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

JASON R. BLEVINS
The Ohio State University
July 20, 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 where we obtain a sharp characterization of the identified set and propose a consistent set estimator, establishing 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. 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 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...
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 for a general class of partially identified models with limited support regressors and provides conditions for consistency, obtaining rates of convergence, and constructing confidence regions. 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 several models and show that depending on the assumptions made on the support of the regressors, the set 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 \( \theta \) which lie in some parameter space \( \Theta \). 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 and are not of interest themselves. 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 \( \theta_0 \) is the only element of \( \Theta \) such that the model would be consistent with the population distribution \( P_{\theta_0} \), 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 \( \theta_0 \), that is, such that \( P_{\theta} = P_{\theta_0} \). The set of all such \( \theta \) is the identified set and is denoted \( \Theta_I \). 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 inference results for two new classes of models: models with continuous but potentially bounded regressors

\(^1\)That is, they can achieve rates arbitrarily close to \( 1/\sqrt{n} \) as in Chernozhukov, Hong, and Tamer (2007).
which may have non-standard rates of convergence and models 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 et al. (2007) in that we propose criterion-function-based set estimators, derive their rates of convergence, and propose a subsampling-based (Politis, Romano, and Wolf, 1999) procedure for obtaining confidence regions. We obtain these results under new conditions which are applicable to the specific cases 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. In addition to nonparametric estimation, they also consider 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 construct confidence regions using subsampling. They
apply their general results to models characterized by moment equalities and inequalities. Romano and Shaikh (2008, 2010) 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.

A second, 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 $\theta_0$. 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 arises due to the inherent properties of the model, not as a result of an implicit assumption.

There are numerous other areas where partially identified econometric models have arisen including, 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.

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 discuss 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 observations are available at times \( t = 0, \ldots, T - 1 \) for each individual. An individual in the model is described completely by 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 \( y \equiv (y_0, \ldots, y_{T-1}) \) 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
and a set of weak semiparametric assumptions on \( F \) to determine the identified set of parameters of interest. We let \( \theta \) denote the finite vector of parameters of interest and we will denote the set of possible values of \( \theta \) by \( \Theta \). We assume \( F \) is induced by some true unknown parameter \( \theta_0 \).

In the models we consider, the distribution of the available regressors may not be rich enough to point identify \( \theta_0 \) without additional assumptions. Therefore, we focus instead on the identified set \( \Theta_I \) which contains \( \theta_0 \) itself, as well as all other parameter vectors which cannot be distinguished from \( \theta_0 \). 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 \( \theta \). First, note that we can always write \( F \) as the product of conditional distributions \( F = F_{y|xu}F_{u|x}F_{c|x}F_x \). 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 \( \theta \) 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 \( F_{y|xu} \) will be fully specified. For example, in panel data discrete choice models, \( F_{y|xu} \) is determined by an underlying latent variable model. Following the fixed effects literature, \( F_{c|x} \) will not be restricted in any way. We will, however, restrict \( F_{u|x} \) with a standard stationarity assumption used in the literature.

3.1. Basic Fixed Effects Panel Data Model

We begin with the fundamental restriction on \( F_{y|xu} \) which defines the basic linear-index fixed effects binary response model.

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

\[
y_t = 1\{x_t' \beta + c + u_t \geq 0\}
\]

where \( x_t \) is a random variable with support \( \mathcal{X} \subseteq \mathbb{R}^k \), \( c \) 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 \( c \), \( F_{u|xc} \) satisfies the following:

a. \( F_{u|xc} = F_{u0|xc} \) 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 $\theta$.

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 $\hat{\beta}$ 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 $\beta$ 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 $\beta$ 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 $\mathcal{X} \subset \mathbb{R}$. 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 the vector \( x_1 - x_0 \) has positive density everywhere on a set \( W_1 \subseteq \mathbb{R} \) for almost every value of the remaining components.

Note that this assumption does not rule out the possibility that \( W_1 = \mathbb{R} \) but it also includes cases where 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 \( t = 1, \ldots, T \),

\[
(2) \quad y_t = 1 \{ x_t' \beta + \gamma y_{t-1} + c + u_t \geq 0 \}
\]

where \( x_t \) is a random vector with support \( \mathcal{X} \), \( c \) 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 \( u_t \) 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 \( t \),

\[
(3) \quad \Lambda(y_t) = x_t' \beta + c + u_t
\]

where \( \Lambda \) is a strictly monotonic function, \( x_t \) is a random vector with support \( \mathcal{X} \), \( c \) 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 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 $F_{u|xc}(u) = 1 - \exp(-e^u)$ and $\Lambda$ 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 $\theta$ be a vector of parameters of interest and let $\Theta$ be the parameter space, the set of all feasible values of $\theta$. 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 $\theta$ and $F_v$. Let $\pi(\cdot \mid \theta, F_v, x)$ denote the distribution of $y \mid x$ implied by the model under $\theta$ 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 $\theta$ given $F_{yx}$ is

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

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 $\Theta_I$, 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 $\Theta_I$ characterizes this set, but when $x_1 - x_0$ has richer support, $\Theta_I$ naturally becomes smaller. For example, in the fixed effects model with discrete regressors, considered below, the number of equality constraints defining $\Theta_I$ 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, \( \Theta_I \) 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 \( \beta, \ F_{u_0|x_c}, \) and \( F_{c|x}. \) We now provide a characterization of \( \Theta_I \) in terms of observables and show that it is equivalent to the identified set defined above. Since \( c \) is unobserved, in order to estimate \( \beta \) 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, \( \Theta_I, \) in terms of observables: \( P(y_0 = 1 \mid x), P(y_1 = 1 \mid x), \) and \( F_x. \)

**Theorem 1.** In Model 1,

\[
\Theta_I = \{ \theta \in \Theta : \text{sgn} \left( P(y_1 = 1 \mid x) - P(y_0 = 1 \mid x) \right) = \text{sgn} \left( (x_1 - x_0)' \beta \right) F_x - a.s. \}.
\]

**Proof.** See Appendix D. \( \Box \)

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, our characterization of the identified set 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 $F_{u_t|x}$ for simplicity, the corresponding choice probabilities are:

$$
P(A \mid x, c, x_2 = x_3) = p_0(x, c)^{1-d_0} (1 - p_0(x, c))^{d_0} G(-x_1' \beta - \gamma d_0 - c)$$

$$\times \left[ 1 - G(-x_2' \beta - c) \right] G(-x_2' \beta - \gamma - c)^{1-d_3}$$

$$\times \left[ 1 - G(-x_2' \beta - \gamma - c) \right]^{d_3},$$

$$P(B \mid x, c, x_2 = x_3) = p_0(x, c)^{1-d_0} (1 - p_0(x, c))^{d_0} \left[ 1 - G(-x_1' \beta - \gamma d_0 - c) \right]$$

$$\times G(-x_2' \beta - \gamma - c)G(-x_2' \beta - c)^{1-d_3}$$

$$\times \left[ 1 - G(-x_2' \beta - c) \right]^{d_3}.$$

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

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

$$\times \left[ \frac{G(-x_2' \beta - \gamma - c)}{G(-x_2' \beta - c)} \right]^{1-d_3} \times \left[ \frac{1 - G(-x_2' \beta - \gamma - c)}{1 - G(-x_2' \beta - c)} \right]^{d_3}.$$

When $d_3 = 0$,

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

and when $d_3 = 1$,

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

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 $c$,

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

This condition provides the foundation for our characterization of the identified set and the results of the derivation above are formalized in the following theorem.

**Theorem 2.** In Model 2,

(6) $\Theta_1 \subseteq \tilde{\Theta}_1 = \{ \theta \in \Theta : \text{sgn} \left( P(A \mid x, x_2 = x_3) - P(B \mid x, x_2 = x_3) \right) = \text{sgn} \left( (x_1 - x_2)' \beta + \gamma (d_3 - d_0) \right) \} \ F_x - \text{a.s.} \ \forall d_0, d_3 \in \{0, 1\}.$
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 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 for 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 $\Theta_I$ 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 $\tau_n$:

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

The “slackness” sequence $\tau_n$ 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 $\tau_n$ such that $\hat{\Theta}_n$ is a consistent estimator of $\Theta_I$.

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 $\Theta$. Let $(\Theta, d)$ 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

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

where, in a slight abuse of notation, $d(\theta, A) \equiv \inf_{\theta' \in A} d(\theta, \theta')$ is the distance between a point $\theta$ and a set $A$. This is illustrated in Figure 1.
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 $\Theta_I$. Note that once $\Theta_I$ is found, constructing $Q$ is straightforward. 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 $\Theta_I$ 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. $\Theta$ is a nonempty subset of $\mathbb{R}^k$ and is compact with respect to the Euclidean metric.

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

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

d. There exists a function $Q_n : \Theta \times X^T \times Y^T \rightarrow \mathbb{R}$, 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 $b_n \rightarrow \infty$.

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.

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

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 $\sup_{\theta \in \Theta_n} d(\theta, \Theta_I) \xrightarrow{p} 0$.

This corollary guarantees that asymptotically, without slackness, $\hat{\Theta}_n$ is close to $\Theta_I$. 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 $x_i$. 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 $x_i$ that imply the required properties of these functions.
In particular, we show that when $Q$ has a discrete jump at the boundary of $\Theta_I$, then $\hat{\Theta}_n$ converges arbitrarily fast in probability to $\Theta_I$. 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 stochastically bounded from above by a polynomial in $d(\theta, \Theta_I)$, 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 $\tau_n$ converges to zero.

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

**Assumption 5 (Existence of a Constant Majorant).** There exists a positive constant $\delta$

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

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

When the above condition holds, $\hat{\Theta}_n$ converges arbitrarily fast to $\Theta_I$. This result is due to the discrete jump in $Q$ at the boundary of $\Theta_I$. 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 determine a partition of $\Theta$ 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 $\tau_n$, 

![Figure 3. Infinite curvature of $Q(\theta)$](image-url)
so that it includes all values of $\theta$ for which $Q_n(\theta) \geq \sup Q_n - \tau_n$, there are four possible estimates: $\Theta_2$, $\Theta_2 \cup \Theta_3$, $\Theta_2 \cup \Theta_3 \cup \Theta_1$, 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 $\Theta_I$ with probability approaching one. Intuitively, as we obtain more data, we are able to detect which values of $\theta$ belong to $\Theta_I$ with increasing accuracy since there is a discrete jump in $Q(\theta)$ for all $\theta$ not in $\Theta_I$. Furthermore, since $\tau_n$ converges to zero in probability slower than $Q_n$ converges uniformly to $Q$, $\hat{\Theta}_n$ converges to $\Theta_I$. 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 $\Theta_I$, $Q_n$ is bounded in probability 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 $\tau_n \overset{p}{\to} 0$ and $\tau_n b_n \overset{p}{\to} \infty,$ then $d_H(\hat{\Theta}_n, \Theta_I) = O_p(\tau_n^{1/\gamma_2}).$

**Proof.** See Appendix B.

### 5.3. Fixed Effects Binary Choice Model

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 regressor, 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):

$$Q(\theta) = E[(y_1 - y_0) \sgn((x_1 - x_0)\beta)].$$

The corresponding finite sample analog objective function is

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_{1i} - y_{0i}) \sgn((x_{1i} - x_{0i})\beta).$$

Note that although essentially the same objective function is used for maximum score estimation in the point identified case, the set estimators proposed here are fundamentally different since they are defined as 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 $\theta.$

**Lemma 1** below establishes the equivalence between the identified set $\Theta_I$ and the set of maximizers of the population objective function.
Lemma 1. Under the maintained assumptions of Model 1,
\[ \arg \max_{\theta \in \Theta} Q(\theta) = \Theta_1. \]

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 \( F, \) indexed by \( \Theta, \) such that \( Q(\theta) = P f_\theta \) and \( Q_n(\theta) = P_n f_\theta \) for \( f_\theta \in F. \) We begin by defining \( F \) 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 \( F = \{ f(\cdot, \cdot, \theta) : \theta \in \Theta \} \) 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. \( F \) 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 \( \tau_n \to 0, \) and \( \tau_n n^{1/2} \to \infty, \) then \( d_H(\hat{\Theta}_n, \Theta_1) \to 0. \)

Proof. See Appendix D.
5.3.3. Rates of Convergence

The rate of convergence of \( \hat{\Theta}_n \) to \( \Theta_I \) in Model 1 depends on the support of \( x \). We obtain the rate under both Assumption 1 and Assumption 2. We show that when the support of \( x \) is finite, \( \hat{\Theta}_n \) converges arbitrarily fast in probability to \( \Theta_I \). 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 \( \tau_n \) 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, \( \tau_n \approx \sqrt{\ln n/n} \).

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 \( \Theta_I \).

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:

\[
Q(\theta) = E_x E_{y|x} \left[ (y_1 - y_0) \text{sgn} \left( (x_1 - x_0)'\beta \right) \right] = \sum_{x \in \mathcal{X}} P(x) \left[ P(y_1 = 1 \mid x) - P(y_0 = 1 \mid x) \right] \text{sgn} \left( (x_1 - x_0)'\beta \right).
\]

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, \( \delta \) 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 \( \tau_n \) such that \( \tau_n \xrightarrow{P} 0 \) and \( n^{1/2}\tau_n \xrightarrow{P} \infty \), then \( \hat{\Theta}_n \) converges to \( \Theta_I \) 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 in the continuous covariate case 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 \( \beta \) 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 \( S^{k-1} \equiv \{ x \in \mathbb{R}^k : \| x \| = 1 \} \) in \( \mathbb{R}^k \) and assume that the angular component of \( x \) has continuous, bounded density with respect to the surface measure on \( S^{k-1} \).
Then for any sequence $\tau_n$, the limiting objective function is approximately quadratic near the identified set: $\Gamma(\theta) = O(d^2(\theta, \Theta))$. The variance of the empirical process component is $O_p(d(\theta, \Theta_1)/n)$. When the trend overtakes the noise, $\Gamma_n$ 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 $\sqrt{d(\theta, \Theta_1)/n} > d^2(\theta, \Theta_1)$, or, $d(\theta, \Theta_1) < n^{-1/3}$. Therefore, $\hat{\Theta}_n$ the set of near maximizers of $\Gamma_n$, should be within an $n^{-1/3}$ neighborhood of $\Theta_1$. 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 $\tau_n$ can be chosen arbitrarily close to $n^{-1/2}$, the rate of convergence can be made arbitrarily close to $(n^{-1/2})^{1/\gamma} = n^{-1/3}$. 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 $\tilde{w}w'$ have finite first absolute moments.

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

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 $\partial F(v | \tilde{w}, w_1)/\partial w_1$ exists and $|\partial F(v | \tilde{w}, w_1)/\partial w_1| < M$.

e. $\Theta_1$ is contained in the interior of $\Theta$.

f. The matrix $V(\theta) \equiv E[2f(0 | \tilde{w}, -\tilde{w}'\beta)g(-\tilde{w}'\beta | \tilde{w})\tilde{w}w']$ is positive semidefinite for all $\theta \in \text{bd}(\Theta_1)$.

Then for any sequence $\tau_n$ such that $\tau_n \overset{P}{\to} 0$ and $n^{1/2}\tau_n \overset{P}{\to} \infty$, $d_H(\hat{\Theta}_n, \Theta_1) = O_p(\tau_n^{2/3})$.

**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 (a 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

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

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

\[
Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left( x_{i2} = x_{i3} \right) \cdot (y_{i2} - y_{i1}) \cdot \text{sgn}\left( (x_{i2} - x_{i1}) \beta + \gamma (y_{i3} - y_{i0}) \right).
\]

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 \( \hat{\Theta}_I \)).** Under the maintained assumptions of Model 2, \( \arg \max_{\theta \in \Theta} Q(\theta) = \hat{\Theta}_I \) (as defined in Theorem 2).

**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 \( \hat{\Theta}_n \) for \( \hat{\Theta}_I \). As in the fixed effects model, we begin by establishing that the objective function belongs to a Euclidean class of functions indexed by \( \theta \) 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 \( f_\theta(x,y) = 1\{x_2 = x_3\}(y_2 - y_1) \cdot 2 \cdot 1\{(x_2 - x_1) \beta + \gamma (y_3 - y_0) \geq 0\} - 1 \}, \) 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 \( d_H(\hat{\Theta}_n, \hat{\Theta}_I) \xrightarrow{p} 0 \).

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 \( \hat{\Theta}_I \) 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 \( r_n, r_n d_H(\hat{\Theta}_n, \hat{\Theta}_I) \xrightarrow{p} 0 \).

Proof. See Appendix E.

6. Confidence Regions

Confidence regions for \( \Theta_I \) 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

\[
(9) \quad C_n(\kappa_n) = \{ \theta \in \Theta : b_n Q_n(\theta) \geq \sup_{\Theta} b_n Q_n - \kappa_n \}.
\]

Inference is based on the statistic

\[
Q_n = \sup_{\Theta} b_n Q_n - \inf_{\Theta_I} b_n Q_n
\]

and the following equivalence:

\[
\{ \Theta_I \subseteq C_n(\kappa_n) \} \iff \{ Q_n \leq \kappa_n \}.
\]

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 \( Q_n \). As a result, the sequence \( \kappa_n \) is analogous to \( b_n \tau_n \). Thus, while in Theorem 3 we required \( \tau_n \xrightarrow{p} 0 \) and \( b_n \tau_n \xrightarrow{p} \infty \) 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, \( \kappa_n \) 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 \( Q_n \) is continuous, Chernozhukov et al. (2007) provide methods of constructing confidence regions which cover \( \Theta_I \) 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 \( Q_n \). 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 $Q$. 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 $Q_n$, the limiting distribution of $Q_n$. 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 $Q_n$ has a nondegenerate limiting distribution.

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

Lemma 6. Suppose Assumption 7 holds. Then, for any sequence $\hat{c}_n$ such that $\hat{c}_n \overset{P}{\to} c(1 - \alpha) \equiv \inf\{c : P\{Q \leq c\} \geq 1 - \alpha\}$ 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. \:\box

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 $\kappa_n$ be any sequence such that such that $C_n(\kappa_n)$ is a consistent estimator of $\Theta_I$ (e.g., $\kappa_n \propto \sqrt{\ln n}$).

2. Compute $\hat{c}_n$ as the $1 - \alpha$ quantile of the values $\{\hat{Q}_{n,m,j}\}_{j=1}^{M_n}$ where

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

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 $Q$ may not be continuous. As a result, our confidence regions are conservative since we cannot place an upper bound on the coverage probability.
3. Report $C_n(\kappa_n)$ as a consistent estimate of $\Theta_I$ 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 $a$.

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

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

Proof. See Appendix C ■

6.2. Fixed Effects Binary Choice 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 $Q_n$ required by Assumption 7 and the approximability of $Q_n$ based on a sequence of estimates $\hat{O}_n$, 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.

---

4See Appendix A.1 for a precise definition of $a_n \downarrow a$, both deterministically and in probability.
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 $f_\theta$ and both functions satisfy Assumption 5. That is, in both cases, for the appropriate class of functions $F = \{f_\theta : \theta \in \Theta\}$, we have $Q(\theta) = Pf_\theta$ and $Q_n(\theta) = P_n f_\theta$.

Since both classes of functions $F$ 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 $x_0' \beta$:

$$P(y_1 \geq y_0 | x, c) \geq P(y_0 \geq y_1 | x, c) \iff P(x_1' \beta + u_1 \geq x_0' \beta + u_0 | x, c) \geq P(x_0' \beta + u_0 \geq x_1' \beta + u_1 | x, c)$$

$$\iff P(u_0 - u_1 \leq (x_1 - x_0)' \beta | x, c) \geq P(u_1 - u_0 \leq (x_0 - x_1)' \beta | x, c)$$

$$\iff (x_1 - x_0)' \beta \geq 0$$
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:

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

This set is guaranteed to contain \( \Theta_I \) 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:

\[
Q(\theta) = \mathbb{E} \left[ \text{sgn}(y_1 - y_0) \cdot \text{sgn} \left( (x_1 - x_0)'\beta \right) \right]
\]

\[
Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \text{sgn}(y_{i1} - y_{i0}) \cdot \text{sgn} \left( (x_{i1} - x_{i0})'\beta \right)
\]

Due to the 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 the Transformation Function

In this model, in addition to \( \beta \), one might be interested in estimating the transformation function \( \Lambda \). This section discusses estimating bounds for \( \Lambda(\tilde{y}) \) at particular values of \( \tilde{y} \).

Since we can only identify \( \Lambda(\tilde{y}) \) up to differences with respect to \( \Lambda(\tilde{y}_0) \) at some value \( \tilde{y}_0 \), we normalize \( \Lambda(\tilde{y}_0) = 0 \).

Suppose first that \( \theta_0 \) is known. Then, again following the maximum score principle, we could estimate bounds for \( \Lambda(\tilde{y}) \) by collecting all values of \( \lambda \) which maximize

\[
\frac{1}{n} \sum_{i=1}^{n} (1\{y_{i1} > \tilde{y}\} - 1\{y_{i0} > \tilde{y}_0\}) 1\{(x_{i0} - x_{i1})'\beta_0 \leq \lambda\}.
\]

Estimating the set above is infeasible because \( \theta_0 \) 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

\[
\Gamma_n(\tilde{y}, \lambda, \theta) = \frac{1}{n} \sum_{i=1}^{n} (1\{y_{i1} > \tilde{y}\} - 1\{y_{i0} > \tilde{y}_0\}) 1\{(x_{i0} - x_{i1})'\beta \leq \lambda\}
\]
and forming a set estimate \( \hat{\Lambda}_n(y) \) of \( \Lambda(y) \) which consists of all values of \( \lambda \) which maximize \( \Gamma_n \) for some \( \theta \in \hat{\Theta}_n \). That is,

\[
\hat{\Lambda}_n(y) = \{ \lambda : \lambda \in \arg \max \Gamma_n(y, \lambda, \hat{\theta}) \text{ for some } \hat{\theta} \in \hat{\Theta}_n \}.
\]

Establishing the asymptotic properties of two-stage estimators such as \( \hat{\Lambda}_n(y) \), 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 choose \( C \in \{0.20, 0.10, 0.05, 0.01\} \). These values were chosen to be roughly around the same magnitude as \( Q_n \). 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 \( \kappa_n \) 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.
\[ \kappa_n \quad n \quad \text{Mean } \hat{\Theta}_n \quad \text{St. Dev.} \quad \text{Coverage} \]

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

Table 1. Fixed effects model estimates. $\theta_0 = -0.150$, $\Theta_i = [-0.163, -0.148]$. 
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 $\tau_n$ 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, as do the corresponding confidence regions.

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

<table>
<thead>
<tr>
<th>$\kappa_n$</th>
<th>$m$</th>
<th>$n$</th>
<th>0.750</th>
<th>0.900</th>
<th>0.950</th>
<th>0.990</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>250</td>
<td>0.700</td>
<td>0.941</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0.672</td>
<td>0.948</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1000</td>
<td>0.621</td>
<td>0.960</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2000</td>
<td>0.660</td>
<td>0.979</td>
<td>0.992</td>
</tr>
<tr>
<td>$0.20\sqrt{\ln n}$</td>
<td>$n^{2/5}$</td>
<td>4000</td>
<td>0.824</td>
<td>0.990</td>
<td>0.993</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8000</td>
<td>0.866</td>
<td>0.991</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16000</td>
<td>0.968</td>
<td>0.991</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32000</td>
<td>0.994</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64000</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>$0.10\sqrt{\ln n}$</td>
<td>$n^{2/5}$</td>
<td>4000</td>
<td>0.508</td>
<td>0.944</td>
<td>0.985</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8000</td>
<td>0.653</td>
<td>0.978</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16000</td>
<td>0.834</td>
<td>0.986</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32000</td>
<td>0.926</td>
<td>0.997</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64000</td>
<td>0.984</td>
<td>0.998</td>
<td>0.999</td>
</tr>
<tr>
<td>$0.05\sqrt{\ln n}$</td>
<td>$n^{2/5}$</td>
<td>4000</td>
<td>0.362</td>
<td>0.739</td>
<td>0.888</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8000</td>
<td>0.446</td>
<td>0.811</td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16000</td>
<td>0.558</td>
<td>0.903</td>
<td>0.970</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32000</td>
<td>0.728</td>
<td>0.966</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64000</td>
<td>0.837</td>
<td>0.978</td>
<td>0.993</td>
</tr>
<tr>
<td>$0.01\sqrt{\ln n}$</td>
<td>$n^{2/5}$</td>
<td>4000</td>
<td>0.308</td>
<td>0.342</td>
<td>0.366</td>
<td>0.414</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8000</td>
<td>0.355</td>
<td>0.461</td>
<td>0.541</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16000</td>
<td>0.420</td>
<td>0.494</td>
<td>0.550</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32000</td>
<td>0.515</td>
<td>0.604</td>
<td>0.661</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64000</td>
<td>0.640</td>
<td>0.764</td>
<td>0.797</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>$\kappa_n$</th>
<th>$m$</th>
<th>$n$</th>
<th>0.750</th>
<th>0.900</th>
<th>0.950</th>
<th>0.990</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>0.943</td>
<td>0.979</td>
<td>0.987</td>
<td>0.992</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.934</td>
<td>0.978</td>
<td>0.994</td>
<td>0.993</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.910</td>
<td>0.976</td>
<td>0.984</td>
<td>0.996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.936</td>
<td>0.989</td>
<td>0.991</td>
<td>0.997</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$0.20\sqrt{\ln n}$</th>
<th>$n^{3/5}$</th>
<th>$n$</th>
<th>0.750</th>
<th>0.900</th>
<th>0.950</th>
<th>0.990</th>
</tr>
</thead>
<tbody>
<tr>
<td>4000</td>
<td>0.978</td>
<td>0.989</td>
<td>0.990</td>
<td>0.995</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td>0.985</td>
<td>0.991</td>
<td>0.994</td>
<td>0.995</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td>0.986</td>
<td>0.994</td>
<td>0.997</td>
<td>0.997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32000</td>
<td>0.997</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$0.10\sqrt{\ln n}$</th>
<th>$n^{3/5}$</th>
<th>$n$</th>
<th>0.750</th>
<th>0.900</th>
<th>0.950</th>
<th>0.990</th>
</tr>
</thead>
<tbody>
<tr>
<td>4000</td>
<td>0.922</td>
<td>0.972</td>
<td>0.982</td>
<td>0.987</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td>0.945</td>
<td>0.979</td>
<td>0.993</td>
<td>0.992</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td>0.965</td>
<td>0.989</td>
<td>0.995</td>
<td>0.994</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32000</td>
<td>0.984</td>
<td>0.996</td>
<td>0.996</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64000</td>
<td>0.994</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$0.05\sqrt{\ln n}$</th>
<th>$n^{3/5}$</th>
<th>$n$</th>
<th>0.750</th>
<th>0.900</th>
<th>0.950</th>
<th>0.990</th>
</tr>
</thead>
<tbody>
<tr>
<td>4000</td>
<td>0.859</td>
<td>0.926</td>
<td>0.939</td>
<td>0.957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td>0.900</td>
<td>0.918</td>
<td>0.926</td>
<td>0.945</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td>0.898</td>
<td>0.954</td>
<td>0.965</td>
<td>0.974</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32000</td>
<td>0.939</td>
<td>0.975</td>
<td>0.980</td>
<td>0.982</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64000</td>
<td>0.969</td>
<td>0.983</td>
<td>0.992</td>
<td>0.995</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$0.01\sqrt{\ln n}$</th>
<th>$n^{3/5}$</th>
<th>$n$</th>
<th>0.750</th>
<th>0.900</th>
<th>0.950</th>
<th>0.990</th>
</tr>
</thead>
<tbody>
<tr>
<td>4000</td>
<td>0.307</td>
<td>0.412</td>
<td>0.432</td>
<td>0.447</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td>0.554</td>
<td>0.629</td>
<td>0.649</td>
<td>0.667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td>0.583</td>
<td>0.609</td>
<td>0.612</td>
<td>0.617</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32000</td>
<td>0.665</td>
<td>0.671</td>
<td>0.676</td>
<td>0.688</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64000</td>
<td>0.798</td>
<td>0.800</td>
<td>0.804</td>
<td>0.807</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** Fixed effects model confidence regions, $m = n^{3/5}$. 
<table>
<thead>
<tr>
<th>( \kappa_n )</th>
<th>( m )</th>
<th>( n )</th>
<th>0.750</th>
<th>0.900</th>
<th>0.950</th>
<th>0.990</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>0.952</td>
<td>0.977</td>
<td>0.989</td>
<td>0.990</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.961</td>
<td>0.987</td>
<td>0.992</td>
<td>0.997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.966</td>
<td>0.989</td>
<td>0.996</td>
<td>0.997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.979</td>
<td>0.994</td>
<td>0.995</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20\sqrt{\ln n} \quad n^{4/5}</td>
<td>4000</td>
<td>0.989</td>
<td>0.996</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td>0.987</td>
<td>0.997</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td>0.991</td>
<td>0.996</td>
<td>0.998</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32000</td>
<td>0.997</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10\sqrt{\ln n} \quad n^{4/5}</td>
<td>4000</td>
<td>0.960</td>
<td>0.991</td>
<td>0.995</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td>0.968</td>
<td>0.992</td>
<td>0.995</td>
<td>0.996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td>0.977</td>
<td>0.994</td>
<td>0.998</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32000</td>
<td>0.993</td>
<td>0.997</td>
<td>0.997</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64000</td>
<td>0.998</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05\sqrt{\ln n} \quad n^{4/5}</td>
<td>4000</td>
<td>0.890</td>
<td>0.905</td>
<td>0.916</td>
<td>0.934</td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td>0.889</td>
<td>0.918</td>
<td>0.934</td>
<td>0.955</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td>0.939</td>
<td>0.961</td>
<td>0.965</td>
<td>0.968</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32000</td>
<td>0.965</td>
<td>0.980</td>
<td>0.979</td>
<td>0.982</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64000</td>
<td>0.974</td>
<td>0.991</td>
<td>0.995</td>
<td>0.995</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01\sqrt{\ln n} \quad n^{4/5}</td>
<td>4000</td>
<td>0.396</td>
<td>0.410</td>
<td>0.413</td>
<td>0.423</td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td>0.575</td>
<td>0.606</td>
<td>0.616</td>
<td>0.634</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16000</td>
<td>0.589</td>
<td>0.601</td>
<td>0.606</td>
<td>0.616</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32000</td>
<td>0.666</td>
<td>0.679</td>
<td>0.679</td>
<td>0.685</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64000</td>
<td>0.798</td>
<td>0.801</td>
<td>0.806</td>
<td>0.806</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Fixed effects model confidence regions, \( m = n^{4/5} \).
\begin{table}
\centering
\begin{tabular}{cccc}
\hline
\text{\(n\)} & \text{Mean \(\hat{\Theta}_n\)} & \text{St. Dev.} & \text{Coverage} \\
\hline
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 \\
\hline
\end{tabular}
\caption{Fixed effects model estimates \((\kappa_n = 0, m = n^{3/5})\).}
\end{table}

\begin{table}
\centering
\begin{tabular}{cccccc}
\hline
\text{\(n\)} & \text{Empirical Coverage} \\
\hline
& \text{0.750} & \text{0.900} & \text{0.950} & \text{0.990} \\
\hline
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 \\
\hline
\end{tabular}
\caption{Fixed effects model confidence regions \((\kappa_n = 0, m = n^{3/5})\).}
\end{table}
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.

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 $a \lor b$ to denote $\max\{a, b\}$ and $a \land b$ 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 $B^\varepsilon$ denote an $\varepsilon$-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 \textit{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 \( x \in A \),

\[ f(x) - \sup_{y \in A} g(y) \leq f(x) - g(x) \leq |f(x) - g(x)| \]

and

\[ \sup_{y \in A} f(y) - g(x) \geq f(x) - g(x) \geq -|f(x) - g(x)|. \]

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):

\[ \sup_{x \in A} f(x) - \sup_{x \in A} g(x) = \sup_{x \in A} f(x) + \inf_{x \in A} (-g(x)) \]

\[ = \inf_{x \in A} \sup_{y \in A} f(y) - g(x) \]

\[ \geq \inf_{x \in A} -|f(x) - g(x)| \]

\[ = -\sup_{x \in A} |f(x) - g(x)| \]

Together, these two inequalities imply the result.

\[ \blacksquare \]

B. Consistent Estimation

Proof of Theorem 3. The proof proceeds in two steps. In the first step, we show that \( \sup_{\theta \in \Theta_n} d(\theta, \Theta_I) \overset{p}{\to} 0 \). 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 \( B^c \) denote an \( \epsilon \)-expansion of a set \( B \), as defined in Subsubsection A.1.

Step 1. For any \( \epsilon > 0 \),

\[ \sup_{\Theta \setminus \Theta_I^c} Q_n \leq \sup_{\Theta \setminus \Theta_I^c} Q + o_p(1) \leq \sup_{\Theta} Q - \delta_\epsilon + o_p(1), \]

\[ \sup_{\Theta} Q - \delta_\epsilon + o_p(1), \]

35
where $\delta_\epsilon > 0$. 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 $\Theta_I$ maximizes $Q$. Similarly,

$$\inf Q_n \geq \sup Q_n - \tau_n \geq \sup Q - \tau_n + o_p(1)$$

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 $\delta_\epsilon > 0$, with probability approaching one, $\tau_n < \delta_\epsilon$, or equivalently, $\sup_{\Theta} Q - \tau_n + o_p(1) \geq \sup_{\Theta} Q - \delta_\epsilon + o_p(1)$. Given the inequalities above, this implies $\inf_{\Theta_n} Q_n \geq \sup_{\Theta \setminus \Theta_I} Q_n$, which in turn implies that $\Theta_n \subseteq \Theta_I$, and so $\sup_{\theta \in \Theta_n} d(\theta, \Theta_I) \leq \epsilon$.

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

$$\sup_{\Theta} Q_n - \inf_{\Theta_I} Q_n = \left[ \sup_{\Theta} Q_n - \sup_{\Theta} Q \right] + \left[ \sup_{\Theta} Q - \inf_{\Theta_I} Q_n \right]$$

$$\leq \left| \sup_{\Theta} Q_n - \sup_{\Theta} Q \right| + \left| \sup_{\Theta} Q - \inf_{\Theta_I} Q_n \right|$$

$$= \left| \sup_{\Theta} Q_n - \sup_{\Theta} Q \right| + \left| \sup_{\Theta} Q - \inf_{\Theta_I} Q_n \right|$$

$$\leq \sup_{\Theta} |Q_n - Q| + \sup_{\Theta_I} |Q_n - Q|$$

$$\leq \sup_{\Theta} |Q_n - Q| + \sup_{\Theta} |Q_n - Q|$$

These steps follow by, respectively, adding and subtracting $\sup_{\Theta} Q$, taking the absolute value, noting that $\Theta_I$ 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 $\tau_n b_n \overset{p}{\to} \infty$ (i.e., that $\tau_n$ 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

$$\sup_{\Theta \setminus \Theta_I} Q_n \leq \sup_{\Theta \setminus \Theta_I} Q + O_p(1/b_n) \leq \sup_{\Theta} Q + O_p(1/b_n) \leq \sup_{\Theta} Q_n - \delta + O_p(1/b_n),$$
where the second inequality follows from Assumption 5.

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

**Proof of Theorem 5.** For any \( \varepsilon > 0 \), let \( \delta, \kappa, \gamma_1, \gamma_2, \kappa_\varepsilon \) and \( n_\varepsilon \) satisfy Assumption 6 and define

\[
v_n = \left( \frac{\kappa \cdot \kappa_\varepsilon \vee b_n \cdot \tau_n}{b_n \cdot \kappa} \right)^{1/2},
\]

where \( b_n \) is given by Assumption 4.d. Then, since \( v_n = o_p(1) \), \( v_n = O_p(\tau_n^{1/\gamma_2}) \), and \( \tau_n b_n \to \infty \), there is an \( n'_\varepsilon > n_\varepsilon \) such that for all \( n > n'_\varepsilon \), with probability at least \( 1 - \varepsilon \), we have both \( v_n \leq \delta \) and \( v_n \geq (\kappa_\varepsilon / b_n)^{1/\gamma_2} \). On a set \( \Theta \setminus \Theta_t^{n_\varepsilon} \), the distance satisfies \( d(\theta, \Theta_t) \geq v_n \), so \( \min\{d(\theta, \Theta_t), \delta\} \geq \min\{v_n, \delta\} \). Therefore, by Assumption 6,

\[
\sup_{\Theta \setminus \Theta_t^{n_\varepsilon}} Q_n \leq \sup_{\Theta} Q_n - \kappa \cdot (v_n \wedge \delta)^{\gamma_1/\gamma_2}
\]

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

**C. Confidence Regions**

**Proof of Lemma 6.** Observe that

\[
P\{\Theta_t \subseteq C_n(\hat{\epsilon}_n)\} = P\{Q_n \leq \hat{\epsilon}_n\} = P\{Q \leq c(1 - \alpha)\} + o_p(1) \geq (1 - \alpha) + o_p(1).
\]

The first equality holds by definition of \( C_n \) and \( Q_n \), the second by Assumption 7 and \( \hat{\epsilon}_n \to c(1 - \alpha) \), 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 \( \hat{Q}_{n,m,j} \) such that \( \underline{Q}_{n,m,j} \leq \hat{Q}_{n,m,j} \leq \overline{Q}_{n,m,j} \) with probability approaching one.
Next, we prove that the empirical distribution function of $\hat{Q}_{n,m,j}$ converges in probability to the distribution function of $Q$, the limiting distribution of $Q_n$. Finally, we show that $\hat{c}_n$ converges in probability to $c(1 - \alpha)$, the desired quantile of the distribution of $Q$.

Step 1 By Theorem 4, we have $d_H(C_n(\kappa_n), \Theta_1) = 0$ with probability approaching one. Thus, $d_H(C_n(\kappa_n), \Theta_1) \leq \varepsilon_n$ for some sequence $\varepsilon_n \downarrow 0$ with probability approaching one. For a fixed subsample $j$, let $\mathcal{Q}_{n,m,j} \equiv \sup_{\theta \in \Theta} b_m Q_{n,m,j}(\theta) - \inf_{\theta \in \Theta} b_m Q_{n,m,j}(\theta)$. Let $\mathcal{K}_n$ be the collection of all subsets $K \subseteq \Theta$ such that $d_H(K, \Theta_1) \leq \varepsilon_n$ and define $\overline{\mathcal{Q}}_{n,m,j} \equiv \sup_{K \in \mathcal{K}_n} \left[ \sup_{\theta \in K} b_m Q_{n,m,j}(\theta) - \inf_{\theta \in K} b_m Q_{n,m,j}(\theta) \right]$. 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 $C_n(\kappa_n) \subseteq \Theta_1$ and $C_n(\kappa_n) \in \mathcal{K}_n$, we have $\overline{\mathcal{Q}}_{n,m,j} \leq \mathcal{Q}_{n,m,j} \leq \underline{\mathcal{Q}}_{n,m,j}$ for all $j = 1, \ldots, M_n$.

Step 2 From Step 1, with probability approaching one,

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

We will show that $G_{n,m}(x) \overset{p}{\rightarrow} P\{Q \leq x\}$ and $\overline{G}_{n,m}(x) \overset{p}{\rightarrow} P\{Q \leq x\}$ as $n \to \infty$ (and thus, $m \to \infty$). Therefore, $\hat{G}_{n,m}(x) \overset{p}{\rightarrow} P\{Q \leq x\}$ for each $x \in \mathbb{R}$.

Let $I_m(x)$ denote the cdf of $\underline{\mathcal{Q}}_{n,m,j}$. Note that $G_{n,m}(x)$ is a U-statistic of degree $m$ with $0 \leq G_{n,m}(x) \leq 1$ (i.e., it is bounded). Furthermore, $E[G_{n,m}(x)] = E[\mathbb{1}\{\mathcal{Q}_{n,m,j} \leq x\}] = I_m(x)$, 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 $t > 0$,

$$P\{G_{n,m}(x) - I_m(x) \geq t\} \leq \exp \left[ -2t^2 \frac{n}{m}\right].$$

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

A similar argument shows that $\overline{G}_{n,m}(x) \overset{p}{\rightarrow} P\{Q \leq x\}$ as well, and therefore, $\hat{G}_{n,m}(x) \overset{p}{\rightarrow} P\{Q \leq x\}$.

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, $\hat{c}_n = \inf\{x : \hat{G}(x) \geq 1 - \alpha\} \overset{p}{\rightarrow} c(1 - \alpha)$. $\blacksquare$
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 \( \Theta_1 \) denote the identified set as defined in (4) and let \( \tilde{\Theta}_1 \) denote the set on the right side of (5). We first show \( \Theta_1 \subseteq \tilde{\Theta}_1 \), and then \( \tilde{\Theta}_1 \subseteq \Theta_1 \).

Step 1 Let \( \theta \in \Theta_1 \). By definition of \( \Theta_1 \), 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 \( F_{u_0|xc} \),

\[
P(y_1 = 1 \mid x, c) \geq P(y_0 = 1 \mid x, c) \iff 1 - F_{u_0|xc}(-x_1'\beta - c) \geq 1 - F_{u_0|xc}(-x_0'\beta - c)
\]

\[
\iff F_{u_0|xc}(-x_1'\beta - c) \leq F_{u_0|xc}(-x_0'\beta - c)
\]

\[
\iff -x_1'\beta - c \leq -x_0'\beta - c
\]

\[
\iff (x_1 - x_0)'\beta \geq 0
\]

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,

\[
sgn (P(y_1 = 1 \mid x) - P(y_0 = 1 \mid x)) = sgn ((x_1 - x_0)'\beta).
\]

Therefore, \( \theta \in \Theta_1 \Rightarrow \theta \in \tilde{\Theta}_1 \).

Step 2 Now, suppose \( \theta \in \tilde{\Theta}_1 \). We will show that for each such \( \theta \), 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:

\[
P(y_0 = 0 \mid x) = \pi(y_0 = 0 \mid x; \theta, F_{u_0|xc}, F_{c|x})
\]

\[
P(y_1 = 0 \mid x) = \pi(y_1 = 0 \mid x; \theta, F_{u_0|xc}, F_{c|x})
\]
For a given $x$ and for primitives $(\theta, F_{u0|x}, F_{c|x})$, the model implications are:

\[
\pi(y_0 = 0 \mid x; \theta, F_{u0|x}, F_{c|x}) = \int F_{u0|x}(-x'_0\beta - c) \, dF_{c|x}
\]

\[
\pi(y_1 = 0 \mid x; \theta, F_{u0|x}, F_{c|x}) = \int F_{u0|x}(-x'_1\beta - c) \, dF_{c|x}
\]

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 $x$):

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

Suppose that $P(y_1 = 1 \mid x) < P(y_0 = 1 \mid x)$ (the opposite case follows similarly). Then our choice of $\theta \in \Theta$ guarantees that $\beta$ 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 $F_{u0|x}$ is a valid cdf:

\[
F_{u0|x}(u) = \begin{cases} 
0 & \text{if } u < -x'_1\beta - c^*(x), \\
P(y_0 = 0 \mid x) & \text{if } -x'_0\beta - c^*(x) \leq u \leq -x'_1\beta - c^*(x), \\
P(y_1 = 0 \mid x) & \text{if } -x'_1\beta - c^*(x) \leq u < \bar{u}, \\
1 & \text{if } u \geq \bar{u},
\end{cases}
\]

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:

\[
\pi(y_0 = 0 \mid x; \theta, F_{u0|x}, F_{c|x}) = F_{u0|x}(-x'_0\beta - c^*(x)) = P(y_0 = 0 \mid x),
\]

\[
\pi(y_1 = 0 \mid x; \theta, F_{u0|x}, F_{c|x}) = F_{u0|x}(-x'_1\beta - c^*(x)) = P(y_1 = 0 \mid x).
\]

Therefore $\theta \in \Theta$, and since $\theta \in \overline{\Theta}$ was chosen arbitrarily, $\overline{\Theta} \subseteq \Theta$. 

---

**Figure 8.** Two distributions $F_{u0|x}$ with equivalent observable implications under $F_{c|x}$. 

---

40
Proof of Lemma 1. 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 \( \theta_1 \) and \( \theta_2 \), \( Q(\theta_1) \geq Q(\theta_2) \).

Consider the difference

\[
Q(\theta_1) - Q(\theta_2) = E \left[ z \cdot \operatorname{sgn}(w'\beta_1) \right] - E \left[ z \cdot \operatorname{sgn}(w'\beta_2) \right]
\]

\[
= E \left[ z \left( \operatorname{sgn}(w'\beta_1) - \operatorname{sgn}(w'\beta_2) \right) \right]
\]

\[
= 2 \int_{D(\theta_1, \theta_2)} \operatorname{sgn}(w'\beta_1) E [z | x, c] dF_{xc}
\]

where \( D(\theta_1, \theta_2) = \{(x, c) : \operatorname{sgn}(w_1 \beta_1) \neq \operatorname{sgn}(w_2 \beta_2) \} \) is the set of values of \( x \) and \( c \) where \( \operatorname{sgn}(w_1 \beta_1) \) and \( \operatorname{sgn}(w_2 \beta_2) \) differ. The last equality above follows from the fact that the integrand vanishes on complement of \( D(\theta_1, \theta_2) \), and that on \( D(\theta_1, \theta_2) \) we have \( \operatorname{sgn}(w_1 \beta_1) = -\operatorname{sgn}(w_2 \beta_2) \), implying that \( \operatorname{sgn}(w_1 \beta_1) - \operatorname{sgn}(w_2 \beta_2) = 2 \operatorname{sgn}(w_1 \beta_1) \). Since \( \theta_1 \in \Theta_I \), Theorem 1 guarantees that

\[
\operatorname{sgn}(w_1 \beta_1) = \operatorname{sgn}(P(y_1 = 1 | x, c) - P(y_0 = 1 | x, c)) = \operatorname{sgn} E(z | x, c)
\]

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

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

for all \( \theta_2 \). Therefore, \( \Theta_I \subseteq \Theta^* \).

Step 2 Now, let \( \theta_1 \in \Theta_I \) and suppose there exists a \( \theta_2 \in \Theta_I \cap \Theta^* \), where \( \Theta_I^c \) denotes the complement of \( \Theta_I \). We will use the definition of \( \Theta_I \) 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:

\[
Q(\theta) = E[z \cdot \operatorname{sgn}(w'\beta)]
\]

\[
= E_{xc} E_{z|wc} [z \cdot \operatorname{sgn}(w'\beta)]
\]

\[
= E_{xc} \left[ (P(y_1 = 1 | x, c) - P(y_0 = 1 | x, c)) \left( \{w' \beta \geq 0\} - \{w' \beta < 0\} \right) \right]
\]

\[
= \int_{\{w' \beta \geq 0\}} (P(y_1 = 1 | x, c) - P(y_0 = 1 | x, c)) dF_{xc}
\]

\[
+ \int_{\{w' \beta < 0\}} (P(y_0 = 1 | x, c) - P(y_1 = 1 | x, c)) dF_{xc}
\]

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 \( Q(\theta_2) - Q(\theta_1) \):

\[
Q(\theta_2) - Q(\theta_1) = \int_{\{w^\prime \beta_2 \geq 0\}} \left( P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c) \right) dF_x \\
+ \int_{\{w^\prime \beta_2 < 0\}} \left( P(y_0 = 1 \mid x, c) - P(y_1 = 1 \mid x, c) \right) dF_x \\
- \int_{\{w^\prime \beta_1 \geq 0\}} \left( P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c) \right) dF_x \\
- \int_{\{w^\prime \beta_1 < 0\}} \left( P(y_0 = 1 \mid x, c) - P(y_1 = 1 \mid x, c) \right) dF_x
\]

Over regions where \( w^\prime \beta_2 \) and \( w^\prime \beta_1 \) have the same sign, the difference is zero, therefore

\[
Q(\theta_2) - Q(\theta_1) = \int_{\{w^\prime \beta_2 \geq 0, w^\prime \beta_1 < 0\}} \left( P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c) \right) dF_x \\
- \int_{\{w^\prime \beta_2 < 0, w^\prime \beta_1 \geq 0\}} \left( P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c) \right) dF_x
\]

From the proof of Theorem 1, we know that for \( \theta_1 \in \Theta_1 \),

\[
P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c) \geq 0 \iff w^\prime \beta_1 \geq 0
\]

and for \( \theta_2 \in \Theta_1^* \),

\[
P(y_1 = 1 \mid x, c) - P(y_0 = 1 \mid x, c) < 0 \iff w^\prime \beta_2 \geq 0.
\]

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 \( \theta_2 \), meaning that \( \Theta_1^* \cap \Theta^* = \emptyset \) and therefore it must be the case
that \( \Theta^* \subseteq \Theta_1 \).

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

\[
g(z, w, t, \alpha, \gamma, \delta) = at + \gamma z + \delta^t w
\]

and define

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

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 $C_1$ and $C_2$ are VC classes, then so are $\{C_1 \cap C_2 : C_1 \in C_1, C_2 \in C_2\}$, $\{C_1 \cup C_2 : C_1 \in C_1, C_2 \in C_2\}$, and $\{C_1 : C_1 \in C_1\}$.

First, note that we can rewrite $f$ as

$$f(z, w, \theta) = (1\{z > 0\} - 1\{z < 0\}) \cdot (1\{w' \beta \geq 0\} - 1\{w' \beta < 0\})$$

$$= 1\{z > 0, w' \beta \geq 0\} - 1\{z > 0, w' \beta < 0\}$$

$$- 1\{z < 0, w' \beta \geq 0\} + 1\{z < 0, w' \beta < 0\}.$$

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

$$\text{subgraph}(f(\cdot, \cdot, \theta)) = \{(z, w, t) \in X \times R : 0 < t < f(z, w, \theta) \text{ or } 0 > t > f(z, w, \theta)\}$$

$$= \{(z > 0) \cap \{w' \beta \geq 0\} \cap \{t \geq 1\} \cap \{t > 0\}\}$$

$$\cup \{(z > 0) \cap \{w' \beta \geq 0\}^c \cap \{t \geq -1\} \cap \{t \geq 0\}^c\}$$

$$\cup \{(z \geq 0)^c \cap \{w' \beta \geq 0\} \cap \{t \geq -1\} \cap \{t \geq 0\}^c\}$$

$$\cup \{(z \geq 0)^c \cap \{w' \beta \geq 0\} \cap \{t \geq 1\} \cap \{t > 0\}\}$$

$$= (\{g_1 > 0\} \cap \{g_2 \geq 0\} \cap \{g_3 \geq 1\}^c \cap \{g_3 > 0\})$$

$$\cup (\{g_1 > 0\} \cap \{g_2 \geq 0\}^c \cap \{g_3 \geq -1\} \cap \{g_3 \geq 0\}^c)$$

$$\cup (\{g_1 \geq 0\}^c \cap \{g_2 \geq 0\} \cap \{g_3 \geq -1\} \cap \{g_3 \geq 0\}^c)$$

$$\cup (\{g_1 \geq 0\}^c \cap \{g_2 \geq 0\}^c \cap \{g_3 \geq 1\}^c \cap \{g_3 > 0\})$$

where $g_k(z, w, t) = a_k t + \gamma_k z + \delta_k w \in \mathcal{G}$ for each $k$ with, $a_1 = 0$, $\gamma_1 = 1$, $\delta_1 = 0$, $a_2 = 0$, $\gamma_2 = 0$, $\delta_2 = \beta$, $a_3 = 1$, $\gamma_3 = 0$, 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
\[ n^{1/2}Q_n(\theta) = n^{1/2}(P_nf_\theta - Pf_\theta) + n^{1/2}Pf_\theta = G_n(f_\theta) + n^{1/2}Pf_\theta, \]
and therefore,
\[ Q_n \equiv \inf_{\theta \in \Theta} n^{1/2}Q_n(\theta) = \inf_{\theta \in \Theta} \left( G_n(f_\theta) + n^{1/2}Pf_\theta \right). \]
Supposing $Q$ is normalized so that it is identically zero on $\Theta_1$, since the map $\inf_{\Theta_1}$, which takes real functions on $\Theta$ into $\mathbb{R}$, is continuous in $\ell^\infty(\mathcal{F})$, the continuous mapping theorem gives $Q_n \overset{d}{\to} \inf_{\theta \in \Theta_1} G(f_\theta) \equiv Q$.

Verification of Assumption 8  For any sequence of subsets $\Theta_n$ of $\Theta$ such that $d_H(\Theta_n, \Theta_1) \downarrow 0$ in probability, define $Q'_n \equiv \inf_{\Theta_n} n^{1/2}Q_n(\theta)$. For all $\varepsilon > 0$, there exists an $n_\varepsilon$ such that for all $n \geq n_\varepsilon$, $P(\Theta_n = \Theta_1) \geq 1 - \varepsilon$. Then, $P(\inf_{\Theta_n} n^{1/2}Q_n = \inf_{\Theta_1} n^{1/2}Q_n) \geq 1 - \varepsilon$. Recall from above that $\inf_{\Theta_n} n^{1/2}Q_n \overset{d}{\to} Q$. Therefore, $Q'_n \overset{d}{\to} Q$.  

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 $\theta \in \text{bd}(\Theta_1)$. Therefore, in a neighborhood $\mathcal{N}$ of $\Theta_1$, $Q$ is approximately quadratic and for some $C > 0$, $Q(\theta) \leq \sup Q - C \cdot d^2(\theta, \Theta_1)$.

Let $\eta > 0$ and define $\mathcal{F}_\eta \equiv \{ f_\theta \in \mathcal{F} : d(\theta, \Theta_1) < \eta \}$. Again, following the arguments of Abrevaya and Huang (2005), $\mathcal{F}_\eta$ is a VC subgraph class with envelope $F_\eta$ such that $P F_\eta^2 = O_p(\eta)$. Then, by Lemma 4.1 of Kim and Pollard (1990), for all $\varepsilon > 0$, there exists a sequence $M_n = O_p(1)$ such that
\[ (12) \quad P_nf_\theta - Pf_\theta \leq \varepsilon d^2(\theta, \Theta_1) + n^{-2/3}M_n^2 \]
for $d(\theta, \Theta_1) \leq \eta$.

Let $G_n(\theta) \equiv n^{1/2}(P_nf_\theta - Pf_\theta)$ denote the standardized empirical process. For $\theta \in \mathcal{N}$,
\[ Q_n(\theta) \leq n^{-1/2}G_n(\theta) + \sup_{\Theta} Q - \delta \]
\[ \leq n^{-1/2}O_p(1) + \sup_{\Theta} Q - \delta \]
\[ \leq \sup_{\Theta} Q - \tilde{\delta} \]
for sufficiently large $n$. The final inequality is a result of the Donsker property which implies $\sup_{\theta \in \Theta} |G_n(f_\theta)| = O_p(1)$.

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

$$Q_n(\theta) = (P_n f_\theta - P_f_\theta) + P_f_\theta$$

$$\leq (P_n f_\theta - P_f_\theta) + \sup_{\theta} Q - C \cdot d^2(\theta, \Theta_1)$$

$$\leq n^{-2/3} M_n^2 + \sup_{\theta} Q - \frac{1}{2} C \cdot d^2(\theta, \Theta_1).$$

Note that the term $n^{-2/3} M_n^2$ is smaller than $\frac{1}{2} C d^2(\theta, \Theta_1)$ whenever $d(\theta, \Theta_1) \geq \frac{4M_n^2}{Cn^{-1/3}}$.

For any $\epsilon > 0$ we can choose $\delta, \kappa_\epsilon$ and $n_\epsilon$ such that for all $n \geq n_\epsilon$, with probability at least $1 - \epsilon$,

$$Q_n(\theta) \leq \sup_{\theta} Q - C \cdot (d(\theta, \Theta_1) \wedge \delta)^2$$

uniformly on $\{\theta \in \Theta : d(\theta, \Theta_1) \geq (\kappa_\epsilon / n^{1/2})^{2/3}\}$. This follows since we can choose $n_\epsilon$ large enough to guarantee that set $D_n \equiv \{\theta \in \Theta : d(\theta, \Theta_1) \geq (\kappa_\epsilon / b_n)^{1/2}\}$ 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 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 $w_1 \equiv x_t - x_{t-1}$, $z \equiv y_2 - y_1$, and $v \equiv y_3 - y_0$.

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

We have

$$Q(\theta_1) - Q(\theta_2) = E \left[1\{w_3 = 0\} \cdot z \cdot (\text{sgn}(w_2^\prime \beta_1 + \gamma_1 v) - \text{sgn}(w_2^\prime \beta_2 + \gamma_2 v)) \right]$$

$$= \int E[z \mid x, c, y_0, y_3, w_3 = 0] (\text{sgn}(w_2^\prime \beta_1 + \gamma_1 v) - \text{sgn}(w_2^\prime \beta_2 + \gamma_2 v)) \ dF_{x, c, y_0, y_3 \mid w_3 = 0}$$

$$= 2 \int_{D(\theta_1, \theta_2)} \text{sgn}(w_2^\prime \beta_1 + \gamma_1 v) E[z \mid x, c, y_0, y_3, w_3 = 0] \ dF_{x, c, y_0, y_3 \mid w_3 = 0}$$

where $D(\theta_1, \theta_2)$ is defined as the set of all $(x, c, v)$ where $\text{sgn}(w_2^\prime \beta_1 + \gamma_1 v)$ and $\text{sgn}(w_2^\prime \beta_2 + \gamma_2 v)$ 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 $D(\theta_1, \theta_2)$,
\[ \text{sgn}(w_2^\prime \beta_2 + \gamma_2 v) = -\text{sgn}(w_2^\prime \beta_1 + \gamma_1 v). \] Note that on the integrand above vanishes on the complement of \( D(\theta_1, \theta_2). \)

Now, since \( \theta_1 \in \Theta_1 \), from Theorem 2 we have that for all \( d_0, d_3 \),

\[ \text{sgn}(w_2^\prime \beta_1 + \gamma_1 v) = \text{sgn} \left( P(A \mid x, x_2 = x_3) - P(B \mid x, x_2 = x_3) \right) \]

\[ = \text{sgn} \left( P(y_1 = 0, y_2 = 1 \mid x, x_2 = x_3, y_0 = d_0, y_3 = d_3) - P(y_1 = 1, y_2 = 0 \mid x, x_2 = x_3, y_0 = d_0, y_3 = d_3) \right) \]

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

\[ \text{E}[z \mid x, c, y_0, y_3, w_3 = 0] = P(y_1 = 0, y_2 = 1 \mid x, c, y_0, y_3, w_3 = 0) \]

\[ - P(y_1 = 1, y_2 = 0 \mid x, c, y_0, y_3, w_3 = 0). \]

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

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

**Step 2** Let \( \theta_1 \in \Theta_1 \) and suppose there exists a \( \theta_2 \in \Theta_1^c \cap \Theta^* \). 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 \( \Theta_1^c \cap \Theta^* \) must be empty.

Note that we can express \( Q \) as

\[ Q(\theta) = \int_{\{w_2^\prime \beta + \gamma v \geq 0\}} \left[ P(y_1 = 0, y_2 = 1 \mid x, c, y_0, y_3, w_3 = 0) \right] dF_{x,c,y_0,y_3 \mid w_3=0} \]

\[ + \int_{\{w_2^\prime \beta + \gamma v < 0\}} \left[ P(y_1 = 1, y_2 = 0 \mid x, c, y_0, y_3, w_3 = 0) \right] dF_{x,c,y_0,y_3 \mid w_3=0}. \]

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 \( w_2^\prime \beta_1 + \gamma_1 v \) and \( w_2^\prime \beta_2 + \gamma_2 v \) have the same sign, the difference is zero, so we only need to consider regions where the sign differs:

\[ D_1 \equiv \{(x, c, y_0, y_3) : w_2^\prime \beta_2 + \gamma_2 v \geq 0, w_2^\prime \beta_1 + \gamma_1 v < 0\}, \]

\[ D_2 \equiv \{(x, c, y_0, y_3) : w_2^\prime \beta_2 + \gamma_2 v < 0, w_2^\prime \beta_1 + \gamma_1 v \geq 0\}. \]
Hence,

\[
Q(\theta_2) - Q(\theta_1) = \int_{D_1} \left[ P(y_1 = 0, y_2 = 1 \mid x, c, y_0, y_3, w_3 = 0) 
- P(y_1 = 1, y_2 = 0 \mid x, c, y_0, y_3, w_3 = 0) \right] dF_{x,c,y_0,y_3|w_3=0} 
+ \int_{D_2} \left[ P(y_1 = 1, y_2 = 0 \mid x, c, y_0, y_3, w_3 = 0) 
- P(y_1 = 0, y_2 = 1 \mid x, c, y_0, y_3, w_3 = 0) \right] dF_{x,c,y_0,y_3|w_3=0}.
\]

Since \(\theta_1 \in \Theta_1\) and \(\theta_2 \not\in \Theta_1\), 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 \(w_t = x_t - x_{t-1}\), \(z \equiv y_2 - y_1\), and \(v \equiv y_3 - y_0\), and let \(f(w_2, w_3, z, v, \theta) = 1\{w_3 = 0\} \cdot z_2 \cdot [2 \cdot 1\{w_2^2 + \gamma v \geq 0\} - 1]\). First, note that \(f\) can be rewritten as

\[
f(w_2, w_3, z, v, \theta) = 1\{w_3 \geq 0\} \cdot 1\{w_3 \leq 0\} \cdot (1\{z_2 > 0\} - 1\{z < 0\}) 
\cdot (1\{w_2^2 + \gamma v \geq 0\} - 1\{w_2^2 + \gamma v < 0\})
\]

Upon expanding this expression, it is clear that, as before, for any \(\theta\) 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 \(\alpha\) 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:

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


Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley. [38]

Shorack, G. R. (2000). *Probability for Statisticians*. Springer. [38]

