C\sqrt{\ln n}$ (recall that $\tau_n = \kappa_n/\sqrt{n}$). We present results for $C \in \{0.20, 0.10, 0.05, 0.01\}$. These values were chosen to be roughly around the same magnitude as $\mathcal{Q}_n$. For each sample size, the standard deviation of the endpoints of the estimated sets and the coverage frequency are also reported. By definition of consistency, the coverage probability should asymptotically approach one. Note that only observations for which $y_0 \neq y_1$ are used in estimation. The effective sample size for this specification is about $0.307n$.

As seen in Table 1, smaller constants C used to construct κ_n produce smaller estimated sets, but only at the expense of lower empirical coverage for small sample values of n. One interesting point to note about the estimates in the first panel of Table 1, with $C = 0.20$, is that the upper bound of the estimated interval plateaus at -0.003 for the small sample sizes shown. This corresponds to the large jump in the objective function at $\beta = -0.003$

[5] Fortran 95 source code to reproduce all figures and tables in this section is available from the author's website at `http://jblevins.org/research/panel`.

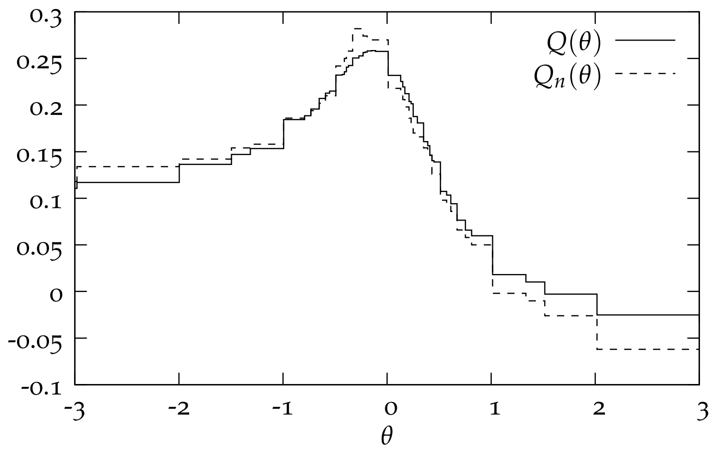

FIGURE 7: $Q(\theta)$ and one realization of $Q_n(\theta)$ for $n = 500$.

that can be seen in Figure 7 and is even larger in the sample analog objective function. Since the sequence $\kappa_n = 0.20\sqrt{\ln n}$ is large relative to the other panels, the cutoff value does not rise above this jump as quickly.

Tables 2, 3, and 4 list, for $m = n^{2/5}$, $m = n^{3/5}$, and $m = n^{4/5}$ respectively, the empirical coverage frequencies of 1000 confidence regions for $1 - \alpha \in \{0.75, 0.90, 0.95, 0.99\}$. For each of the 1000 datasets used for estimation and for each value of $1 - \alpha$, a confidence region was constructed using Algorithm 1 of Section 6. These regions are based on the estimated sets from the same 1000 datasets as before. Increasing the subsample size from $n^{2/5}$ to $n^{3/5}$ seems to increase the speed of convergence of the lower quantiles. The results for the upper quantiles are largely the same for $n^{2/5}$, $n^{3/5}$, and $n^{4/5}$. Note that when the level of τ_n used for estimation is large, the finite sample confidence regions tend to have too little coverage, although it seems that larger subsample sizes are able to mitigate this to some extent.

Finally, in Tables 5 and 6, we present similar estimates and confidence regions with $\kappa_n = 0$ (for $m = n^{3/5}$ only). The estimates obtained with $\kappa_n = 0$ are tight, but have poor coverage in finite samples. The corresponding confidence regions have equally poor coverage

9. Conclusion

We have developed new conditions for establishing both regular and irregular rates of convergence for set estimators in partially identified econometric models and proposed a method for performing inference in models whose estimators exhibit arbitrarily fast convergence. We have applied these general results to a standard binary choice panel data models with fixed effects. First we characterize the sharp identified set and we propose a consistent estimator which converges arbitrarily fast with fully discrete regressors and can achieve rates arbitrarily close to $n^{-1/3}$ when a continuous regressor is present. The validity of a subsampling-based inference procedure is established in the discrete regressor case. We also consider extensions to a lagged dependent variable and panel data duration models. Finally, a series of Monte Carlo experiments illustrates the estimation and inference procedures, which perform as expected.

κ_n	n	Mean $\hat{\Theta}_n$	St. Dev.	Coverage
$0.20\sqrt{\ln n}$	250	[-0.450, 0.077]	[0.149, 0.105]	0.96
	500	[-0.401, 0.044]	[0.096, 0.076]	0.98
	1000	[-0.365, 0.019]	[0.069, 0.051]	0.99
	2000	[-0.328, -0.001]	[0.045, 0.016]	0.99
	4000	[-0.305, -0.003]	[0.037, 0.003]	1.00
	8000	[-0.277, -0.003]	[0.032, 0.005]	1.00
	16000	[-0.256, -0.003]	[0.019, 0.000]	1.00
	32000	[-0.246, -0.003]	[0.010, 0.000]	1.00
	64000	[-0.239, -0.003]	[0.015, 0.000]	1.00
$0.10\sqrt{\ln n}$	250	[-0.318, -0.003]	[0.124, 0.103]	0.75
	500	[-0.301, -0.007]	[0.085, 0.070]	0.85
	1000	[-0.293, -0.005]	[0.064, 0.044]	0.94
	2000	[-0.268, -0.011]	[0.047, 0.034]	0.96
	4000	[-0.252, -0.012]	[0.036, 0.033]	0.99
	8000	[-0.236, -0.011]	[0.026, 0.031]	0.99
	16000	[-0.226, -0.012]	[0.024, 0.032]	0.99
	32000	[-0.216, -0.019]	[0.022, 0.040]	1.00
	64000	[-0.204, -0.026]	[0.015, 0.046]	1.00
$0.05\sqrt{\ln n}$	250	[-0.242, -0.079]	[0.114, 0.121]	0.39
	500	[-0.258, -0.041]	[0.083, 0.085]	0.66
	1000	[-0.247, -0.033]	[0.061, 0.065]	0.76
	2000	[-0.231, -0.044]	[0.048, 0.064]	0.81
	4000	[-0.215, -0.050]	[0.038, 0.065]	0.84
	8000	[-0.203, -0.055]	[0.031, 0.064]	0.88
	16000	[-0.196, -0.064]	[0.027, 0.062]	0.95
	32000	[-0.194, -0.078]	[0.021, 0.060]	0.97
	64000	[-0.188, -0.096]	[0.017, 0.051]	0.99
$0.01\sqrt{\ln n}$	250	[-0.242, -0.079]	[0.114, 0.121]	0.39
	500	[-0.210, -0.096]	[0.080, 0.097]	0.34
	1000	[-0.192, -0.109]	[0.058, 0.083]	0.29
	2000	[-0.178, -0.122]	[0.047, 0.068]	0.31
	4000	[-0.171, -0.126]	[0.039, 0.061]	0.30
	8000	[-0.173, -0.120]	[0.031, 0.056]	0.46
	16000	[-0.166, -0.129]	[0.027, 0.042]	0.50
	32000	[-0.168, -0.137]	[0.021, 0.027]	0.62
	64000	[-0.166, -0.138]	[0.018, 0.017]	0.77

TABLE 1: Fixed effects model estimates. $\theta_0 = -0.150$, $\Theta_I = [-0.163, -0.148]$.

			Empirical Coverage			
κ_n	m	n	0.750	0.900	0.950	0.990
$0.20\sqrt{\ln n}$	$n^{2/5}$	250	0.700	0.941	0.973	0.992
		500	0.672	0.948	0.984	0.993
		1000	0.621	0.960	0.987	0.995
		2000	0.660	0.979	0.992	0.995
		4000	0.824	0.990	0.993	0.997
		8000	0.866	0.991	0.992	0.996
		16000	0.968	0.991	0.992	0.998
		32000	0.994	0.998	0.998	1.000
		64000	0.998	0.998	0.998	1.000
$0.10\sqrt{\ln n}$	$n^{2/5}$	250	0.464	0.662	0.793	0.933
		500	0.438	0.719	0.863	0.967
		1000	0.391	0.819	0.935	0.980
		2000	0.425	0.899	0.972	0.991
		4000	0.508	0.944	0.985	0.991
		8000	0.653	0.978	0.990	0.993
		16000	0.834	0.986	0.992	0.996
		32000	0.926	0.997	0.998	0.999
		64000	0.984	0.998	0.999	1.000
$0.05\sqrt{\ln n}$	$n^{2/5}$	250	0.399	0.464	0.525	0.573
		500	0.384	0.556	0.689	0.857
		1000	0.321	0.594	0.787	0.927
		2000	0.338	0.674	0.857	0.942
		4000	0.362	0.739	0.888	0.958
		8000	0.446	0.811	0.927	0.973
		16000	0.558	0.903	0.970	0.983
		32000	0.728	0.966	0.984	0.988
		64000	0.837	0.978	0.993	0.996
$0.01\sqrt{\ln n}$	$n^{2/5}$	250	0.404	0.465	0.535	0.574
		500	0.347	0.404	0.468	0.530
		1000	0.295	0.336	0.398	0.481
		2000	0.311	0.344	0.382	0.446
		4000	0.308	0.342	0.366	0.414
		8000	0.355	0.461	0.541	0.636
		16000	0.420	0.494	0.550	0.615
		32000	0.515	0.604	0.661	0.688
		64000	0.640	0.764	0.797	0.808

TABLE 2: Fixed effects model confidence regions, $m = n^{2/5}$.

κ_n	m	n	Empirical Coverage 0.750	0.900	0.950	0.990
		250	0.943	0.979	0.987	0.992
		500	0.934	0.978	0.994	0.993
		1000	0.910	0.976	0.984	0.996
		2000	0.936	0.989	0.991	0.997
$0.20\sqrt{\ln n}$	$n^{3/5}$	4000	0.978	0.989	0.990	0.995
		8000	0.985	0.991	0.994	0.995
		16000	0.986	0.994	0.997	0.997
		32000	0.997	1.000	1.000	1.000
		64000	1.000	1.000	1.000	1.000
		250	0.709	0.884	0.949	0.960
		500	0.776	0.923	0.952	0.970
		1000	0.838	0.931	0.965	0.983
		2000	0.903	0.961	0.978	0.990
$0.10\sqrt{\ln n}$	$n^{3/5}$	4000	0.922	0.972	0.982	0.987
		8000	0.945	0.979	0.993	0.992
		16000	0.965	0.989	0.995	0.994
		32000	0.984	0.996	0.996	1.000
		64000	0.994	0.999	1.000	1.000
		250	0.484	0.574	0.574	0.573
		500	0.608	0.775	0.877	0.918
		1000	0.721	0.853	0.897	0.936
		2000	0.808	0.912	0.931	0.946
$0.05\sqrt{\ln n}$	$n^{3/5}$	4000	0.859	0.926	0.939	0.957
		8000	0.900	0.918	0.926	0.945
		16000	0.898	0.954	0.965	0.974
		32000	0.939	0.975	0.980	0.982
		64000	0.969	0.983	0.992	0.995
		250	0.480	0.572	0.574	0.574
		500	0.434	0.519	0.535	0.538
		1000	0.378	0.468	0.497	0.498
		2000	0.387	0.421	0.453	0.465
$0.01\sqrt{\ln n}$	$n^{3/5}$	4000	0.367	0.412	0.432	0.447
		8000	0.554	0.629	0.649	0.667
		16000	0.583	0.609	0.612	0.617
		32000	0.665	0.671	0.676	0.688
		64000	0.798	0.800	0.804	0.807

Table 3: Fixed effects model confidence regions, $m = n^{3/5}$.

			Empirical Coverage			
κ_n	m	n	0.750	0.900	0.950	0.990
$0.20\sqrt{\ln n}$	$n^{4/5}$	250	0.952	0.977	0.989	0.990
		500	0.961	0.987	0.992	0.997
		1000	0.966	0.989	0.996	0.997
		2000	0.979	0.994	0.995	0.999
		4000	0.989	0.996	0.999	1.000
		8000	0.987	0.997	0.999	1.000
		16000	0.991	0.996	0.998	1.000
		32000	0.997	1.000	1.000	0.999
		64000	1.000	1.000	1.000	1.000
$0.10\sqrt{\ln n}$	$n^{4/5}$	250	0.807	0.876	0.893	0.904
		500	0.829	0.918	0.946	0.951
		1000	0.888	0.958	0.970	0.982
		2000	0.949	0.981	0.985	0.988
		4000	0.960	0.991	0.995	0.990
		8000	0.968	0.992	0.995	0.996
		16000	0.977	0.994	0.998	0.995
		32000	0.993	0.997	0.997	1.000
		64000	0.998	1.000	1.000	1.000
$0.05\sqrt{\ln n}$	$n^{4/5}$	250	0.544	0.551	0.556	0.556
		500	0.743	0.822	0.838	0.856
		1000	0.775	0.861	0.892	0.920
		2000	0.868	0.917	0.931	0.941
		4000	0.890	0.905	0.916	0.934
		8000	0.889	0.918	0.934	0.955
		16000	0.939	0.961	0.965	0.968
		32000	0.965	0.980	0.979	0.982
		64000	0.974	0.991	0.995	0.995
$0.01\sqrt{\ln n}$	$n^{4/5}$	250	0.545	0.553	0.555	0.555
		500	0.480	0.499	0.504	0.505
		1000	0.437	0.458	0.464	0.464
		2000	0.429	0.436	0.444	0.446
		4000	0.396	0.410	0.413	0.423
		8000	0.575	0.606	0.616	0.634
		16000	0.589	0.601	0.606	0.616
		32000	0.666	0.679	0.679	0.685
		64000	0.798	0.801	0.806	0.806

TABLE 4: Fixed effects model confidence regions, $m = n^{4/5}$.

n	Mean $\hat{\Theta}_n$	St. Dev.	Coverage
125	[-0.281, -0.019]	[0.181, 0.156]	0.53
250	[-0.242, -0.079]	[0.114, 0.121]	0.39
500	[-0.210, -0.096]	[0.080, 0.097]	0.34
1000	[-0.192, -0.109]	[0.058, 0.083]	0.29
2000	[-0.178, -0.122]	[0.047, 0.068]	0.31
4000	[-0.171, -0.126]	[0.039, 0.061]	0.30
8000	[-0.166, -0.132]	[0.031, 0.047]	0.35
16000	[-0.162, -0.135]	[0.027, 0.037]	0.41
32000	[-0.163, -0.143]	[0.022, 0.020]	0.50
64000	[-0.161, -0.143]	[0.017, 0.014]	0.61

TABLE 5: Fixed effects model estimates ($\kappa_n = 0$, $m = n^{3/5}$).

	Empirical Coverage			
n	0.750	0.900	0.950	0.990
125	0.603	0.651	0.652	0.651
250	0.480	0.571	0.574	0.573
500	0.429	0.517	0.534	0.536
1000	0.377	0.461	0.495	0.501
2000	0.381	0.424	0.458	0.465
4000	0.372	0.416	0.430	0.447
8000	0.399	0.421	0.433	0.440
16000	0.442	0.457	0.459	0.461
32000	0.514	0.517	0.518	0.521
64000	0.622	0.622	0.622	0.622

TABLE 6: Fixed effects model confidence regions ($\kappa_n = 0$, $m = n^{3/5}$).

A. Notation and Preliminary Results

A.1. Notation

First we introduce some notation. We shall make use of a modified signum function $\text{sgn}(x)$ where

$$\text{sgn}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0. \end{cases}$$

This definition, which is standard in the maximum score literature, differs from the common definition only at zero, where we define $\text{sgn}(0) = 1$ instead of $\text{sgn}(0) = 0$. We write $a \vee b$ to denote $\max\{a, b\}$ and $a \wedge b$ to denote $\min\{a, b\}$.

In a slight abuse of notation, define the distance between a point x and a set B to be

$$d(x, B) = \inf_{x' \in B} d(x, x'),$$

where d denotes the Euclidean distance. For any set B, we let B^ε denote an ε-expansion of B, defined as

$$B^\varepsilon = \{x \in B : d(x, B) \leq \varepsilon\}.$$

We write $a_n \downarrow a$ to a sequence which eventually equals a, or in other words, a sequence for which there exists an $N < \infty$ such that $a_n = a$ for all $n \geq N$. We also say that such a sequence converges *arbitrarily fast* to a since for any sequence r_n, $r_n |a_n - a| \to 0$. This includes all polynomials of n such as $r_n = n^{1/2}$. In particular, when a_n is a stochastic process, we say a_n converges arbitrarily fast to a in probability, or a_n is eventually a in probability, when $P\{\omega \in \Omega : a_n(\omega) = a\} \to 1$. In such cases we write $a_n \downarrow a$ in probability.

A.2. Preliminary Results

Lemma 8. *Let f and g be bounded real functions on $A \subset \mathbb{R}^n$. Then*

$$\left| \sup_{x \in A} f(x) - \sup_{x \in A} g(x) \right| \leq \sup_{x \in A} |f(x) - g(x)|.$$

Proof of Lemma 8. First, note that for all $x \in A$,

$$f(x) - \sup_{y \in A} g(y) \leq f(x) - g(x) \leq |f(x) - g(x)| \tag{10}$$

and

$$\sup_{y \in A} f(y) - g(x) \geq f(x) - g(x) \geq -|f(x) - g(x)|. \tag{11}$$

We prove the result by showing that

$$-\sup_{x\in A}|f(x)-g(x)| \leq \sup_{x\in A} f(x) - \sup_{x\in A} g(x) \leq \sup_{x\in A}|f(x)-g(x)|.$$

For the right hand side:

$$\sup_{x\in A} f(x) - \sup_{x\in A} g(x) = \sup_{x\in A}\left[f(x) - \sup_{y\in A} g(y)\right] \leq \sup_{x\in A}|f(x)-g(x)|.$$

The equality holds since $\sup g$ is constant with respect to x and the inequality follows from (10), since it holds for all x. Similarly, the left hand side follows from (11):

$$\begin{aligned}\sup_{x\in A} f(x) - \sup_{x\in A} g(x) &= \sup_{x\in A} f(x) + \inf_{x\in A}(-g(x)) \\ &= \inf_{x\in A}\left[\sup_{y\in A} f(y) - g(x)\right] \\ &\geq \inf_{x\in A} -|f(x)-g(x)| \\ &= -\sup_{x\in A}|f(x)-g(x)|\end{aligned}$$

Together, these two inequalities imply the result. ■

B. Consistent Estimation

Proof of Theorem 3. The proof proceeds in two steps. First, we show that $\sup_{\theta\in\hat{\Theta}_n} d(\theta,\Theta_I) \xrightarrow{p} 0$. The second step shows that $\lim_{n\to\infty} P(\Theta_I \subset \hat{\Theta}_n) = 1$. Combining these steps and using the definition of the Hausdorff distance yields the final conclusion of the theorem. Let B^ε denote an ε-expansion of a set B, as defined in Subsubsection A.1.

Step 1 For any $\varepsilon > 0$,

$$\sup_{\Theta\backslash\Theta_I^\varepsilon} Q_n \leq \sup_{\Theta\backslash\Theta_I^\varepsilon} Q + o_p(1) \leq \sup_{\Theta} Q - \delta_\varepsilon + o_p(1),$$

where $\delta_\varepsilon > 0$. The first inequality above follows from Assumption 4.d, giving uniform convergence in probability of Q_n to Q, and the second inequality follows from Assumption 4.c, since Θ_I maximizes Q. Similarly,

$$\inf_{\hat{\Theta}_n} Q_n \geq \sup_{\Theta} Q_n - \tau_n \geq \sup_{\Theta} Q - \tau_n + o_p(1)$$

The first inequality follows from the definition of $\hat{\Theta}_n$ and the second follows again from uniform convergence. By assumption, $\tau_n = o_p(1)$, and since $\delta_\varepsilon > 0$, with probability approaching one, $\tau_n < \delta_\varepsilon$, or equivalently, $\sup_\Theta Q - \tau_n + o_p(1) \geq \sup_\Theta Q - \delta_\varepsilon + o_p(1)$. Given the inequalities above, this implies $\inf_{\hat{\Theta}_n} Q_n \geq \sup_{\Theta\backslash\Theta_I^\varepsilon} Q_n$, which in turn implies that $\Theta_n \subseteq \Theta_I^\varepsilon$, and so $\sup_{\theta\in\hat{\Theta}_n} d(\theta,\Theta_I) \leq \varepsilon$.

Step 2 By definition of $\hat{\Theta}_n$ and τ_n, we know that if $b_n\tau_n \geq \sup_\Theta b_n Q_n - \inf_{\Theta_I} b_n Q_n$, then $\Theta_I \subseteq \hat{\Theta}_n$. We have

$$\begin{aligned}
\sup_\Theta Q_n - \inf_{\Theta_I} Q_n &= \left[\sup_\Theta Q_n - \sup_\Theta Q\right] + \left[\sup_\Theta Q - \inf_{\Theta_I} Q_n\right] \\
&\leq \left|\sup_\Theta Q_n - \sup_\Theta Q\right| + \left|\sup_\Theta Q - \inf_{\Theta_I} Q_n\right| \\
&= \left|\sup_\Theta Q_n - \sup_\Theta Q\right| + \left|\sup_{\Theta_I} Q - \inf_{\Theta_I} Q_n\right| \\
&\leq \sup_\Theta |Q_n - Q| + \sup_{\Theta_I} |Q_n - Q| \\
&\leq \sup_\Theta |Q_n - Q| + \sup_\Theta |Q_n - Q|
\end{aligned}$$

These steps follow by, respectively, adding and subtracting $\sup_\Theta Q$, taking the absolute value, noting that Θ_I maximizes Q, using the fact that $\inf f = -\sup -f$, and applying Lemma 8 (see Appendix A) twice, noting that $\Theta_I \subseteq \Theta$. By Assumption 4.d, $\sup_\Theta |Q_n - Q| = O_p(1/b_n)$ and so the requirement that $\tau_n b_n \xrightarrow{p} \infty$ (i.e., that τ_n approaches zero in probability slower than $1/b_n$) implies that $\tau_n \geq 2\sup_\Theta |Q_n - Q| \geq \sup_\Theta Q_n - \inf_{\Theta_I} Q_n$ with probability approaching one. ∎

Proof of Theorem 4. From Theorem 3, $\lim_{n\to\infty} P(\Theta_I \subseteq \hat{\Theta}_n) = 1$. We will prove the result by showing that $\lim_{n\to\infty} P(\hat{\Theta}_n \subseteq \Theta_I) = 1$ and therefore the Hausdorff distance $d_H(\hat{\Theta}_n, \Theta_I)$ eventually equals zero with probability approaching one.

Uniform convergence at the $1/b_n$ rate (Assumption 4.d) implies $Q_n(\theta) \leq Q(\theta) + O_p(1/b_n)$ and $Q(\theta) \leq Q_n(\theta) + O_p(1/b_n)$. It follows that $\sup_{\Theta\setminus\Theta_I} Q_n \leq \sup_{\Theta\setminus\Theta_I} Q + O_p(1/b_n) \leq \sup_\Theta Q - \delta + O_p(1/b_n) \leq \sup_\Theta Q_n - \delta + O_p(1/b_n)$, where the second inequality follows from Assumption 5.

Since τ_n converges to zero in probability and $\delta > 0$ is constant, with probability approaching one, $\tau_n < \delta$. Thus, with probability approaching one, $-\delta < -\tau_n$, $\sup_{\Theta\setminus\Theta_I} Q_n \leq \sup_\Theta Q_n - \tau_n + O_p(1/b_n) \leq \inf_{\hat{\Theta}_n} Q_n + O_p(1/b_n)$, and therefore, $\hat{\Theta}_n \subseteq \Theta_I$. ∎

Proof of Theorem 5. For any $\varepsilon > 0$, let δ, κ, γ_1, γ_2, κ_ε, and n_ε satisfy Assumption 6 and define

$$\nu_n \equiv \left(\frac{\kappa \cdot \kappa_\varepsilon \vee b_n \cdot \tau_n}{b_n \cdot \kappa}\right)^{\frac{1}{\gamma_2}}$$

where b_n is given by Assumption 4.d. Then, since $\nu_n = o_p(1)$, $\nu_n = O_p(\tau_n^{1/\gamma_2})$, and $\tau_n b_n \xrightarrow{p} \infty$, there is an $n'_\varepsilon > n_\varepsilon$ such that for all $n > n'_\varepsilon$, with probability at least $1 - \varepsilon$, we have both $\nu_n \le \delta$ and $\nu_n \ge (\kappa_\varepsilon / b_n)^{1/\gamma_2}$. On a set $\Theta \setminus \Theta_I^{\nu_n}$, $d(\theta, \Theta_I) \ge \nu_n$, so $\min\{d(\theta,\Theta_I),\delta\} \ge \min\{\nu_n, \delta\}$. Therefore, by Assumption 6,

$$\begin{aligned}
\sup_{\Theta \setminus \Theta_I^{\nu_n}} Q_n &\le \sup_{\Theta} Q_n - \kappa \cdot (\nu_n \wedge \delta)^{\gamma_1} \\
&\le \sup_{\Theta} Q_n - \kappa \cdot \nu_n^{\gamma_1} \\
&\le \sup_{\Theta} Q_n - \left(\frac{\kappa \cdot \kappa_\varepsilon}{b_n} \vee \tau_n \right)^{\gamma_1/\gamma_2} \\
&\le \sup_{\Theta} Q_n - \tau_n^{\gamma_1/\gamma_2} \\
&\le \inf_{\hat{\Theta}_n} Q_n .
\end{aligned}$$

The above implies that $\hat{\Theta}_n \cap (\Theta \setminus \Theta_I^{\nu_n})$ is empty, or equivalently, that $\hat{\Theta}_n \subseteq \Theta_I^{\nu_n}$. Therefore, in light of Step 1 of the proof of Theorem 3, which shows that $\lim_{n\to\infty} P(\Theta_I \subseteq \hat{\Theta}_n) = 1$, we have $d_H(\hat{\Theta}_n, \Theta_I) = O_p(\tau_n^{1/\gamma_2})$ (since τ_n is slower than $1/b_n$ by assumption). ■

C. Confidence Regions

Proof of Lemma 6. Observe that

$$P\{\Theta_I \subseteq C_n(\hat{c}_n)\} = P\{\mathcal{Q}_n \le \hat{c}_n\} = P\{\mathcal{Q} \le c(1-\alpha)\} + o_p(1) \ge (1-\alpha) + o_p(1).$$

The first equality holds by definition of C_n and $\mathcal{Q}_n$, the second by Assumption 7 and $\hat{c}_n \xrightarrow{p} c(1-\alpha)$, and the third by definition of $c(1-\alpha)$. ■

Proof of Theorem 11. The proof proceeds in three steps. First, we derive upper and lower bounds for $\hat{\mathcal{Q}}_{n,m,j}$ such that $\underline{\mathcal{Q}}_{n,m,j} \le \hat{\mathcal{Q}}_{n,m,j} \le \overline{\mathcal{Q}}_{n,m,j}$ with probability approaching one. Next, we prove that the empirical distribution function of $\hat{\mathcal{Q}}_{n,m,j}$ converges in probability to the distribution function of $\mathcal{Q}$, the limiting distribution of $\mathcal{Q}_n$. Finally, we show that $\hat{c}_n$ converges in probability to $c(1-\alpha)$, the desired quantile of the distribution of $\mathcal{Q}$.

Step 1 By Theorem 4, we have $d_H(C_n(\kappa_n), \Theta_I) = 0$ with probability approaching one. Thus, $d_H(C_n(\kappa_n), \Theta_I) \le \varepsilon_n$ for some sequence $\varepsilon_n \downarrow 0$ with probability approaching one. For a fixed subsample $1 \le j \le M_n$, let $\underline{\mathcal{Q}}_{n,m,j} \equiv \sup_{\theta\in\Theta} b_m Q_{n,m,j}(\theta) - \inf_{\theta \in \Theta_I^{\varepsilon_n}} b_m Q_{n,m,j}(\theta)$. Let $\mathcal{K}_n$ be the collection of all subsets $K \subseteq \Theta$ such that $d_H(K, \Theta_I) \le \varepsilon_n$ and define $\overline{\mathcal{Q}}_{n,m,j} \equiv \sup_{K \in \mathcal{K}_n} \left[\sup_{\theta\in\Theta} b_m Q_{n,m,j}(\theta) - \inf_{\theta\in K} b_m Q_{n,m,j}(\theta)\right]$. There exists a set $\Theta_{n,m,j} \in \mathcal{K}_n$ such that $\mathcal{Q}_{n,m,j}$ is equal to $\inf_{\theta \in \Theta_{n,m,j}} b_m Q_{n,m,j}(\theta)$. With probability approaching one, since $C_n(\kappa_n) \subseteq \Theta_I^{\varepsilon_n}$ and $C_n(\kappa_n) \in \mathcal{K}_n$, we have $\underline{\mathcal{Q}}_{n,m,j} \le \hat{\mathcal{Q}}_{n,m,j} \le \overline{\mathcal{Q}}_{n,m,j}$ for all $j = 1, \ldots, M_n$.

Step 2 From Step 1, with probability approaching one,

$$\underline{G}_{n,m}(x) \equiv M_n^{-1}\sum_{j=1}^{M_n} 1\{\overline{\mathcal{Q}}_{n,m,j} \le x\} \le \hat{G}_{n,m}(x) \equiv M_n^{-1}\sum_{j=1}^{M_n} 1\{\hat{\mathcal{Q}}_{n,m,j} \le x\}$$
$$\le \overline{G}_{n,m}(x) \equiv M_n^{-1}\sum_{j=1}^{M_n} 1\{\overline{\mathcal{Q}}_{n,m,j} \le x\}.$$

We will show that $\underline{G}_{n,m}(x) \xrightarrow{\text{p}} P\{\mathcal{Q} \le x\}$ and $\overline{G}_{n,m}(x) \xrightarrow{\text{p}} P\{\mathcal{Q} \le x\}$ as $n \to \infty$ (and thus, $m \to \infty$). Therefore, $\hat{G}_{n,m}(x) \xrightarrow{\text{p}} P\{\mathcal{Q} \le x\}$ for each $x \in \mathbb{R}$.

Let $\overline{J}_m(x)$ denote the cdf of $\overline{\mathcal{Q}}_{n,m,j}$. Note that $\underline{G}_{n,m}(x)$ is a U-statistic of degree m with $0 \le \underline{G}_{n,m}(x) \le 1$ (i.e., it is bounded). Furthermore, $\mathrm{E}[\underline{G}_{n,m}(x)] = \mathrm{E}[1\{\overline{\mathcal{Q}}_{n,m,j} \le x\}] = \overline{J}_m(x)$, where the last equality holds by nonreplacement sampling, since each subsample of size m is itself an iid sample. By the Hoeffding inequality for bounded U-statistics for iid data (Serfling, 1980, Theorem A, p. 201), for any $t > 0$,

$$P\left\{\underline{G}_{n,m}(x) - \overline{J}_m(x) \ge t\right\} \le \exp\left[-2t^2\frac{n}{m}\right].$$

A similar inequality follows for $t < 0$ by considering the U-process $-\underline{G}_{n,m}(x)$. Therefore, $\underline{G}_{n,m}(x) = \overline{J}_m(x) + o_p(1)$ for fixed m. Finally, since $\overline{\mathcal{Q}}_{n,m,j}$ is obtained from sets satisfying Assumption 8, $\overline{J}_m(x) = P\{\overline{\mathcal{Q}}_{n,m,j} \le x\} = P\{\overline{\mathcal{Q}} \le x\} + o_p(1)$.

A similar argument shows that $\overline{G}_{n,m}(x) \xrightarrow{\text{p}} P\{\mathcal{Q} \le x\}$ as well, and therefore, $\hat{G}_{n,m}(x) \xrightarrow{\text{p}} P\{\mathcal{Q} \le x\}$.

Step 3 Convergence of the distribution function at continuity points implies convergence of the quantile function at continuity points (cf. Shorack, 2000, Proposition 3.1). Therefore, $\hat{c}_n = \inf\{x : \hat{G}(x) \ge 1-\alpha\} \xrightarrow{\text{p}} c(1-\alpha)$. ■

D. Fixed Effects Model

Below we provide proofs for results pertaining to Model 1. First we present results which are independent of assumptions on state space $\mathcal{X}$, followed by results for discrete regressors and continuous regressors.

D.1. General Results

Proof of Theorem 1. For the proof, let Θ_I denote the identified set as defined in (4) and let $\tilde{\Theta}_I$ denote the set on the right side of (5). We first show $\Theta_I \subseteq \tilde{\Theta}_I$, and then $\tilde{\Theta}_I \subseteq \Theta_I$.

Step 1 Let $\theta \in \Theta_I$. By definition of Θ_I, there exist distributions $F_{u_0|xc}$ and $F_{c|x}$ such that $\pi(y_t = 1 \mid x; \beta, F_{u_0|xc}, F_{c|x}) = P(y_t = 1 \mid x)$ F_x-almost surely for $t = 0, 1$. Conditioning on c, we have $P(y_0 = 1 \mid x, c) = 1 - F_{u_0|xc}(-x_0'\beta - c)$ and $P(y_1 = 1 \mid x, c) = 1 - F_{u_0|xc}(-x_1'\beta - c)$. By the monotonicity of $F_{u_0|xc}$,

$$\begin{aligned} P(y_1 = 1 \mid x, c) \geq P(y_0 = 1 \mid x, c) &\iff 1 - F_{u_0|xc}(-x_1'\beta - c) \geq 1 - F_{u_0|xc}(-x_0'\beta - c) \\ &\iff F_{u_0|xc}(-x_1'\beta - c) \leq F_{u_0|xc}(-x_0'\beta - c) \\ &\iff -x_1'\beta - c \leq -x_0'\beta - c \\ &\iff (x_1 - x_0)'\beta \geq 0 \end{aligned}$$

Since this event is independent of c, we have

$$P(y_1 = 1 \mid x) - P(y_0 = 1 \mid x) \geq 0 \iff (x_1 - x_0)'\beta \geq 0,$$

or, equivalently,

$$\operatorname{sgn}\left(P(y_1 = 1 \mid x) - P(y_0 = 1 \mid x)\right) = \operatorname{sgn}\left((x_1 - x_0)'\beta\right).$$

Therefore, $\theta \in \Theta_I \Rightarrow \theta \in \tilde{\Theta}_I$.

Step 2 Now, suppose $\theta \in \tilde{\Theta}_I$. We will show that for each such θ, given population distributions $P(y_t \mid x)$ for $t = 0, 1$, there are values of the remaining free model primitives—the cdfs $F_{u_0|xc}$ and $F_{c|x}$—such that the implications of the model coincide with the true population values $P(y_0 = 0 \mid x)$ and $P(y_1 = 0 \mid x)$.

First, note that we do not need to consider the events $y_0 = 1$ or $y_1 = 1$ since in each time period, the (binary) choice probabilities must sum to one. Thus, we need to show that there exist distributions $F_{u_0|xc}$ and $F_{c|x}$ such that for F_x-almost every x the model implications align with the population choice probabilities:

$$\begin{aligned} P(y_0 = 0 \mid x) &= \pi(y_0 = 0 \mid x; \theta, F_{u_0|xc}, F_{c|x}) \\ P(y_1 = 0 \mid x) &= \pi(y_1 = 0 \mid x; \theta, F_{u_0|xc}, F_{c|x}) \end{aligned}$$

For a given x and for primitives $(\theta, F_{u_0|xc}, F_{c|x})$, the model implications are:

$$\begin{aligned} \pi(y_0 = 0 \mid x; \theta, F_{u_0|xc}, F_{c|x}) &= \int F_{u_0|xc}(-x_0'\beta - c)\, dF_{c|x} \\ \pi(y_1 = 0 \mid x; \theta, F_{u_0|xc}, F_{c|x}) &= \int F_{u_0|xc}(-x_1'\beta - c)\, dF_{c|x} \end{aligned}$$

Fix x. It will suffice to construct a distribution $F_{c|x}$ with only a single mass point $c^*(x)$ (conditional on each fixed value of x):

$$F_{c|x}(c) = \begin{cases} 0 & \text{if } c < c^*(x), \\ 1 & \text{if } c \geq c^*(x). \end{cases}$$

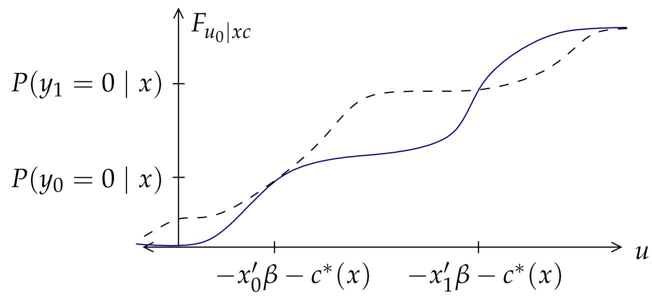

Figure 8: Two distributions $F_{u_0|xc}$ with equivalent observable implications under $F_{c|x}$.

Suppose that $P(y_1 = 1 \mid x) < P(y_0 = 1 \mid x)$ (the opposite case follows similarly). Then our choice of $\theta \in \tilde{\Theta}_I$ guarantees that β is such that $x_1'\beta < x_0'\beta$. We can rewrite these two inequalities equivalently as $P(y_0 = 0 \mid x) < P(y_1 = 0 \mid x)$ and $-x_0'\beta < -x_1'\beta$. Thus, the following choice for $F_{u_0|xc}$ is a valid cdf:

$$F_{u_0|xc}(u) = \begin{cases} 0 & \text{if } u < -x_1'\beta - c^*(x), \\ P(y_0 = 0 \mid x) & \text{if } -x_0'\beta - c^*(x) \leq u \leq -x_1'\beta - c^*(x), \\ P(y_1 = 0 \mid x) & \text{if } -x_1'\beta - c^*(x) \leq u < \bar{u}, \\ 1 & \text{if } u \geq \bar{u}, \end{cases}$$

for any $\bar{u} > -x_1'\beta - c^*(x)$. Essentially, we only need to choose a cdf that passes through the two points $(-x_0'\beta - c^*(x), P(y_0 = 0 \mid x))$ and $(-x_1'\beta - c^*(x), P(y_1 = 0 \mid x))$ and there are an infinite number of such cdfs, as illustrated by Figure 8.

Given the above cdfs, we have:

$$\pi(y_0 = 0 \mid x; \theta, F_{u_0|xc}, F_{c|x}) = F_{u_0|xc}(-x_0'\beta - c^*(x)) = P(y_0 = 0 \mid x),$$
$$\pi(y_1 = 0 \mid x; \theta, F_{u_0|xc}, F_{c|x}) = F_{u_0|xc}(-x_1'\beta - c^*(x)) = P(y_1 = 0 \mid x).$$

Therefore $\theta \in \Theta_I$, and since $\theta \in \tilde{\Theta}_I$ was chosen arbitrarily, $\tilde{\Theta}_I \subseteq \Theta_I$. ■

Proof of Lemma 1. For simplicity, define $w = x_1 - x_0$, $z = y_1 - y_0$, and $\Theta^* = \arg\max_{\theta \in \Theta} Q(\theta)$.

Step 1 Let $\theta_1 \in \Theta_I$ and $\theta_2 \in \Theta$. We will show that $\Theta_I \subseteq \Theta^*$ by proving that, for arbitrary choices of θ_1 and θ_2, $Q(\theta_1) \geq Q(\theta_2)$.

Consider the difference

$$\begin{aligned} Q(\theta_1) - Q(\theta_2) &= \mathrm{E}\left[z \operatorname{sgn}(w'\beta_1)\right] - \mathrm{E}\left[z \operatorname{sgn}(w'\beta_2)\right] \\ &= \mathrm{E}\left[z \left(\operatorname{sgn}(w'\beta_1) - \operatorname{sgn}(w'\beta_2)\right)\right] \\ &= 2 \int_{D(\theta_1, \theta_2)} \operatorname{sgn}(w'\beta_1)\, \mathrm{E}\left[z \mid x, c\right]\, dF_{xc} \end{aligned}$$

where $D(\theta_1, \theta_2) = \{(x, c) : \operatorname{sgn}(w'\beta_1) \neq \operatorname{sgn}(w'\beta_2)\}$ is the set of values of x and c where $\operatorname{sgn}(w'\beta_1)$ and $\operatorname{sgn}(w'\beta_2)$ differ. The last equality above follows from the fact that the integrand vanishes on complement of $D(\theta_1, \theta_2)$, and that on $D(\theta_1, \theta_2)$, $\operatorname{sgn}(w'\beta_1) = -\operatorname{sgn}(w'\beta_2)$, implying that $\operatorname{sgn}(w'\beta_1) - \operatorname{sgn}(w'\beta_2) = 2\operatorname{sgn}(w'\beta_1)$. Since $\theta_1 \in \Theta_I$, Theorem 1 guarantees that

$$\operatorname{sgn}(w'\beta_1) = \operatorname{sgn}\left(P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c)\right) = \operatorname{sgn} \mathrm{E}(z \mid x, c)$$

F_{xc}-almost surely. Rewriting the above difference,

$$Q(\theta_1) - Q(\theta_2) = 2 \int_{D(\theta_1, \theta_2)} |\mathrm{E}[z \mid x, c]| \, dF_{xc} \geq 0$$

for all θ_2. Therefore, $\Theta_I \subseteq \Theta^*$.

Step 2 Now, let $\theta_1 \in \Theta_I$ and suppose there exists a $\theta_2 \in \Theta_I^c \cap \Theta^*$, where Θ_I^c denotes the complement of Θ_I. We will use the definition of Θ_I to show that $Q(\theta_2) < Q(\theta_1)$, contradicting the assumption that $\theta_2 \in \Theta^*$, and guaranteeing that $\Theta_I^c \cap \Theta^* = \emptyset$, or equivalently, $\Theta^* \subseteq \Theta_I$.

First, note that we can rewrite $Q(\theta)$ as follows:

$$\begin{aligned}
Q(\theta) &= \mathrm{E}[z \operatorname{sgn}(w'\beta)] \\
&= \mathrm{E}_{xc} \, \mathrm{E}_{z|wc}[z \operatorname{sgn}(w'\beta)] \\
&= \mathrm{E}_{xc} \left[\left(P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c)\right) \left(1\{w'\beta \geq 0\} - 1\{w'\beta < 0\}\right) \right] \\
&= \int_{\{w'\beta \geq 0\}} \left(P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c)\right) dF_{xc} \\
&\quad + \int_{\{w'\beta < 0\}} \left(P(y_0 = 1 \mid x, c) - P(y_1 = 1 \mid x, c)\right) dF_{xc}
\end{aligned}$$

The first equality is definitional, the second is an application of the law of iterated expectations, and the third follows from the definition of z and the signum function. In the fourth line, the expectations of the indicator functions are expressed as integrals over the corresponding regions of the support of x.

Now, consider the difference $Q(\theta_2) - Q(\theta_1)$:

$$\begin{aligned}
Q(\theta_2) - Q(\theta_1) = &\int_{\{w'\beta_2 \geq 0\}} \left(P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c)\right) dF_{xc} \\
&+ \int_{\{w'\beta_2 < 0\}} \left(P(y_0 = 1 \mid x, c) - P(y_1 = 1 \mid x, c)\right) dF_{xc} \\
&- \int_{\{w'\beta_1 \geq 0\}} \left(P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c)\right) dF_{xc} \\
&- \int_{\{w'\beta_1 < 0\}} \left(P(y_0 = 1 \mid x, c) - P(y_1 = 1 \mid x, c)\right) dF_{xc}
\end{aligned}$$

Over regions where $w'\beta_2$ and $w'\beta_1$ have the same sign, the difference is zero, therefore

$$Q(\theta_2) - Q(\theta_1) = \int_{\{w'\beta_2 \geq 0, w'\beta_1 < 0\}} (P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c)) \, dF_{xc} - \int_{\{w'\beta_2 < 0, w'\beta_1 \geq 0\}} (P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c)) \, dF_{xc}$$

From the proof of Theorem 1, we know that for $\theta_1 \in \Theta_I$,

$$P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c) \geq 0 \iff w'\beta_1 \geq 0$$

and for $\theta_2 \in \Theta_I^c$,

$$P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c) < 0 \iff w'\beta_2 \geq 0.$$

This implies that the first term in the difference above is strictly negative and the second term, which is being subtracted, is weakly non-negative. Thus, $Q(\theta_2) < Q(\theta_1)$. This contradicts the choice of θ_2, meaning that $\Theta_I^c \cap \Theta^* = \emptyset$ and therefore it must be the case that $\Theta^* \subseteq \Theta_I$. ■

Proof of Lemma 2. Let $\mathcal{D} \subset \mathbb{R}^d$ denote the support of w and let $\mathcal{X} = \{-1, 0, 1\} \times \mathcal{D}$ denote the support of (z, w). For each $(z, w) \in \mathcal{X}$ and for each real number t, α, and γ, and real vector $\delta \in \mathbb{R}^d$, define

$$g(z, w, t, \alpha, \gamma, \delta) = \alpha t + \gamma z + \delta' w$$

and define

$$\mathcal{G} = \left\{ g(\cdot, \cdot, \cdot, \alpha, \gamma, \delta) : \alpha, \gamma \in \mathbb{R} \text{ and } \delta \in \mathbb{R}^d \right\}.$$

Since $\mathcal{G}$ is a vector space of real-valued functions on $\mathcal{X} \times \mathbb{R}$, by Lemma 2.4 of Pakes and Pollard (1989), classes of sets of the form $\{g \geq r\}$ or $\{g > r\}$ with $g \in \mathcal{G}$ and $r \in \mathbb{R}$ are VC classes. We will show that $\mathcal{F}$ is Euclidean by showing that it is a VC subgraph class, that is, that the collection of subgraphs of functions in $\mathcal{F}$ is a VC class. To accomplish this, we will use Lemma 2.5 of Pakes and Pollard (1989) which states that, in particular, if $\mathcal{C}_1$ and $\mathcal{C}_2$ are VC classes, then so are $\{C_1 \cap C_2 : C_1 \in \mathcal{C}_1, C_2 \in \mathcal{C}_2\}$, $\{C_1 \cup C_2 : C_1 \in \mathcal{C}_1, C_2 \in \mathcal{C}_2\}$, and $\{C_1^c : C_1 \in \mathcal{C}_1\}$.

First, note that we can rewrite f as

$$\begin{aligned} f(z, w, \theta) &= (1\{z > 0\} - 1\{z < 0\}) \cdot (1\{w'\beta \geq 0\} - 1\{w'\beta < 0\}) \\ &= 1\{z > 0,\, w'\beta \geq 0\} - 1\{z > 0,\, w'\beta < 0\} \\ &\quad - 1\{z < 0,\, w'\beta \geq 0\} + 1\{z < 0,\, w'\beta < 0\}. \end{aligned}$$

Now, for any $\theta \in \Theta$,

$$\begin{aligned}
\text{subgraph}(f(\cdot,\cdot,\theta)) &= \{(z,w,t) \in \mathcal{X} \times \mathbb{R} : 0 < t < f(z,w,\theta) \text{ or } 0 > t > f(z,w,\theta)\} \\
&= \big(\{z > 0\} \cap \{w'\beta \geq 0\} \cap \{t \geq 1\}^c \cap \{t > 0\}\big) \\
&\quad \cup \big(\{z > 0\} \cap \{w'\beta \geq 0\}^c \cap \{t \geq -1\} \cap \{t \geq 0\}^c\big) \\
&\quad \cup \big(\{z \geq 0\}^c \cap \{w'\beta \geq 0\} \cap \{t \geq -1\} \cap \{t \geq 0\}^c\big) \\
&\quad \cup \big(\{z \geq 0\}^c \cap \{w'\beta \geq 0\}^c \cap \{t \geq 1\}^c \cap \{t > 0\}\big) \\
&= (\{g_1 > 0\} \cap \{g_2 \geq 0\} \cap \{g_3 \geq 1\}^c \cap \{g_3 > 0\}) \\
&\quad \cup (\{g_1 > 0\} \cap \{g_2 \geq 0\}^c \cap \{g_3 \geq -1\} \cap \{g_3 \geq 0\}^c) \\
&\quad \cup (\{g_1 \geq 0\}^c \cap \{g_2 \geq 0\} \cap \{g_3 \geq -1\} \cap \{g_3 \geq 0\}^c) \\
&\quad \cup (\{g_1 \geq 0\}^c \cap \{g_2 \geq 0\}^c \cap \{g_3 \geq 1\}^c \cap \{g_3 > 0\})
\end{aligned}$$

where $g_k(z,w,t) = \alpha_k t + \gamma_k z + \delta_k' w \in \mathcal{G}$ for each k with, $\alpha_1 = 0$, $\gamma_1 = 1$, $\delta_1 = 0$, $\alpha_2 = 0$, $\gamma_2 = 0$, $\delta_2 = \beta$, $\alpha_3 = 1$, $\gamma_3 = 0$, and $\delta_3 = 0$. The collection of sets of the form $\{g \geq 0\}$ or $\{g > 0\}$ is a VC class by Lemma 2.4 of Pakes and Pollard (1989). Furthermore, this property is preserved over complements, unions, and intersections of VC classes by their Lemma 2.5. Therefore, $\{\text{subgraph}(f) : f \in \mathcal{F}\}$ is a VC class, and by Lemma 2.12 of Pakes and Pollard (1989), $\mathcal{F}$ is Euclidean for any envelope. In particular, $\mathcal{F}$ is Euclidean for the constant envelope $F = 1$. ∎

Proof of Theorem 6. We shall verify each of the conditions of Assumption 4. Condition a is satisfied by definition of Model 1, condition b holds as a result of Lemma 1, and condition d is satisfied with $b_n = \sqrt{n}$ as a result of Lemma 3. ∎

D.2. Discrete Regressors

Proof of Lemma 7. Note that the first part of this proof is independent of assumption Assumption 1.

Verification of Assumption 7 First, note that we can rewrite $n^{1/2}Q_n$ as

$$n^{1/2}Q_n(\theta) = n^{1/2}(P_n f_\theta - P f_\theta) + n^{1/2} P f_\theta = \mathbb{G}_n(f_\theta) + n^{1/2} P f_\theta,$$

and therefore,

$$\mathcal{Q}_n \equiv \inf_{\theta \in \Theta_I} n^{1/2} Q_n(\theta) = \inf_{\theta \in \Theta_I} \left(\mathbb{G}_n(f_\theta) + n^{1/2} P f_\theta \right).$$

Supposing Q is normalized so that it is identically zero on Θ_I, since the map $\inf_{\Theta_I}$, which takes real functions on Θ into $\mathbb{R}$, is continuous in $\ell^\infty(\mathcal{F})$, the continuous mapping theorem gives $\mathcal{Q}_n \xrightarrow{d} \inf_{\theta \in \Theta_I} \mathbb{G}(f_\theta) \equiv \mathcal{Q}$.

Verification of Assumption 8 For any sequence of subsets Θ_n of Θ such that $d_H(\Theta_n, \Theta_I) \downarrow 0$ in probability, define $\mathcal{Q}'_n \equiv \inf_{\theta \in \Theta_n} n^{1/2} Q_n(\theta)$. For all $\varepsilon > 0$, there exists an n_ε such that for all $n \geq n_\varepsilon$, $P(\Theta_n = \Theta_I) \geq 1 - \varepsilon$. Then, $P\left(\inf_{\Theta_n} n^{1/2} Q_n = \inf_{\Theta_I} n^{1/2} Q_n\right) \geq 1 - \varepsilon$. Recall from above that $\inf_{\theta \in \Theta_I} n^{1/2} Q_n \xrightarrow{\text{d}} \mathcal{Q}$. Therefore, $\mathcal{Q}'_n \xrightarrow{\text{d}} \mathcal{Q}$. ∎

D.3. Continuous Regressors

Proof of Theorem 8. Lemma 2 established that $\mathcal{F}$ is Euclidean and the conditions of Assumption 4 have been shown to hold previously with $b_n = n^{1/2}$. We will show that Assumption 6 holds with $\gamma_1 = 2$ and $\gamma_2 = 3/2$ and then use Theorem 5 to obtain the resulting rate.

Abrevaya and Huang (2005) show that $\nabla_{\theta\theta'} Q(\theta_0) = -V(\theta_0)$. Generalizing their argument to the set identified case yields $\nabla_{\theta\theta'} Q(\theta) = -V(\theta)$ for all $\theta \in \text{bd}(\Theta_I)$. Therefore, in a neighborhood $\mathcal{N}$ of Θ_I, Q is approximately quadratic and for some $C > 0$, $Q(\theta) \leq \sup Q - C \cdot d^2(\theta, \Theta_I)$.

Let $\eta > 0$ and define $\mathcal{F}_\eta \equiv \{f_\theta \in \mathcal{F} : d(\theta, \Theta_I) < \eta\}$. Again, following the arguments of Abrevaya and Huang (2005), $\mathcal{F}_\eta$ is a VC subgraph class with envelope F_η such that $PF_\eta^2 = O_p(\eta)$. Then, by Lemma 4.1 of Kim and Pollard (1990), for all $\epsilon > 0$, there exists a sequence $M_n = O_p(1)$ such that

$$P_n f_\theta - P f_\theta \leq \epsilon d^2(\theta, \Theta_I) + n^{-2/3} M_n^2 \tag{12}$$

for $d(\theta, \Theta_I) \leq \eta$.

Let $\mathbb{G}_n(\theta) \equiv n^{1/2}(P_n f_\theta - P f_\theta)$ denote the standardized empirical process. For $\theta \in \mathcal{N}^c$,

$$\begin{aligned} Q_n(\theta) &\leq n^{-1/2} \mathbb{G}_n(\theta) + \sup_\Theta Q - \delta \\ &\leq n^{-1/2} O_p(1) + \sup_\Theta Q - \delta \\ &\leq \sup_\Theta Q - \tilde{\delta} \end{aligned}$$

for sufficiently large n. The final inequality is a result of the Donsker property which implies $\sup_{\theta \in \Theta} |\mathbb{G}_n(f_\theta)| = O_p(1)$.

For $\theta \in \mathcal{N}$, we can choose $\epsilon = \frac{1}{2} C$ in (12) and combine this with the quadratic approximation above to obtain

$$\begin{aligned} Q_n(\theta) &= (P_n f_\theta - P f_\theta) + P f_\theta \\ &\leq (P_n f_\theta - P f_\theta) + \sup_\Theta Q - C \cdot d^2(\theta, \Theta_I) \\ &\leq n^{-2/3} M_n^2 + \sup_\Theta Q - \frac{1}{2} C \cdot d^2(\theta, \Theta_I). \end{aligned}$$

Note that the term $n^{-2/3}M_n^2$ is smaller than $\frac{1}{4}Cd^2(\theta,\Theta_I)$ whenever $d(\theta,\Theta_I) \geq \frac{4M_n^2}{Cn^{-1/3}}$.

For any $\varepsilon > 0$ we can choose δ, κ_ε and n_ε such that for all $n \geq n_\varepsilon$, with probability at least $1-\varepsilon$,

$$Q_n(\theta) \leq \sup_{\Theta} Q - C \cdot (d(\theta,\Theta_I) \wedge \delta)^2$$

uniformly on $\{\theta \in \Theta : d(\theta,\Theta_I) \geq (\kappa_\varepsilon/n^{1/2})^{2/3}\}$. This follows since we can choose n_ε large enough to guarantee that set $D_n \equiv \{\theta : d(\theta,\Theta_I) \geq (\kappa_\varepsilon/b_n)^{1/\gamma_2}\}$ intersects the neighborhood $\mathcal{N}$.

Thus, we have verified Assumption 6 with $\gamma_1 = 2$, $\gamma_2 = 3/2$, and $b_n = n^{1/2}$. The conclusion then follows by applying Theorem 5. ■

E. Lagged Dependent Variable Model

Proof of Theorem 2. Let $\tilde{\Theta}_I$ denote the right hand side of (6). The derivation in the text shows that $\theta \in \Theta_I$ implies $\theta \in \tilde{\Theta}_I$.

Now, let $\theta \in \tilde{\Theta}_I$ and let $P(y_{0:3} \mid x)$ denote the population distribution of (y_0, y_1, y_2, y_3) given x. That is, the distribution of observable outcomes conditional on observable covariates. We will show that there exist distributions $F_{u_0|xc}$, $F_{c|x}$, and $p_0(x,c)$ such that $\pi(y_{0:3} \mid x; \theta, F_{u_0|xc}, F_{c|x}, p_0(x,c)) = P(y_{0:3} \mid x)$ F_x-almost surely. By the structure of the model, we only need to consider the following probabilities: $P(y_0 = 0 \mid x)$, $P(y_1 = 0 \mid x, y_0)$, $P(y_2 = 0 \mid x, y_1)$, and $P(y_3 = 0 \mid x, y_2)$. Conditioning on additional lags of y_t or considering events for which $y_t = 1$ would be redundant. Let d_t denote a realized value of y_t. For a fixed x, the model implications corresponding to the aforementioned probabilities are:

$$\begin{aligned}
\pi(y_0 = 0 \mid x; \theta, F_{u_0|xc}, F_{c|x}, p_0(x,c)) &= \int p_0(x,c)\, dF_{c|x}, \\
\pi(y_1 = 0 \mid x, d_0; \theta, F_{u_0|xc}, F_{c|x}, p_0(x,c)) &= \int F_{u_0|xc}(-x_1'\beta - \gamma d_0 - c)\, dF_{c|x}, \\
\pi(y_2 = 0 \mid x, d_1; \theta, F_{u_0|xc}, F_{c|x}, p_0(x,c)) &= \int F_{u_0|xc}(-x_2'\beta - \gamma d_1 - c)\, dF_{c|x}, \\
\pi(y_3 = 0 \mid x, d_2; \theta, F_{u_0|xc}, F_{c|x}, p_0(x,c)) &= \int F_{u_0|xc}(-x_3'\beta - \gamma d_2 - c)\, dF_{c|x}.
\end{aligned}$$

We must consider the ordering of the population probabilities in order to construct valid cdfs. Suppose, for example, for the fixed values of x and y that $P(y_0 = 0 \mid x) < P(y_1 = 0 \mid x, d_0 = 0) < P(y_2 = 0 \mid x, d_1 = 0) < P(y_3 = 0 \mid x, d_2 = 0) < P(y_1 = 0 \mid x, d_0 = 1) < P(y_2 = 0 \mid x, d_1 = 1) < P(y_3 = 0 \mid x, d_2 = 1)$. Then, using the assumption that $\theta \in \tilde{\Theta}_I$ and letting (d_0, d_3) vary over $\{0,1\}^2$ implies that $-x_1'\beta < -x_2'\beta < -x_3'\beta < -x_1'\beta - \gamma < -x_2'\beta - \gamma < -x_3'\beta - \gamma$.

If we choose $F_{c|x}$ such that it places a point mass at $c^*(x)$, as in the proof of Theorem 1, then we can choose G as follows:

$$G(u) = \begin{cases} 0 & \text{if } u < -x_1'\beta - c^*(x), \\ P(y_0 = 0 \mid x) & \text{if } -x_0'\beta - c^*(x) \leq u \leq -x_1'\beta - c^*(x), \\ P(y_1 = 0 \mid x, d_0 = 0) & \text{if } -x_1'\beta - c^*(x) \leq u < -x_2'\beta - c^*(x), \\ P(y_2 = 0 \mid x, d_1 = 0) & \text{if } -x_2'\beta - c^*(x) \leq u < -x_3'\beta - c^*(x), \\ P(y_3 = 0 \mid x, d_2 = 0) & \text{if } -x_3'\beta - c^*(x) \leq u < -x_1'\beta - \gamma - c^*(x), \\ P(y_1 = 0 \mid x, d_0 = 1) & \text{if } -x_1'\beta - \gamma - c^*(x) \leq u < -x_2'\beta - \gamma - c^*(x), \\ P(y_2 = 0 \mid x, d_1 = 1) & \text{if } -x_2'\beta - \gamma - c^*(x) \leq u < -x_3'\beta - \gamma - c^*(x), \\ P(y_3 = 0 \mid x, d_1 = 1) & \text{if } -x_3'\beta - \gamma - c^*(x) \leq u < \bar{u}, \\ 1 & \text{if } u \geq \bar{u}, \end{cases}$$

Finally, if we choose $p_0(x, c) = P(y_0 \mid x)$ for all c then the model implications match the population choice probabilities for the fixed value of x. A similar procedure can be carried out for each x. ■

Proof of Lemma 4. This proof parallels the proof of Lemma 1, the corresponding result for Model 1. Let $\Theta^* \equiv \arg\max_\Theta Q$ and define $w_t \equiv x_t - x_{t-1}$, $z \equiv y_2 - y_1$, and $v \equiv y_3 - y_0$.

Step 1 Let $\theta_1 \in \Theta_I$ and $\theta_2 \in \Theta$. We will show that $Q(\theta_1) \geq Q(\theta_2)$ and therefore, $\theta_1 \in \Theta^*$. We have

$$\begin{aligned} Q(\theta_1) - Q(\theta_2) &= \mathrm{E}\left[1\{w_3 = 0\} \cdot z \cdot \left(\mathrm{sgn}(w_2'\beta_1 + \gamma_1 v) - \mathrm{sgn}(w_2'\beta_2 + \gamma_2 v)\right)\right] \\ &= \int \mathrm{E}[z \mid x, c, y_0, y_3, w_3 = 0] \left(\mathrm{sgn}(w_2'\beta_1 + \gamma_1 v) - \mathrm{sgn}(w_2'\beta_2 + \gamma_2 v)\right) dF_{x,c,y_0,y_3|w_3=0} \\ &= 2\int_{D(\theta_1,\theta_2)} \mathrm{sgn}(w_2'\beta_1 + \gamma_1 v)\, \mathrm{E}[z \mid x, c, y_0, y_3, w_3 = 0]\, dF_{x,c,y_0,y_3|w_3=0} \end{aligned}$$

where $D(\theta_1, \theta_2)$ is defined as the set of all (x, c, v) where $\mathrm{sgn}(w_2'\beta_1 + \gamma_1 v)$ and $\mathrm{sgn}(w_2'\beta_2 + \gamma_2 v)$ differ. The first equality follows by definition of Q, the second is an application of the law of iterated expectations, and the third is due to the fact that on $D(\theta_1, \theta_2)$, $\mathrm{sgn}(w_2'\beta_2 + \gamma_2 v) = -\mathrm{sgn}(w_2'\beta_1 + \gamma_1 v)$. Note that on the integrand above vanishes on the complement of $D(\theta_1, \theta_2)$.

Now, since $\theta_1 \in \Theta_I$, from Theorem 2 we have that for all d_0, d_3,

$$\begin{aligned} \mathrm{sgn}(w_2'\beta_1 + \gamma_1 v) &= \mathrm{sgn}\left(P(A \mid x, x_2 = x_3) - P(B \mid x, x_2 = x_3)\right) \\ &= \mathrm{sgn}\big(P(y_1 = 0, y_2 = 1 \mid x, x_2 = x_3, y_0 = d_0, y_3 = d_3) \\ &\quad - P(y_1 = 1, y_2 = 0 \mid x, x_2 = x_3, y_0 = d_0, y_3 = d_3)\big) \end{aligned}$$

for all $d_0, d_3 \in \{0,1\}$. The second line follows because the common factor which was removed, $P(y_0 = d_0, y_3 = d_3 \mid x, x_2 = x_3)$, is always positive. Furthermore, we can write

$$\begin{aligned} \mathrm{E}[z \mid x, c, y_0, y_3, w_3 = 0] = P(y_1 = 0, y_2 = 1 \mid x, c, y_0, y_3, w_3 = 0) \\ - P(y_1 = 1, y_2 = 0 \mid x, c, y_0, y_3, w_3 = 0). \end{aligned}$$

So, the sign above times the conditional expectation of z simplifies to the absolute value of the conditional expectation. Returning to the objective function,

$$Q(\theta_1) - Q(\theta_2) = 2 \int_{D(\theta_1,\theta_2)} |E[z \mid x, c, y_0, y_3, w_3 = 0]| \, dF_{x,c,y_0,y_3|w_3=0} \geq 0.$$

Step 2 Let $\theta_1 \in \Theta_I$ and suppose there exists a $\theta_2 \in \Theta_I^c \cap \Theta^*$. We will show that this implies $Q(\theta_2) < Q(\theta_1)$, which is a contradiction of the choice of $\theta_2 \in \Theta^*$, and therefore $\Theta_I^c \cap \Theta^*$ must be empty.

Note that we can express Q as

$$\begin{aligned} Q(\theta) = \int_{\{w_3'\beta + \gamma v \geq 0\}} \big[& P(y_1 = 0, y_2 = 1 \mid x, c, y_0, y_3, w_3 = 0) \\ & - P(y_1 = 1, y_2 = 0 \mid x, c, y_0, y_3, w_3 = 0)\big] \, dF_{x,c,y_0,y_3|w_3=0} \\ + \int_{\{w_3'\beta + \gamma v < 0\}} \big[& P(y_1 = 1, y_2 = 0 \mid x, c, y_0, y_3, w_3 = 0) \\ & - P(y_1 = 0, y_2 = 1 \mid x, c, y_0, y_3, w_3 = 0)\big] \, dF_{x,c,y_0,y_3|w_3=0}. \end{aligned}$$

Again we consider a difference $Q(\theta_2) - Q(\theta_1)$. Using the linearity of integrals, we can partition the range of each integral into disjoint sets and subtract the corresponding integrands on each set. When $w_3'\beta_1 + \gamma_1 v$ and $w_3'\beta_2 + \gamma_2 v$ have the same sign, the difference is zero, so we only need to consider regions where the sign differs:

$$\begin{aligned} D_1 &\equiv \{(x, c, y_0, y_3) : w_3'\beta_2 + \gamma_2 v \geq 0, w_3'\beta_1 + \gamma_1 v < 0\}, \\ D_2 &\equiv \{(x, c, y_0, y_3) : w_3'\beta_2 + \gamma_2 v < 0, w_3'\beta_1 + \gamma_1 v \geq 0\}. \end{aligned}$$

Hence,

$$\begin{aligned} Q(\theta_2) - Q(\theta_1) = \int_{D_1} \big[& P(y_1 = 0, y_2 = 1 \mid x, c, y_0, y_3, w_3 = 0) \\ & - P(y_1 = 1, y_2 = 0 \mid x, c, y_0, y_3, w_3 = 0)\big] \, dF_{x,c,y_0,y_3|w_3=0} \\ + \int_{D_2} \big[& P(y_1 = 1, y_2 = 0 \mid x, c, y_0, y_3, w_3 = 0) \\ & - P(y_1 = 0, y_2 = 1 \mid x, c, y_0, y_3, w_3 = 0)\big] \, dF_{x,c,y_0,y_3|w_3=0}. \end{aligned}$$

Since $\theta_1 \in \Theta_I$ and $\theta_2 \notin \Theta_I$, first term is strictly negative and the second term is weakly non-positive. ■

Proof of Lemma 5. We follow the same strategy as in the proof of Lemma 2. Define $w_t \equiv x_t - x_{t-1}$, $z \equiv y_2 - y_1$, and $v \equiv y_3 - y_0$, and let $f(w_2, w_3, z, v, \theta) = 1\{w_3 = 0\} \cdot z_2 \cdot [2 \cdot 1\{w_2'\beta + \gamma v \geq 0\} - 1]$. First, note that f can be rewritten as

$$f(w_2, w_3, z, v, \theta) = 1\{w_3 \geq 0\} \cdot 1\{w_3 \leq 0\} \cdot (1\{z_2 > 0\} - 1\{z < 0\}) \\ \cdot \left(1\{w_2'\beta + \gamma v \geq 0\} - 1\{w_2'\beta + \gamma v < 0\}\right).$$

Upon expanding this expression, it is clear that, as before, for any θ we can express subgraph $f(\cdot, \cdot, \cdot, \cdot, \theta)$ as series of intersections and unions (and complements thereof) of the form $\{g \geq 0\}$ and $\{g > 0\}$ for specific coefficient values α of some polynomial

$$g(w_2, w_3, z, v, t, \alpha) = \alpha_1 t + \alpha_2 w_2 + \alpha_3 w_3 + \alpha_4 z + \alpha_5 v.$$

It then follows that $\{\text{subgraph}(f) : f \in \mathcal{F}\}$ is a VC class, and, therefore, $\mathcal{F}$ is Euclidean for any envelope. In particular, it is Euclidean for the envelope $F = 1$. ■

Proof of Theorem 9. We verify each of the conditions of Assumption 4. Condition a is satisfied by definition of Model 2, condition b holds as a result of Lemma 4, and condition d is satisfied with $b_n = \sqrt{n}$ as a result of Lemma 3, since the objective function is of the same form as that of Model 1—only the indexing class of functions $\mathcal{F}$ is different but both are Euclidean with envelope $F = 1$. ■

Proof of Theorem 10. When the support of x is a finite set, henceforth $\mathcal{X}$, the objective function $Q(\theta)$ can be rewritten as follows:

$$\begin{aligned} Q(\theta) = \sum_{y_0 \in \{0,1\}} \sum_{y_3 \in \{0,1\}} \sum_{x \in \mathcal{X}} & P(x) P(y_0 \mid x) P(y_3 \mid x, y_0) \\ & \times [P(y_2 = 1 \mid x, y_0, y_3) - P(y_1 = 1 \mid x, y_0, y_3)] \\ & \times \operatorname{sgn}\left((x_2 - x_1)'\beta + \gamma(y_3 - y_0)\right). \end{aligned}$$

Therefore, $Q(\theta)$ is a step function and there exists a real number $\delta > 0$ with

$$\delta \geq \inf_{(x, y_0, y_3)} P(x) P(y_0 \mid x) P(y_3 \mid x, y_0) \left[P(y_2 = 1 \mid x, y_0, y_3) - P(y_1 = 1 \mid x, y_0, y_3)\right]$$

such that for all $\theta \in \Theta \setminus \Theta_I$, $Q(\theta) \leq \sup_\Theta Q - \delta$. This verifies Assumption 5. Since conditions of Assumption 4 were already established under the current assumptions in the proof of Theorem 9, the result follows by applying Theorem 4. ■

References

Abrevaya, J. (2000). Rank estimation of a generalized fixed-effects regression model. *Journal of Econometrics 95*, 1–23. [2, 25]

Abrevaya, J. and J. Huang (2005). On the bootstrap of the maximum score estimator. *Econometrica 73*, 1175–1204. [19, 20, 45]

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press. [2]

Andrews, D. W. K., S. Berry, and P. Jia (2004). Confidence regions for parameters in discrete games with multiple equilibria, with an application to discount chain store location. Unpublished manuscript, Yale University. [4]

Andrews, D. W. K. and X. Shi (2009). Inference based on conditional moment inequalities. Working paper, Yale University. [4]

Aradillas-Lopez, A. and E. Tamer (2008). The identification power of equilibrium in simple games. *Journal of Business and Economic Statistics 26*, 261–283. [4]

Bajari, P., J. T. Fox, and S. P. Ryan (2008). Evaluating wireless carrier consolidation using semiparametric demand estimation. *Quantitative Marketing and Economics 6*, 299–338. [5]

Beresteanu, A., I. Molchanov, and F. Molinari (2009). Sharp identification regions in models with convex predictions: games, individual choice, and incomplete data. Working Paper CWP27/09, CEMMAP. [4]

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

Bierens, H. J. and J. Hartog (1988). Non-linear regression with discrete explanatory variables, with an application to the earnings function. *Journal of Econometrics 38*, 269–299. [4]

Bugni, F. A. (2008). Bootstrap inference in partially identified models. Working paper, Northwestern University. [4, 12]

Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In J. J. Heckman and B. S. Singer (Eds.), *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press. [10]

Chen, S. (2002). Rank estimation of transformation models. *Econometrica 70*, 1683–1697. [25]

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2009). Identification and estimation of marginal effects in nonlinear panel models. Unpublished manuscript, Massachusetts Institute of Technology. [5]

Chernozhukov, V., H. Hong, and E. Tamer (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica 75*, 1243–1284. [3, 12, 22, 23]

Ciliberto, F. and E. Tamer (2009). Market structure and multiple equilibria in the airline industry. *Econometrica*, forthcoming. [4]

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics 35*, 303–316. [2, 7, 25]

Honoré, B. E. and E. Kyriazidou (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica 68*, 839–874. [5, 10, 21]

Honoré, B. E. and A. Lleras-Muney (2006). Bounds in competing risks models and the war on cancer. *Econometrica 74*, 1675–1698. [4, 25]

Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica 74*, 611–629. [5]

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica 60*, 505–531. [2, 7]

Horowitz, J. L. (1998). *Semiparametric Methods in Econometrics*. New York: Springer. [2, 4]

Horowitz, J. L. and C. F. Manski (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association 95*, 77–88. [4]

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*, 1845–1857. [4]

Khan, S. and E. Tamer (2009). Inference on endogenously censored regression models using conditional moment inequalities. *Journal of Econometrics 152*, 104–119. [4]

Kim, J. and D. Pollard (1990). Cube root asymptotics. *The Annals of Statistics 18*, 191–219. [18, 19, 45]

Kim, K. (2009). Set estimation and inference with models characterized by conditional moment inequalities. Working paper, University of Minnesota. [4, 12]

Komarova, T. (2008). Binary choice models with discrete regressors: Identification and misspecification. Unpublished manuscript, London School of Economics. [5, 9]

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press. [2]

Magnac, T. and E. Maurin (2008). Partial identification in monotone binary models: Discrete regressors and interval data. *Review of Economic Studies 75*, 835–864. [4]

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*, 205–228. [2, 17]

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics 27*, 313–333. [2, 7, 17]

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica 55*, 357–362. [5, 7, 10, 17]

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer-Verlag. [2]

Manski, C. F. and E. Tamer (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica 70*, 519–546. [3, 4, 12]

McFadden, D. L. (1974). Conditional logit analysis of qualitative choice analysis. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press. [2]

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–1057. [18, 43, 44]

Pakes, A., J. Porter, K. Ho, and J. Ishii (2006). Moment inequalities and their application. Unpublished manuscript, Harvard University. [4]

Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer. [3]

Pollard, D. (1989). Asymptotics via empirical processes. *Statistical Science 4*, 341–366. [18]

Ridder, G. (1990). The non-parametric identification of generalized accelerated failure-time models. *Review of Economic Studies 57*, 167–182. [9, 25]

Romano, J. P. and A. M. Shaikh (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference 138*, 2786–2807. [3]

Romano, J. P. and A. M. Shaikh (2009). Inference for the identified set in partially identified econometric models. *Econometrica*, forthcoming. [3]

Rosen, A. (2009). Set identification via quantile restrictions in short panels. Working Paper CWP26/09, CEMMAP. [5]

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley. [39]

Shorack, G. R. (2000). *Probability for Statisticians*. Springer. [39]

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica 77*, 1299–1315. [4]

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies 70*, 147–165. [4]

Tamer, E. (2009). Partial identification in econometrics. Unpublished manuscript, Northwestern University. [2]