Econometrica, Vol. 73, No. 4 (July, 2005), 1175–1204

ON THE BOOTSTRAP OF THE MAXIMUM SCORE ESTIMATOR

BY JASON ABREVAYA AND JIAN HUANG[1]

This paper shows that the bootstrap does not consistently estimate the asymptotic distribution of the maximum score estimator. The theory developed also applies to other estimators within a cube-root convergence class. For some single-parameter estimators in this class, the results suggest a simple method for inference based upon the bootstrap.

KEYWORDS: Maximum score estimation, bootstrap, cube-root asymptotics.

1. INTRODUCTION

THE MAXIMUM SCORE ESTIMATOR of Manski (1975) was the first semiparametric estimator proposed for the (latent-variable) binary response model. The model considered by Manski (1975) replaced parametric assumptions on the error disturbance with a conditional median restriction,

$$(1)\qquad y=\{x'\beta_0-\epsilon\geq 0\},\quad \mathrm{Median}(\epsilon|x)=0,$$

where $\{\cdot\}$ denotes an indicator function. If $(y_1,x_1),\ldots,(y_n,x_n)$ are the observed data that satisfy (1), the maximum score estimator β_n is defined as the maximizer of the objective function[2]

$$(2)\qquad n^{-1}\sum_{i=1}^{n}(2y_i-1)\{x_i'\beta\geq 0\}$$

over a suitably normalized parameter space (since β is only identified up to scale). Manski (1985) proved consistency of the maximum score estimator under mild regularity conditions. The estimator is appealing since consistency does not require a parametric assumption or a homoskedasticity assumption on the error disturbance.

The maximum score estimator is not $\sqrt{n}$ consistent, but rather converges at the slower $\sqrt[3]{n}$ rate, as shown by Kim and Pollard (1990). Unfortunately, the asymptotic distribution derived by Kim and Pollard (1990) is too complicated to be used for inference purposes. As a result, resampling methods have proven to be the only feasible tools for performing inference with the maximum score estimator. The bootstrap has been used in various applications of

[1]The co-editor, two anonymous referees, and seminar participants at Ohio State University and University of Chicago provided many helpful comments. The first author is grateful for support from the National Science Foundation (Grant SES-0451660) and the John and Mary Willis Young Faculty Scholar Fund at Purdue University.

[2]Other equivalent formulations of the objective function can be used, such as $n^{-1}\sum_i(y_i\{x_i'\beta\geq 0\}+(1-y_i)\{x_i'\beta<0\})$ or $n^{-1}\sum_i|y_i-\{x_i'\beta\geq 0\}|$.

the maximum score estimator (e.g., Bartik, Butler, and Liu (1992), Goodwin (1996), Mayer et al. (2002)) despite the fact that the bootstrap's validity in this context has never been proven. Early simulations by Manski and Thompson (1986) suggested that the bootstrap provides a reasonable approximation for the estimator's root-mean-squared error (RMSE).[3] Somewhat surprisingly, no other simulation evidence regarding the bootstrap appeared in the literature until Delgado, Rodriguez-Poo, and Wolf (2001), who theoretically justify the use of subsampling for the maximum score estimator and provide simulation evidence that suggests inconsistency of the bootstrap. As an alternative approach, Horowitz (1992) proposes a smoothed maximum score estimator that (under additional smoothness assumptions) is asymptotically normal and converges at a rate than can approach $\sqrt{n}$. For this smoothed estimator, Horowitz (2002) has shown that the bootstrap is consistent and even provides asymptotic refinements.

This paper shows that the bootstrap does not consistently estimate the asymptotic distribution of the maximum score estimator. This inconsistency result follows from a more general result concerning inconsistency of the bootstrap for a class of $\sqrt[3]{n}$-consistent estimators. For this class of estimators, Kim and Pollard (1990) show that the asymptotic distribution is the maximizer of a Gaussian process minus a quadratic form. In contrast, the bootstrap distribution converges to the difference between the maximizers of two such processes. In the one-dimensional parameter case, the bootstrap distribution is characterized by a pronounced central peak and long tails, as opposed to the nearly normal-shaped asymptotic distribution.

Section 2 contains the main results on the limiting distribution of the bootstrap and its inconsistency for a class of $\sqrt[3]{n}$-consistent estimators. Section 3 considers the maximum score estimator in detail and derives the asymptotic and bootstrap distributions. In a single-covariate model (where only the intercept is estimated), the maximum score estimator is shown to be asymptotically distributed as the maximum of two-sided Brownian motion minus quadratic drift. Section 4 focuses on other single-parameter cube-root estimators for which the bootstrap-inconsistency result applies. Despite the inconsistency of the bootstrap for these estimators, the asymptotic and bootstrap distributions have a relationship that allows for a method of asymptotic inference based on the bootstrap. The appeal of using the bootstrap for inference in this class of problems is that the asymptotic distribution often involves quantities that must be nonparametrically estimated.

2. MAIN RESULTS

Let $z_1, z_2, \dots$ be a sequence of independent random variables from a common distribution P on a probability space $(\mathcal{Z}, \mathcal{A})$. Suppose that an estimator

[3]Manski and Thompson (1986) suggest doubling the bootstrap estimate of RMSE as a conservative upper bound on the RMSE of the estimator.

based on the first n observations is defined to be the value θ_n that maximizes an objective function $M_n(\theta) \equiv n^{-1}\sum_{i=1}^{n} m(z_i, \theta)$ over a parameter space $\Theta \subseteq \mathbb{R}^d$, where $\{m(\cdot, \theta) : \theta \in \Theta\}$ is a class of functions indexed by Θ.

Let $\hat{z}_1, \ldots, \hat{z}_n$ be a random sample drawn with replacement from the collection of the sample values $z_1, \ldots, z_n$. Then the nonparametric bootstrap estimator of Efron (1979) is the value $\hat{\theta}_n$ that maximizes $\hat{M}_n(\theta) \equiv n^{-1}\sum_{i=1}^{n} m(\hat{z}_i, \theta)$.

Let $M(\theta) \equiv Pm(\cdot, \theta)$ be the population version of $M_n(\theta)$.[4] Suppose $M(\theta)$ has a unique maximizer θ_0. To simplify notation (by avoiding continual recentering), assume without loss of generality that $m(\cdot, \theta_0) \equiv 0$. For $\eta > 0$, define the class of functions

$$\mathcal{M}_\eta \equiv \{m(\cdot, \theta) : |\theta - \theta_0| \leq \eta\}. \tag{3}$$

Define the envelope function of $\mathcal{M}_\eta$ by

$$M_\eta(\cdot) \equiv \sup_{|\theta-\theta_0|\leq\eta} |m(\cdot, \theta)|. \tag{4}$$

Any function that is no less than M_η can also be used as an envelope function. Assume that $M_\eta \in L_2(P)$ and let ϕ be a continuous function that satisfies $\phi(\eta) \geq PM_\eta^2$. The following assumptions, which are similar to those in Theorem 3.2.10 of Van der Vaart and Wellner (1996), are made.

ASSUMPTION 1: *Suppose θ_0 is an interior point of Θ.*

ASSUMPTION 2: *Suppose $M(\theta)$ is twice continuously differentiable at θ_0, with $V \equiv -\nabla_{\theta\theta} M(\theta_0)$.*

ASSUMPTION 3: *Suppose*

$$H(s, t) \equiv \lim_{\alpha\to 0} \frac{Pm(\cdot, \theta_0 + \alpha s)m(\cdot, \theta_0 + \alpha t)}{\phi(\alpha)}$$

exists for each $s, t \in \mathbb{R}^d$.

ASSUMPTION 4: *For every $\eta > 0$,*

$$\lim_{\alpha\to 0} \frac{PM_\alpha^2\{M_\alpha > \eta\alpha^{-2}\phi(\alpha)\}}{\phi(\alpha)} = 0.$$

[4]Here and throughout we use the linear functional notation $Qf = \int f\,dQ$ for a function f and a probability measure Q when the integral is well defined.

ASSUMPTION 5: *For any finite* $k > 0$,

$$\lim_{\epsilon\to 0}\limsup_{\alpha\to 0}\sup_{\substack{|s-t|<\epsilon\\ \max\{|s|,|t|\}\le k}}\frac{P(m(\cdot,\theta_0+\alpha s)-m(\cdot,\theta_0+\alpha t))^2}{\phi(\alpha)}=0.$$

ASSUMPTION 6: *There exists an* $\eta_0 > 0$ *such that for* $\eta \in (0, \eta_0]$, *the classes* $\mathcal{M}_\eta$ *are uniformly manageable for the envelopes* M_η *in the sense of Kim and Pollard* (1990, *p.* 200).

Note that differentiability is not imposed on $m(\cdot, \theta)$ as a function of θ, but rather for smoothed versions of $m(\cdot, \theta)$ or its functions. Assumption 3 defines the covariance function associated with the estimator and determines the order of the function ϕ (e.g., $\phi(\eta) = O(\eta)$ for a cube-root estimator). Assumptions 4 and 5 are technical conditions required to verify stochastic equicontinuity, as discussed by Van der Vaart and Wellner (1996, Section 2.11.3). Assumption 6 is a condition on the entropy of $\mathcal{M}_\eta$. The term "manageable" was coined by Pollard (1989) to describe certain classes of functions with controllable random entropy numbers. If a class of functions has a square integrable envelope function and if the square root of the (covering or bracketing) entropy number is dominated by a function λ with

$$\Lambda(x) \equiv \int_0^x \lambda(s)\,ds \tag{5}$$

continuous, $\Lambda(0) = 0$, and $\Lambda(1) < \infty$, then the class is manageable. The classes of functions discussed in this paper are all Vapnik–Chervonenkis (VC) subgraph and hence manageable. For details on entropy calculation and VC-subgraph classes, see Pollard (1990) and Van der Vaart and Wellner (1996).

When a class is manageable, the following inequalities hold. They will be used in deriving the convergence rate of the estimator and verifying stochastic equicontinuity conditions. We restate them in the following lemma taken from Kim and Pollard (1990).

LEMMA 1: *Let* $\mathcal{F}$ *be a manageable class of functions with an envelope* F, *for which* $PF^2 < \infty$. *Suppose* $0 \in \mathcal{F}$. *Then there exists a function* Λ, *not depending on* n, *such that*

(i) $\sqrt{n}E\sup_{\mathcal{F}}|P_nf - Pf| \le E[\sqrt{P_nF^2}\Lambda(\sup_{\mathcal{F}}P_nf^2/P_nF^2)] \le \Lambda(1)\sqrt{PF^2}$,

(ii) $nE\sup_{\mathcal{F}}|P_nf - Pf|^2 \le E[(P_nF^2)\Lambda^2(\sup_{\mathcal{F}}P_nf^2/P_nF^2)] \le \Lambda(1)PF^2$,

where P_n *is the empirical measure of independent random variables* $z_1, \dots, z_n$ *identically distributed as* P. *The function* Λ *is continuous and increasing with* $\Lambda(0) = 0$ *and* $\Lambda(1) < \infty$.

Let $\hat{P}_n$ be the empirical measure of the bootstrap sample $\hat{z}_1, \dots, \hat{z}_n$, which can be written as

$$\hat{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{\hat{z}_i} = n^{-1} \sum_{i=1}^{n} M_{ni} \delta_{z_i}, \tag{6}$$

where M_{ni} is the number of times that z_i is drawn from the original sample. The bootstrap empirical measure $\hat{P}_n$ can be defined directly to be the term on the far right in (6), where $\mathbf{M} \equiv (M_{n1}, \dots, M_{nn})$ has a multinomial distribution with parameters n and cell probabilities all equal to $1/n$ (and independent of $z_1, \dots, z_n$). To describe this probability structure, suppose $(M_{n1}, \dots, M_{nn})$, $n = 1, 2, \dots$, are defined on a probability space $(\mathcal{T}, \mathcal{B}, \hat{P})$. For the original sample, take z_i as the ith coordinate projection from the probability space $(\mathcal{Z}^\infty, \mathcal{A}^\infty, P^\infty)$. For the joint randomness involving both $\mathbf{M}$ and z_i, $i = 1, 2, \dots$, define the product probability space

$$(\mathcal{Z}^\infty \times \mathcal{T}, \mathcal{A}^\infty \times \mathcal{B}, P_r) \equiv (\mathcal{Z}^\infty, \mathcal{A}^\infty, P^\infty) \times (\mathcal{T}, \mathcal{B}, \hat{P}), \tag{7}$$

where $P_r \equiv P^\infty \times \hat{P}$. This framework is similar to that considered in Wellner and Zhan (1996).

For the asymptotic results to hold and to avoid measurability difficulties, we require only that both θ_n and $\hat{\theta}_n$ nearly maximize $M_n(\theta)$ and $\hat{M}_n(\theta)$ in the sense

$$M_n(\theta_n) \geq \sup_{\theta \in \Theta} M_n(\theta) - o_{P^\infty}(r_n^{-2}) \tag{8}$$

and

$$\hat{M}_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} \hat{M}_n(\theta) - o_{P_r}(r_n^{-2}), \tag{9}$$

where $r_n \to \infty$ as $n \to \infty$ and satisfies

$$r_n^4 \phi(r_n^{-1}) = n. \tag{10}$$

The term r_n turns out to be the rate of convergence of θ_n and $\hat{\theta}_n$, and is completely determined by ϕ. For example, if $\phi(\eta) = O(\eta^2)$, then the rate of convergence $r_n = n^{1/2}$; if $\phi(\eta) = O(\eta)$, then the rate of convergence $r_n = n^{1/3}$.

An important step in deriving the limit distribution of $(\theta_n, \hat{\theta}_n)$ is to consider the limit distributions of the processes

$$Z_n(t) \equiv \begin{cases} r_n^2 n^{-1} \sum_{i=1}^{n} m(z_i, \theta_0 + tr_n^{-1}) = r_n^2 P_n m(\cdot, \theta_0 + tr_n^{-1}), \\ \qquad \text{if } \theta_0 + tr_n^{-1} \in \Theta, \\ 0, \quad \text{otherwise}, \end{cases} \tag{11}$$

and

$$\hat{Z}_n(t) \equiv \begin{cases} r_n^2 n^{-1} \sum_{i=1}^{n} m(\hat{z}_i, \theta_0 + tr_n^{-1}) = r_n^2 \hat{P}_n m(\cdot, \theta_0 + tr_n^{-1}), \\ \qquad \text{if } \theta_0 + tr_n^{-1} \in \Theta, \\ 0, \quad \text{otherwise.} \end{cases} \tag{12}$$

Notice that $(Z_n, \hat{Z}_n)$ is simply a rescaled version of $(M_n, \hat{M}_n)$. Therefore, if $(\theta_n, \hat{\theta}_n)$ maximizes $(M_n, \hat{M}_n)$ componentwise, then $r_n(\theta_n - \theta_0, \hat{\theta}_n - \theta_0)$ maximizes $(Z_n, \hat{Z}_n)$ componentwise. Thus, intuitively, the limit distribution of $r_n(\theta_n - \theta_0, \hat{\theta}_n - \theta_0)$ should be the same as the maximizers of the limit process of $(Z_n, \hat{Z}_n)$ under appropriate conditions. This idea builds upon Kim and Pollard (1990), and a slight extension of their arg max continuous mapping theorem is given at the end of this section to deal with the present problem.

Following Kim and Pollard (1990), Z_n and $\hat{Z}_n$ are considered to be random elements of $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$, the space of all locally bounded real functions on $\mathbb{R}^d$, endowed with the uniform metric on compacta. The limit processes can be restricted to belong to the space $\mathbf{C}_{\max}(\mathbb{R}^d)$, the subset of continuous functions $x(\cdot) \in \mathbf{B}_{\text{loc}}(\mathbb{R}^d)$ for which (i) $x(t) \to -\infty$ as $|t| \to \infty$ and (ii) $x(t)$ achieves its maximum at a unique point in $\mathbb{R}^d$. The processes Z_n and $\hat{Z}_n$ are viewed as random elements in $\mathbf{B}_{\text{loc}}(\mathbb{R}^d)$. Then convergence in distribution of $(Z_n, \hat{Z}_n) \in \mathbf{B}_{\text{loc}}(\mathbb{R}^d) \times \mathbf{B}_{\text{loc}}(\mathbb{R}^d)$ can be characterized by (i) convergence in distribution of $\{(Z_n(t), \hat{Z}_n(t)) : t$ belongs to a finite subset of $\mathbb{R}^d\}$ and, (ii) stochastic equicontinuity of $(Z_n, \hat{Z}_n)$, which requires the process to satisfy the following: for any sequence of positive numbers δ_n converging to zero and each finite k,

$$P_r\Big\{\max\big\{\sup|Z_n(t) - Z_n(s)|, \sup|\hat{Z}_n(t) - \hat{Z}_n(s)|\big\} : \\ |s - t| < \delta_n, \max(|s|, |t|) \le k\Big\} \to 0.$$

The main results are stated in the following two theorems: Theorem 1 characterizes $Z(t)$ and $\hat{Z}(t)$, and Theorem 2 characterizes the asymptotic distributions of θ_n and $\hat{\theta}_n$.

THEOREM 1: *If Assumptions* 2–6 *hold, the process* $(Z_n(t), \hat{Z}_n(t))$ *converges in distribution to a bivariate vector of Gaussian processes* $(Z(t), \hat{Z}(t))$, *where*

$$Z(t) \equiv -\frac{1}{2} t' V t + W(t) \tag{13}$$

and

$$\hat{Z}(t) \equiv -\frac{1}{2}t'Vt + W(t) + \hat{W}(t), \tag{14}$$

where W and $\hat{W}$ are two independent Gaussian processes, both with continuous sample paths, mean zero, and covariance function H.

THEOREM 2: *Suppose that Assumptions* 1–6 *hold and that both θ_n and $\hat{\theta}_n$ converge to θ_0 in P_r probability. Furthermore, suppose that for every constant $c \geq 0$, $\phi(c\eta) \leq c^{\kappa}\phi(\eta)$ for some $\kappa < 4$. If V is positive definite and if Z and $\hat{Z}$ have nondegenerate increments, then for any bounded, uniformly continuous real function h defined on $\mathbb{R}^d \times \mathbb{R}^d$,*

$$\begin{aligned} &P_r h[r_n(\theta_n - \theta_0), r_n(\hat{\theta}_n - \theta_0)] \\ &\quad \to Eh\big[\arg\max(Z(t)), \arg\max(\hat{Z}(t))\big] \quad \textit{as } n \to \infty, \end{aligned} \tag{15}$$

and for any bounded and uniformly continuous real function h_1 defined on $\mathbb{R}^d$,

$$\begin{aligned} &\hat{P}h_1[r_n(\hat{\theta}_n - \theta_n)] \to Eh_1\big[\arg\max(\hat{Z}(t)) - \arg\max(Z(t))\big] \\ &\textit{as } n \to \infty \textit{ in } P^{\infty} \textit{ probability.} \end{aligned} \tag{16}$$

Proofs are given in the Appendix. As suggested in the statement of the theorems, the approach is to first consider the joint convergence in distribution of the bivariate vector of processes $(Z_n, \hat{Z}_n)$ and the joint convergence of $(\theta_n - \theta_0, \hat{\theta}_n - \theta_0)$ with respect to product probability measure P_r, and then to consider the limit distribution of $\hat{\theta}_n - \theta_n$ with respect to the probability measure $\hat{P}$ conditioning on the data. The process W in Theorem 1 corresponds to the (normalized) difference between the empirical and limiting objective functions, whereas the process $\hat{W}$ corresponds to the (normalized) difference between the bootstrap and empirical objective functions. When added together in (14), the resulting process corresponds to the (normalized) difference between the bootstrap and limiting objective functions. The quadratic term $-\frac{1}{2}t'Vt$ in both (13) and (14) approximates the limiting objective function near θ_0.

This method differs from previous approaches in establishing convergence in distribution of a bootstrap M estimator with $n^{1/2}$ rate of convergence and asymptotic normality, in which an asymptotic linear expansion of $\hat{\theta}_n - \theta_n$ is sought directly in terms of the bootstrap empirical measure (see Arcones and Giné (1992) for finite-dimensional M estimators and Wellner and Zhan (1996) for infinite-dimensional Z estimators).[5] It appears that no linear approximation

[5]Hahn (1995) applies the results of Arcones and Giné (1992) to show the consistency of the bootstrap for quantile regression estimators (for the linear and censored models).

exists in the present situation. Instead, the proof uses the following extension of the arg max continuous mapping theorem of Kim and Pollard (1990), proven in Huang and Wellner (1995).[6]

THEOREM 3: *Let* (U_n, V_n) *be random maps onto* $\mathbf{B}_{\text{loc}}(\mathbb{R}^d) \times \mathbf{B}_{\text{loc}}(\mathbb{R}^d)$ *and let* (s_n, t_n) *be random maps onto* $\mathbb{R}^d \times \mathbb{R}^d$ *such that*:

(i) $(U_n, V_n) \Rightarrow (U, V)$, $P\{(U, V) \in \mathbf{C}_{\max}(\mathbb{R}^d) \times \mathbf{C}_{\max}(\mathbb{R}^d)\} = 1$;

(ii) $s_n, t_n = O_p(1)$;

(iii) $U_n(s_n) \geq \sup_t U_n(t) - \alpha_n$, *and* $V_n(t_n) \geq \sup_t V_n(t) - \beta_n$, *where* $\alpha_n, \beta_n = o_p(1)$.

Then $(s_n, t_n) \xrightarrow{d} (\arg\max(U), \arg\max(V))$.

By Theorem 2, the bootstrap estimator $\hat{\theta}_n$ has the same rate of convergence as θ_n. However, the limit distribution of $r_n(\hat{\theta}_n - \theta_n)$ can be different from that of $r_n(\theta_n - \theta_0)$. Consider the cube-root case ($r_n = n^{1/3}$), which occurs when the envelope function satisfies

$$\phi(\alpha) = PM_\alpha^2 = O(\alpha). \tag{17}$$

In this case, the limit distribution of $n^{1/3}(\theta_n - \theta_0)$ is $\tau \equiv \arg\max(-t'Vt/2 + W(t))$, whereas the limit distribution of $n^{1/3}(\hat{\theta}_n - \theta_n)$ is $\hat{\tau} - \tau$, where $\hat{\tau} \equiv \arg\max(-t'Vt/2 + W(t) + \hat{W}(t))$.

Theorem 2 also provides an alternative argument for bootstrap consistency in the (regular) $\sqrt{n}$-consistent case. If the envelope function satisfies

$$\phi(\alpha) = PM_\alpha^2 = O(\alpha^2), \tag{18}$$

it is reasonable to assume in Assumption 3 that there is a positive definite matrix Σ such that $H(s, t) = s'\Sigma t$. Take $W(t)$ and $\hat{W}(t)$ to be $t'\xi$ and $t'\hat{\xi}$, respectively, where ξ and $\hat{\xi}$ are i.i.d. $N(0, \Sigma)$. Then $n^{1/2}(\theta_n - \theta_0)$ is asymptotically distributed as $\arg\max(-t'Vt/2 + t'\xi) = V^{-1}\xi$, and $n^{1/2}(\hat{\theta}_n - \theta_n)$ is distributed as

$$\begin{aligned} &\arg\max(-t'Vt/2 + t'\xi + t'\hat{\xi}) - \arg\max(-t'Vt/2 + t'\xi) \\ &\quad = V^{-1}(\xi + \hat{\xi}) - V^{-1}\xi = V^{-1}\hat{\xi}. \end{aligned}$$

Therefore, in this (regular) asymptotic situation when the rate of convergence is $n^{1/2}$ and the limit distribution is normal, the nonparametric bootstrap estimator is consistent. The key to bootstrap consistency is linearity of $W(t)$.

[6]See also Van der Vaart and Wellner (1996, Chapter 3.2) for a discussion and applications of the arg max continuous mapping theorem.

3. THE MAXIMUM SCORE ESTIMATOR

3.1. *Inconsistency of the Bootstrap*

Consider the latent-variable binary choice model from (1), with a scale normalization incorporated as

$$y = \{\tilde{x}'\beta_0 + x_d - \epsilon \geq 0\}. \tag{19}$$

The random vector x is split into two components, $\tilde{x} \in \mathbb{R}^{d-1}$ and a single continuous covariate x_d; that is, $x \equiv (\tilde{x}, x_d)$ and $z \equiv (y, x)$. The parameter vector β_0 is an element of $\mathbb{R}^{d-1}$. To avoid additional notation, we have fixed the coefficient of x_d to be equal to $+1$. In practice, maximization of the maximum score objective function would be subject to the restriction that the coefficient on x_d is equal to $+1$ or -1, but the rate of convergence of this coefficient estimate (to $+1$ or -1) is faster than that of the other coefficient estimates. As such, we assume without loss of generality that the coefficient on x_d is equal to $+1$.[7,8]

The maximum score objective function, normalized to have $m(\cdot, \beta_0) \equiv 0$, is

$$M_n(\beta) \equiv n^{-1} \sum_{i=1}^{n} (2y_i - 1)(\{\tilde{x}_i'\beta + x_{d,i} \geq 0\} - \{\tilde{x}_i'\beta_0 + x_{d,i} \geq 0\}). \tag{20}$$

Note that $m(z, \beta) = (2y - 1)(\{\tilde{x}'\beta + x_d \geq 0\} - \{\tilde{x}'\beta_0 + x_d \geq 0\})$. The maximum score estimator, denoted β_n, is defined to be any vector that maximizes $M_n(\beta)$. The bootstrap estimator, denoted $\hat{\beta}_n$, is defined to be any vector that maximizes $\hat{M}_n(\beta)$, the bootstrap counterpart to the objective function in (20). The limiting objective function $M(\beta)$ is

$$M(\beta) = E\big[(2y - 1)(\{\tilde{x}'\beta + x_d \geq 0\} - \{\tilde{x}'\beta_0 + x_d \geq 0\})\big]. \tag{21}$$

Some additional notation is required to describe the distributions of the random variables x and ϵ.[9] Let $F(\epsilon|\tilde{x}, x_d)$ and $f(\epsilon|\tilde{x}, x_d)$ denote the distribution and density functions, respectively, of ϵ conditional on $\tilde{x}$ and x_d. Let $g(x_d|\tilde{x})$

[7]If the coefficient on x_d is -1, the model in (19) is satisfied with $-x_d$ replacing x_d. Another benefit of writing the model as (19) is that this form also can be interpreted as a "current-status" model in a duration-data framework, in which one observes whether the spell has ended ($y = 1$) or not ($y = 0$) at a given duration x_d.

[8]Kim and Pollard (1990) analyze the maximum score estimator under the alternative normalization that the $(d \times 1)$ coefficient vector has length 1. This approach turns out to be more complicated than the current approach since it involves the evaluation of surface integrals on the unit sphere. The resulting expressions that we derive for the asymptotic distribution of the maximum score estimator are simpler than those in Kim and Pollard (1990), but the form of the distribution still makes asymptotic inference infeasible.

[9]In what follows, it is always assumed that the quantities are defined only when the necessary derivative exists.

denote the density function of x_d conditional on $\tilde{x}$ and let $g'(x_d|\tilde{x})$ denote its derivative (with respect to the x_d argument).

To show inconsistency of the bootstrap for the maximum score estimator, the conditions of Theorem 2 need to be verified. The assumptions for consistency of β_n and $\hat{\beta}_n$ are similar to those originally made by Manski (1985). A few additional technical assumptions (a subset of those in Horowitz (1992)) are required for the asymptotic distributional theory. To allow for an easy comparison, we follow Horowitz (1992) as closely as possible in stating the following assumptions.

ASSUMPTION MS1: *Suppose* $\{(x_i, \epsilon_i)\}_{i=1}^n$ *is an i.i.d. random sample of* (x, ϵ). *The observed data are* $\{(y_i, x_i)\}_{i=1}^n$, *where* y_i *is generated according to the model in* (19).

ASSUMPTION MS2: (a) *The support of the distribution of* x *is not contained in any proper linear subspace of* $\mathbb{R}^d$; (b) $0 < \Pr(y = 1|x) < 1$ *for almost every* x; (c) *the distribution of* x_d *conditional on* $\tilde{x}$ *has everywhere positive density with respect to Lebesgue measure for almost every* $\tilde{x}$.

ASSUMPTION MS3: $\text{Median}(\epsilon|x) = 0$ (*uniquely*) *for almost every* x.

ASSUMPTION MS4: *Suppose* $\beta_0 \in \mathcal{B}$ *is a compact subset of* $\mathbb{R}^{d-1}$.

ASSUMPTION MS5: *The components of* $\tilde{x}$ *and* $\tilde{x}\tilde{x}'$ *have finite first absolute moments.*

ASSUMPTION MS6: *The function* $g'(x_d|\tilde{x})$ *exists and, for some* $M > 0$, $|g'(x_d|\tilde{x})| < M$ *and* $|g(x_d|\tilde{x})| < M$ *for all* x_d *and almost every* $\tilde{x}$.

ASSUMPTION MS7: *For all* ϵ *in a neighborhood around* 0, *all* x_d *in a neighborhood around* $-\tilde{x}'\beta_0$, *almost every* $\tilde{x}$, *and some* $M > 0$, *the function* $f(\epsilon|\tilde{x}, x_d)$ *exists and* $f(\epsilon|\tilde{x}, x_d) < M$.

ASSUMPTION MS8: *For all* ϵ *in a neighborhood around* 0, *all* x_d *in a neighborhood around* $-\tilde{x}'\beta_0$, *almost every* $\tilde{x}$, *and some* $M > 0$, *the function* $\partial F(\epsilon|\tilde{x}, x_d)/\partial x_d$ *exists and* $|\partial F(\epsilon|\tilde{x}, x_d)/\partial x_d| < M$.

ASSUMPTION MS9: *Suppose* β_0 *is an interior point of* $\mathcal{B}$.

ASSUMPTION MS10: *The matrix*

$$V_{\text{ms}} \equiv E\left[2f(0|\tilde{x}, -\tilde{x}'\beta_0)g(-\tilde{x}'\beta_0|\tilde{x})\tilde{x}\tilde{x}'\right] \tag{22}$$

is positive definite.

Assumptions MS1–MS4 (analogous to Assumptions 1–4 of Horowitz (1992)) are used to show consistency of β_n and $\hat{\beta}_n$. Assumption MS2(b) is satisfied if ϵ has support everywhere along the real line (for almost every x). Assumption MS3 is the conditional median restriction on the error disturbance. Assumptions MS5–MS10 are similar to Assumptions 5 and 8–11 of Horowitz (1992). The main difference is that Horowitz (1992) allows for higher-order derivatives (of g and F) to exist, which leads to an improved rate of convergence for the smoothed estimator.[10] The V_{ms} matrix in Assumption MS10 turns out to be the negative of the second derivative of $M(\beta)$ evaluated at β_0.

The following theorem provides the limit distributions for both the maximum score estimator and its bootstrap counterpart. These limit distributions are different, which implies inconsistency of the bootstrap.

THEOREM 4: *If Assumptions* MS1–MS10 *hold*,

$$\sqrt[3]{n}(\beta_n - \beta_0) \xrightarrow{d} \arg\max(Z(t)) \tag{23}$$

and (*conditioning on the data*)

$$\sqrt[3]{n}(\hat{\beta}_n - \beta_n) \xrightarrow{d} \arg\max(\hat{Z}(t)) - \arg\max(Z(t)) \quad \textit{in } P^{\infty} \textit{ probability}, \tag{24}$$

where $Z(t) \equiv -\frac{1}{2}t'V_{\text{ms}}t + W(t)$, $\hat{Z}(t) \equiv -\frac{1}{2}t'V_{\text{ms}}t + W(t) + \hat{W}(t)$, *and* W *and* $\hat{W}$ *are independent Gaussian processes with mean zero and covariance function* H *given by*

$$H(u, v) = E\big[\{\text{sign}(\tilde{x}'u) = \text{sign}(\tilde{x}'v)\} \min(|\tilde{x}'u|, |\tilde{x}'v|)\, g(-\tilde{x}'\beta_0|\tilde{x})\big]. \tag{25}$$

The proof, which is left for the Appendix, entails verifying the conditions of Theorem 2.

3.2. *One-Covariate Case*

Consider the simplest case of a single covariate x (in addition to the intercept term), so that the model in (19) becomes $y = \{\beta_0 + x - \epsilon \geq 0\}$. Since $\tilde{x}$ is constant, the notation can be simplified by dropping the $\tilde{x}$ arguments from the distribution and density functions defined above. As such, let $f(\epsilon|x)$ denote the density function of ϵ conditional on x and let $g(x)$ denote the density function of x. The expressions for V_{ms} and H simplify considerably, with $V_{\text{ms}} = 2f(0|$

[10]Another difference is that Horowitz (1992) defines F to be conditional on $\tilde{x}$ and $\tilde{x}'\beta_0 + x_d$. The condition in Assumption 9 of Horowitz (1992) essentially entails existence and boundedness of two derivatives, which we have explicitly handled separately in Assumptions MS7 and MS8.

$-\beta_0)g(-\beta_0)$ and $H(u,v)=\{\text{sign}(u)=\text{sign}(v)\}g(-\beta_0)\min(|u|,|v|)$. This covariance function H characterizes two-sided Brownian motion that originates at zero. It follows that Theorem 4 holds with

$$Z(t)=-ct^2+B(t) \quad\text{and}\quad \hat{Z}(t)=-ct^2+B(t)+\hat{B}(t), \tag{26}$$

where $c\equiv f(0|-\beta_0)\sqrt{g(-\beta_0)}$, and $B(t)$ and $\hat{B}(t)$ are independent two-sided Brownian motions that originate at zero.

Let Z_c^* (for $c>0$) denote the random variable given by

$$Z_c^*\equiv\arg\max(-ct^2+B(t)) \tag{27}$$

and denote $Z^*\equiv Z_1^*$. The distribution of Z^* was derived by Groeneboom (1989) and computed by Groeneboom and Wellner (2001). Let $\hat{Z}_c^*$ (for $c>0$) denote the random variable given by

$$\hat{Z}_c^*\equiv\arg\max(-ct^2+B(t))-\arg\max\bigl(-ct^2+B(t)+\hat{B}(t)\bigr) \tag{28}$$

and denote $\hat{Z}^*\equiv\hat{Z}_1^*$. Since no analytical expression is yet available for $\hat{Z}^*$, we used 10 million simulated draws to approximate the distribution of $\hat{Z}^*$.[11] Figure 1 compares the density functions of Z^* and $\hat{Z}^*$.[12] As compared to the distribution of Z^* (which has a shape similar to the normal distribution),[13] the density function of $\hat{Z}^*$ has a pronounced central peak and long tails.

4. SINGLE-PARAMETER ESTIMATION PROBLEMS

Like the single-covariate case for the maximum score estimator, there are a variety of other single-parameter estimators that are $\sqrt[3]{n}$ consistent with a limiting distribution that (up to a scale factor) is equal to Z^*:

$$\sqrt[3]{n}k(\theta_n-\theta_0)\xrightarrow{d}Z^* \tag{29}$$

for some $k>0$. Examples include the nonparametric maximum likelihood estimator (NPMLE) of a distribution function in a binary-choice model (or current-status model) and the isotonic regression estimator. Kim and Pollard

[11]Each draw was conducted by evaluating the two Brownian motions on a grid with points 0.0001 apart on the interval $[-4.0000,+4.0000]$.

[12]Since Brownian scaling implies that $Z_c^*=c^{-2/3}Z^*$ and $\hat{Z}_c^*=c^{-2/3}\hat{Z}^*$, the case $c=1$ is considered without loss of generality. For $c\neq 1$, both distributions would be scaled by the same constant.

[13]In fact, Dykstra and Carolan (1999) suggest the approximation $N(0,(0.52)^2)$ for Z^*. Such an approximation turns out to be fairly accurate, as evidenced by the results of Groeneboom and Wellner (2001, Table 2).

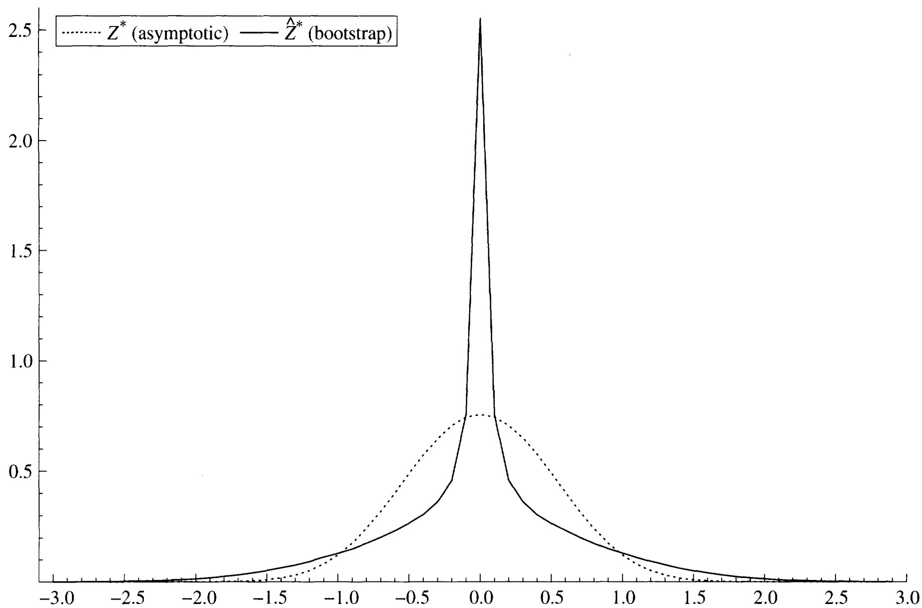

FIGURE 1.—Densities of Z^* and $\hat{Z}^*$.

(1990) and Groeneboom and Wellner (2001) consider several additional examples. Based on the results of Section 2, one would expect that the bootstrap limiting distribution in this class of estimators would be

$$\sqrt[3]{n}k(\hat{\theta}_n - \theta_n) \xrightarrow{d} \hat{Z}^*, \tag{30}$$

where k is the same constant as in the limiting distribution for θ_n.

Since the distribution of Z^* is known, asymptotic inference using the limiting distribution in (29) is feasible. Unfortunately, consistent estimation of the scale factor k requires additional nonparametric estimation. For example, the scale factor k for the NPMLE contains three unknown quantities that must be nonparametrically estimated (see Theorem 5). The bootstrap provides an alternative method for asymptotic inference that avoids additional nonparametric estimation and does not depend on smoothing-parameter choices.

The bootstrap approach entails first forming a bootstrap percentile confidence interval in the usual way and then "correcting" this inconsistent interval by using information about Z^* and $\hat{Z}^*$. Table I is provided as a reference to aid in asymptotic inference. The table contains the quantiles of Z^* (computed using the same computer program as Groeneboom and Wellner (2001)) and $\hat{Z}^*$ (computed from the empirical quantiles from the 10 million simulated draws of $\hat{Z}^*$). Since the $\hat{Z}^*$ quantiles are based on simulations, 99% confidence intervals are also presented in the last two columns.[14] Note that only quantiles above 0.50 are reported, as both Z^* and $\hat{Z}^*$ are symmetric distributions. To illustrate how Table I can be used to "correct" a bootstrap confidence interval, suppose that a 90% interval is constructed using a series of bootstrap estimators. If the 90% bootstrap interval is written as

$$(\theta_n - a, \theta_n + b)$$

(where $a = b$ corresponds to a symmetric bootstrap interval), an asymptotically valid 90% confidence interval (based on (29) and (30)) would be

$$\left(\theta_n - \frac{0.8451}{1.0932}a, \theta_n + \frac{0.8451}{1.0932}b\right) \approx (\theta_n - 0.773a, \theta_n + 0.773b).$$

The values 0.8451 and 1.0932 are the 0.95 quantiles of Z^* and $\hat{Z}^*$, respectively, from Table I.

For the proposed method to be applicable, the limiting distributions in (29) and (30) must hold for a given estimator. As an example, the NPMLE for a binary-choice distribution function is considered next. Although the NPMLE is the only example considered here, it should be possible to verify (29) and (30) for other single-parameter cube-root estimators, as there is no reason to expect otherwise for the estimators considered in both Kim and Pollard (1990) and Groeneboom and Wellner (2001).

EXAMPLE—NPMLE of a Binary-Choice Distribution Function: Suppose $(x_1, y_1), \ldots, (x_n, y_n)$ is an i.i.d. sample, where G denotes the distribution of the random variable x and $\Pr(y = 1|x) = F(x)$ for an increasing function $F: \mathcal{R} \to [0, 1]$. The NPMLE estimator F_n for F maximizes the likelihood function

$$\sum_{i=1}^{n} \big(y_i \log \tilde{F}(x_i) + (1 - y_i)\log(1 - \tilde{F}(x_i))\big)$$

over $\tilde{F}$. Let $x_{(1)}, \ldots, x_{(n)}$ denote the x values in ascending order. Since the maximum likelihood estimator is not unique, assume that F_n is constant on the intervals $[x_{(i-1)}, x_{(i)})$ (that is, F_n can jump only at x_i values). It is well known that the NPMLE estimates $F_n(x_i)$ solve the "isotonic regression" problem, minimizing

$$\sum_{i=1}^{n} (y_i - f_i)^2$$

[14]The confidence intervals are formed using the distribution-free method; see Van der Vaart (1998, p. 309).

TABLE I

QUANTILES OF Z^* AND $\hat{Z}^*$

Quantile	Z^*	$\hat{Z}^*$	99% Interval	
0.510	0.0132	0.0000	0.0000	0.0000
0.520	0.0264	0.0001	0.0001	0.0001
0.530	0.0396	0.0004	0.0004	0.0004
0.540	0.0528	0.0010	0.0010	0.0010
0.550	0.0661	0.0020	0.0020	0.0020
0.560	0.0794	0.0034	0.0034	0.0035
0.570	0.0928	0.0055	0.0054	0.0055
0.580	0.1062	0.0082	0.0081	0.0082
0.590	0.1196	0.0117	0.0116	0.0117
0.600	0.1332	0.0158	0.0158	0.0159
0.610	0.1468	0.0211	0.0210	0.0212
0.620	0.1606	0.0272	0.0271	0.0273
0.630	0.1744	0.0344	0.0342	0.0345
0.640	0.1883	0.0426	0.0424	0.0428
0.650	0.2024	0.0521	0.0520	0.0523
0.660	0.2166	0.0630	0.0627	0.0632
0.670	0.2310	0.0749	0.0747	0.0751
0.680	0.2455	0.0885	0.0883	0.0888
0.690	0.2602	0.1031	0.1028	0.1034
0.700	0.2752	0.1191	0.1188	0.1195
0.710	0.2903	0.1361	0.1357	0.1365
0.720	0.3056	0.1542	0.1538	0.1546
0.730	0.3212	0.1742	0.1738	0.1747
0.740	0.3371	0.1962	0.1956	0.1965
0.750	0.3533	0.2188	0.2183	0.2192
0.760	0.3698	0.2422	0.2417	0.2427
0.770	0.3867	0.2678	0.2672	0.2684
0.780	0.4040	0.2947	0.2942	0.2953
0.790	0.4217	0.3229	0.3224	0.3235
0.800	0.4398	0.3527	0.3521	0.3533
0.810	0.4585	0.3842	0.3835	0.3849
0.820	0.4778	0.4173	0.4166	0.4179
0.830	0.4977	0.4519	0.4512	0.4527
0.840	0.5184	0.4879	0.4872	0.4886
0.850	0.5399	0.5259	0.5251	0.5266
0.860	0.5623	0.5659	0.5651	0.5667
0.870	0.5857	0.6083	0.6075	0.6091
0.880	0.6104	0.6533	0.6525	0.6540
0.890	0.6365	0.7009	0.7001	0.7018
0.900	0.6642	0.7528	0.7519	0.7537
0.910	0.6940	0.8086	0.8077	0.8095
0.920	0.7262	0.8695	0.8686	0.8706
0.925	0.7434	0.9021	0.9013	0.9031
0.930	0.7615	0.9364	0.9354	0.9375
0.935	0.7805	0.9727	0.9717	0.9737

Continues

TABLE I—*Continued*

Quantile	Z^*	$\hat{Z}^*$	99% Interval	
0.940	0.8007	1.0107	1.0097	1.0117
0.945	0.8221	1.0505	1.0495	1.0517
0.950	0.8451	1.0932	1.0921	1.0943
0.955	0.8699	1.1396	1.1385	1.1407
0.960	0.8969	1.1899	1.1888	1.1911
0.965	0.9267	1.2469	1.2457	1.2482
0.970	0.9601	1.3106	1.3093	1.3119
0.975	0.9982	1.3822	1.3808	1.3836
0.980	1.0430	1.4663	1.4649	1.4678
0.985	1.0982	1.5703	1.5687	1.5719
0.990	1.1715	1.7068	1.7050	1.7085
0.995	1.2867	1.9214	1.9189	1.9237

subject to $f_{(1)} \leq f_{(2)} \leq \cdots \leq f_{(n)}$. The NPMLE estimate $F_n(x_{(i)})$ is also equal to the slope in the interval $(i-1, i)$ of the greatest convex minorant of the points $\{(i, \sum_{j\leq i} y_{(j)})\}_{i=1}^n$. The arguments leading to the limit distributions for the estimator $F_n(x_0)$ and its bootstrap counterpart $\hat{F}_n(x_0)$ are somewhat more complicated than for the maximum score estimator. The estimator $F_n(x_0)$ cannot be expressed simply as an M estimator, meaning the results of Section 2 cannot be directly applied. Viewing F_n as the slope of the greatest convex minorant, however, allows an indirect application of the results. The main result is given by the following theorem.

THEOREM 5: *Suppose $(x_1, y_1), \ldots, (x_n, y_n)$ is an i.i.d. sample, where G denotes the distribution of the random variable x and* $\Pr(y=1|x) = F(x)$ *for an increasing function $F : \mathcal{R} \to [0,1]$. For a fixed $x_0 > 0$, if F and G are differentiable at x_0 with $f(x_0) \equiv F'(x_0) > 0$ and $g(x_0) \equiv G'(x_0) > 0$, then*

$$\sqrt[3]{n}\left(\frac{4F(x_0)(1-F(x_0))f(x_0)}{g(x_0)}\right)^{-1/3}(F_n(x_0) - F(x_0)) \xrightarrow{d} Z^* \tag{31}$$

and (conditioning on the data)

$$\sqrt[3]{n}\left(\frac{4F(x_0)(1-F(x_0))f(x_0)}{g(x_0)}\right)^{-1/3} \times (\hat{F}_n(x_0) - F_n(x_0)) \xrightarrow{d} \hat{Z}^* \quad \textit{in } P^\infty \textit{ probability}. \tag{32}$$

Groeneboom and Wellner (1992, Theorem 5.1) provide a proof of (31); see also Van der Vaart and Wellner (1996, Example 3.2.15). Asymptotic inference using (31) requires consistent estimation of $f(x_0)$ and $g(x_0)$, which can be obtained through nonparametric estimation.[15] Alternatively, the adjusted-bootstrap interval described above can be used to avoid nonparametric estimation.

5. CONCLUSION

This paper has shown inconsistency of the bootstrap for a class of cube-root estimators, including the maximum score estimator. Although not pursued here, the results could be applied to other closely related estimators, including the maximum score estimator for the multinomial choice model (Manski (1975)), the maximum score estimator for the binary-choice fixed-effects model (Manski (1987)), and rank estimators of the generalized fixed-effects model (Abrevaya (2000)).

Dept. of Economics, Purdue University, 403 West State Street, West Lafayette, IN 47907, U.S.A.; abrevaya@purdue.edu

and

Dept. of Statistics and Actuarial Science, University of Iowa, 241 Schaeffer Hall, Iowa City, IA 52242, U.S.A.; jian@stat.uiowa.edu.

Manuscript received January, 2004; final revision received February, 2005.

APPENDIX

PROOF OF THEOREM 1: Denote

$$W_n(t) \equiv \begin{cases} r_n^2(P_n - P)m(\cdot, \theta_0 + tr_n^{-1}), & \text{if } \theta_0 + tr_n^{-1} \in \Theta, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\hat{W}_n(t) \equiv \begin{cases} r_n^2(\hat{P}_n - P_n)m(\cdot, \theta_0 + tr_n^{-1}), & \text{if } \theta_0 + tr_n^{-1} \in \Theta, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$r_n^2 P_n m(\cdot, \theta_0 + tr_n^{-1}) = W_n(t) + r_n^2 Pm(\cdot, \theta_0 + tr_n^{-1}) \tag{33}$$

and

$$r_n^2 \hat{P}_n m(\cdot, \theta_0 + tr_n^{-1}) = W_n(t) + \hat{W}_n(t) + r_n^2 Pm(\cdot, \theta_0 + tr_n^{-1}). \tag{34}$$

[15]Hausman, Abrevaya and Scott-Morton (1998) use this approach in a two-stage estimation procedure where the second stage involves the binary-choice NPMLE estimator.

Based on (33) and (34), it suffices to prove the following two parts: (i) both W_n and $\hat{W}_n$ are stochastically equicontinuous and (ii) W_n and $\hat{W}_n$ are asymptotically independent and their finite limit distributions are the same as those corresponding to $\hat{W}$ and W.

Part (i): Stochastic equicontinuity of W_n is proved by Van der Vaart and Wellner (1996, Theorem 3.2.10), so we only consider $\hat{W}_n$. The argument resembles that for W_n used by Kim and Pollard (1990). Let $\delta_n \to 0$ and let K be a finite positive constant. Define

$$\mathcal{G}(n) \equiv \big\{m(\cdot, \theta_0 + t_1 r_n^{-1}) - m(\cdot, \theta_0 + t_2 r_n^{-1}) :$$
$$\max(|t_1|, |t_2|) \le K \text{ and } |t_1 - t_1| \le \delta_n\big\}.$$

An envelope function of this class can be taken to be $G_n = 2M_{\eta(n)}$, where $\eta(n) \equiv K r_n^{-1}$. By the definition of stochastic equicontinuity and Markov's inequality, it suffices to show that

$$r_n^2 P_r \sup_{\mathcal{G}(n)} |(\hat{P}_n - P_n)g| \to 0 \quad \text{almost surely.}$$

Define $\hat{U}_n \equiv n^{-1} r_n^4 \hat{P}_n G_n^2$ and $\hat{V}_n \equiv \sup_{\mathcal{G}(n)} \hat{P}_n g^2$. Let $\epsilon > 0$ be any small positive number. The first maximal inequality implies

$$\begin{aligned}
&r_n^2 \hat{P} \sup_{\mathcal{G}(n)} |(\hat{P}_n - P_n)g| \\
&= \sqrt{n} \hat{P} \sup_{\mathcal{G}(n)} |(\hat{P}_n - P_n) n^{-1/2} r_n^2 g| \\
&\le \hat{P} \sqrt{\hat{U}_n} \Lambda(n^{-1} r_n^4 \hat{V}_n / \hat{U}_n) \\
&= \hat{P} \sqrt{\hat{U}_n} \Lambda(n^{-1} r_n^4 \hat{V}_n / \hat{U}_n)\{\hat{U}_n \le \epsilon\} + \hat{P} \sqrt{\hat{U}_n} \Lambda(n^{-1} r_n^4 \hat{V}_n / \hat{U}_n)\{\hat{U}_n > \epsilon\} \\
&\le \sqrt{\epsilon} \Lambda(1) + \sqrt{\hat{P} \hat{U}_n} \sqrt{\hat{P} \Lambda^2(n^{-1} r_n^4 \hat{V}_n / \hat{U}_n)\{\hat{U}_n > \epsilon\}},
\end{aligned}$$

where the last inequality is obtained by using $n^{-1} r_n^4 \hat{V}_n \le \hat{U}_n$. Since $\hat{P}\hat{U}_n = n^{-1} r_n^4 P_n G_n^2$ and $n^{-1} r_n^4 P_n G_n^2 = O(1)$, it suffices to show that $P_r \hat{V}_n = o(n r_n^{-4})$. Let $\eta > 0$ be an arbitrary positive constant. We have

$$\begin{aligned}
\hat{P}\hat{V}_n &= \hat{P} \sup_{\mathcal{G}(n)} \hat{P}_n g^2 \\
&\le \hat{P} \sup_{\mathcal{G}(n)} \hat{P}_n g^2 \{G_n > \eta n r_n^{-2}\} + \hat{P} \sup_{\mathcal{G}(n)} \hat{P}_n g^2 \{G_n \le \eta n r_n^{-2}\}
\end{aligned}$$

$$
\begin{aligned}
&\leq \hat{P}\hat{P}_n G_n^2\{G_n > \eta n r_n^{-2}\} + \hat{P}\sup_{\mathcal{G}(n)}|(\hat{P}_n - P_n)g^2\{G_n \leq \eta n r_n^{-2}\}| \\
&\quad + \hat{P}\sup_{\mathcal{G}(n)}|(P_n - P)g^2\{G_n \leq \eta n r_n^{-2}\}| + \hat{P}\sup_{\mathcal{G}(n)} Pg^2\{G_n \leq \eta n r_n^{-2}\} \\
&= \hat{P}G_n^2\{G_n > \eta n r_n^{-2}\} + \hat{P}\sup_{\mathcal{G}(n)}|(\hat{P}_n - P_n)g^2\{G_n \leq \eta n r_n^{-2}\}| \\
&\quad + \sup_{\mathcal{G}(n)}|(P_n - P)g^2\{G_n \leq \eta n r_n^{-2}\}| + \sup_{\mathcal{G}(n)} Pg^2\{G_n \leq \eta n r_n^{-2}\}.
\end{aligned}
$$

Using the maximal inequality conditioning on $z_1, z_2, \ldots,$ the second term is bounded as

$$
\begin{aligned}
\hat{P}\sup_{\mathcal{G}(n)}|(\hat{P}_n - P_n)g^2\{G_n \leq \eta n r_n^{-2}\}| &\leq C_1 n^{-1/2}\sqrt{P_n G_n^4\{G_n \leq \eta n r_n^{-2}\}} \\
&\leq C_1 \eta n^{1/2} r_n^{-2}\sqrt{P_n G_n^2}.
\end{aligned}
$$

Similarly, the third term is bounded as

$$
\begin{aligned}
\sup_{\mathcal{G}(n)}|(P_n - P)g^2\{G_n \leq \eta n r_n^{-2}\}| &\leq C_2 n^{-1/2}\sqrt{PG_n^4\{G_n \leq \eta n r_n^{-2}\}} \\
&\leq C_2 \eta n^{1/2} r_n^{-2}\sqrt{PG_n^2}.
\end{aligned}
$$

The conditions of the theorem also imply that (see, e.g., Van der Vaart and Wellner (1996, p. 293))

$$
\begin{aligned}
&PG_n^2 = O(n r_n^{-4}), \quad PG_n^2\{G_n > \eta n r_n^{-2}\} = o(n r_n^{-4}), \quad \text{and} \\
&\sup_{\mathcal{G}(n)} Pg^2 = o(n r_n^{-4}).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
P_r\hat{V}_n &= P^\infty \hat{P}\hat{V}_n \\
&\leq PG_n^2\{G_n > \eta n r_n^{-2}\} + C_1 \eta n^{1/2} r_n^{-2} P[P_n G_n^2]^{1/2} \\
&\quad + C_2 \eta n^{1/2} r_n^{-2} P[PG_n^2]^{1/2} + \sup_{\mathcal{G}(n)} Pg^2 \\
&\leq (C_1 + C_2)\eta n^{1/2} r_n^{-2}[PG_n^2]^{1/2} + o(n r_n^{-4}) \\
&\leq O(1)\eta n r_n^{-4} + o(n r_n^{-4}).
\end{aligned}
$$

Since η is arbitrarily small, the proof of part (i) is finished.

Part (ii): For asymptotic independence of $\hat{W}_n$ and W_n, we need to show that $(\hat{W}_n(t_1), \ldots, \hat{W}_n(t_k))$ and $(W_n(s_1), \ldots, W_n(s_\ell))$ are independent, where k and ℓ are arbitrary positive finite integers and $(t_1, \ldots, t_k)$ and $(s_1, \ldots, s_\ell)$ may overlap. For simplicity, we show that $\hat{W}_n(t)$ and $W_n(t)$ are asymptotically independent for any fixed t. Denote the characteristic function of $(\hat{W}_n(t), W_n(t))$ by

$$\phi_n(s_1, s_2) \equiv P_r\big[\exp\big(i(s_1\hat{W}_n(t) + s_2 W_n(t))\big)\big].$$

Denote

$$\phi(s) \equiv E\big[\exp(isW(t))\big] = E\big[\exp(is\hat{W}(t))\big].$$

Van der Vaart and Wellner (1996, Theorem 3.2.10) show that the finite-dimensional distributions of W_n converge to the corresponding finite-dimensional distributions of W. Using the Lindeberg–Feller central limit theorem, it can be shown that the (conditional) finite-dimensional distributions of $\hat{W}_n$ (conditioning on $z_1, z_2, \ldots$) converge to the corresponding finite-dimensional distributions of $\hat{W}$ almost surely. Thus, we have

$$\hat{P}\big[\exp(is\hat{W}_n(t))\big] \xrightarrow{\text{a.s.}} \phi(s). \tag{35}$$

Therefore,

$$\begin{aligned}
&\phi_n(s_1, s_2) - \phi(s_1)\phi(s_2) \\
&\quad = P^\infty\big\{\exp(is_2 W_n(t))\hat{P}\big[\exp(is_1\hat{W}_n(t))\big]\big\} - \phi(s_1)\phi(s_2) \\
&\quad = P^\infty\big\{\exp(is_2 W_n(t))\big\{\hat{P}\big[\exp(is_1\hat{W}_n(t))\big] - \phi(s_1)\big\}\big\} \\
&\qquad + \phi(s_1)\big[P^\infty\big\{\exp(is_2 W_n(t))\big\} - \phi(s_2)\big].
\end{aligned}$$

The first term converges to zero by (35) and the bounded convergence theorem. The second term converges to zero because $W_n(t)$ converges in distribution. The proof of part (ii) is thus complete. *Q.E.D.*

Before proving Theorem 2, we provide two lemmas. The first states relationships among the probability measures P_r, P^∞, and $\hat{P}$ in terms of $o_p(\cdot)$ and $O_p(\cdot)$. The second gives the rate of convergence for $\hat{\theta}_n$ (the same rate as proven by Kim and Pollard (1990) for θ_n), which is needed in applying the arg max continuous mapping theorem.

LEMMA 2: (i) *If Δ_n is defined only on the probability space $(\mathcal{Z}^\infty, \mathcal{A}^\infty, P^\infty)$ and $\Delta_n = o_{P^\infty}(1)$ $(O_{P^\infty}(1))$, then $\Delta_n = o_{P_r}(1)$ $(O_{P_r}(1))$*; (ii) *If $\Delta_n = o_{P_r}(1)$ $(O_{P_r}(1))$, then $\Delta_n = o_{\hat{P}}(1)$ $(O_{\hat{P}}(1))$ in P^∞ probability. (Δ_n is said to be "$o_{\hat{P}}(1)$ in P^∞ probability" if, for any $\epsilon > 0$ and $\eta > 0$, $P^\infty\{\hat{P}\{|\Delta_n| > \epsilon\} > \eta\} \to 0$ as $n \to \infty$.)*

These results are proven by Wellner and Zhan (1996, Theorem 3.1). (The converse of (ii) is also true, but is not needed for the proof of Theorem 2.)

LEMMA 3: *Suppose that both θ_n and $\hat{\theta}_n$ converge to θ_0 in P_r probability. Furthermore, suppose that for every constant $c \geq 0$, $\phi(c\eta) \leq c^{\kappa}\phi(\eta)$ for some $\kappa < 4$ and that Assumptions* 1–6 *hold. Then*

$$\hat{\theta}_n = \theta_0 + O_{P_r}(r_n^{-1}).$$

PROOF: By the definition of $\hat{\theta}_n$,

$$\hat{P}_n m(\cdot, \hat{\theta}_n) \geq \hat{P}_n m(\cdot, \theta_0) - o(r_n^{-2}).$$

Thus,

$$\begin{aligned} o(r_n^{-2}) &\leq \hat{P}_n[m(\cdot, \theta_0) - m(\cdot, \hat{\theta}_n)] \\ &= (\hat{P}_n - P_n)[m(\cdot, \theta_0) - m(\cdot, \hat{\theta}_n)] \\ &\quad + (P_n - P)[m(\cdot, \theta_0) - m(\cdot, \hat{\theta}_n)] + P[m(\cdot, \theta_0) - m(\cdot, \hat{\theta}_n)]. \end{aligned}$$

For the third term on the right-hand side of the last equation, choose $\epsilon > 0$ such that

$$P[m(\cdot, \theta) - m(\cdot, \theta_0)] \leq -3\epsilon|\theta - \theta_0|^2 \tag{36}$$

in a neighborhood of θ_0. For the second term, an argument similar to that of Lemma 4.1 of Kim and Pollard (1990) shows that there exist random variables R_n of order $O_{P^\infty}(1)$ (and thus of order $O_{P_r}(1)$ by Lemma 2(i)) such that

$$\left|(P_n - P)[m(\cdot, \hat{\theta}_n) - m(\cdot, \theta_0)]\right| \leq \epsilon|\hat{\theta}_n - \theta_0|^2 + r_n^{-2}R_n^2. \tag{37}$$

To deal with the first term, we show that there exist $\hat{R}_n$ of order $O_{P_r}(1)$ such that

$$\begin{aligned} &\left|(P_n - P)[m(\cdot, \theta) - m(\cdot, \theta_0)]\right| \\ &\quad \leq \epsilon|\theta - \theta_0|^2 + r_n^{-2}R_n^2 \quad \text{for} \quad |\theta - \theta_0| \leq \eta_0. \end{aligned} \tag{38}$$

This can be done similarly to the proof of Lemma 4.1 of Kim and Pollard (1990) as follows. Let $\hat{R}_n$ be the infimum of those values for which the above inequality holds. Define

$$A(n, j) \equiv \{\theta : (j-1)2^{j-1} \leq r_n|\theta - \theta_0| \leq j2^j\}.$$

Then for any constant R, conditioning on $z_1, \ldots, z_n$,

$$\begin{aligned}
&\hat{P}\{\hat{R}_n > R\} \\
&\quad \leq \hat{P}\big\{\sup|(\hat{P}_n - P_n)m(\cdot,\theta)| > \epsilon|\theta-\theta_0|^2 + r_n^{-2}R^2\big\} \\
&\quad \leq \sum_{j=1}^{\infty} \hat{P}\bigg\{\sup_{\theta\in A(n,j)} r_n^2|(\hat{P}_n - P_n)m(\cdot,\theta)| > \epsilon(j-1)^2 2^{2(j-1)} + R^2\bigg\}.
\end{aligned}$$

By Markov's inequality, the jth summand is bounded by

$$r_n^4 \hat{P} \sup_{|\theta-\theta_0|<j2^j/r_n} |(\hat{P}_n - P_n)m(\cdot,\theta)|^2 / \big[\epsilon(j-1)^2 2^{2(j-1)} + R^2\big]^2.$$

By the maximum inequality, the numerator is bounded:

$$r_n^4 \hat{P} \sup_{|\theta-\theta_0|<j2^j/r_n} |(\hat{P}_n - P_n)m(\cdot,\theta)|^2 \leq Cn^{-1} r_n^4 P_n M_{j2^j/r_n}^2.$$

Therefore,

$$\begin{aligned}
P_r\{\hat{R}_n > R\} &= P^{\infty}\hat{P}\{\hat{R}_n > R\} \\
&\leq \sum_{j=1}^{\infty} Cn^{-1} r_n^4 \phi(j2^j/r_n)/[2^{-2}\epsilon(j-1)^2 2^{2j} + R^2]^2 \\
&\leq C\sum_{j=1}^{\infty} j^{\kappa} 2^{\kappa j}/[2^{-2}\epsilon(j-1)^2 2^{2j} + R^2]^2
\end{aligned}$$

by the definition of r_n and the assumption that $\phi(c\eta) \leq c^{\kappa}\phi(\eta)$. Since $\kappa < 4$, the sum in this equation can be made arbitrarily small by choosing R sufficiently large. Combining (36), (37), and (38) yields

$$o_{P_r}(r_n^{-2}) \leq -\epsilon|\hat{\theta}_n - \theta_0|^2 + r_n^{-2}R_n + r_n^{-2}\hat{R}_n,$$

from which it follows that $|\hat{\theta}_n - \theta_0| = O_{P_r}(r_n^{-1})$ and completes the proof of the lemma. *Q.E.D.*

PROOF OF THEOREM 2: Kim and Pollard (1990, Theorem 4.7) show that $\theta_n - \theta_0 = O_{P^{\infty}}(r_n^{-1})$. Lemma 2(i) implies that

$$\theta_n - \theta_0 = O_{P_r}(r_n^{-1}).$$

Furthermore, by Lemma 3,

$$\hat{\theta}_n - \theta_0 = O_{P_r}(r_n^{-1}).$$

Then, (15) follows directly from Theorem 1 and an application of the continuous mapping theorem (Theorem 3). In particular, Theorem 3 is applied with respect to P_r probability for $s_n \equiv r_n(\theta_n - \theta_0)$, $t_n \equiv r_n(\hat{\theta}_n - \theta_0)$, $U_n(t) \equiv Z_n(t)$, $V_n(t) \equiv \hat{Z}_n(t)$, $U(t) \equiv Z(t)$, and $V(t) \equiv \hat{Z}(t)$. By Lemma 2(ii), note that s_n and t_n are both $O_{\hat{P}}(1)$ in P^∞ probability. Also by Lemma 2(ii), the α_n and β_n in condition (iii) of Theorem 3 are both $o_{\hat{P}}(1)$ in P^∞ probability. Therefore, Theorem 3 can also be applied with respect to the probability measure $\hat{P}$ in P^∞ probability, yielding $(r_n(\theta_n - \theta_0), r_n(\hat{\theta}_n - \theta_0)) \xrightarrow{d} (\arg\max Z(t), \arg\max \hat{Z}(t))$ in P^∞ probability. Therefore, $r_n(\hat{\theta}_n - \theta_n) \xrightarrow{d} \arg\max \hat{Z}(t) - \arg\max Z(t)$ in P^∞ probability, from which (16) follows.

Q.E.D.

PROOF OF THEOREM 4: The proof involves verifying the conditions of Theorem 2, showing that $V = V_{\text{ms}}$ and deriving the expression in (25) for the covariance function H. Throughout the proof, we use $G(\cdot)$ to denote various distribution functions, where the argument(s) will indicate the random variable(s) of interest; for instance, $G(\tilde{x}, x_d)$ is the joint distribution function for $\tilde{x}$ and x_d. Consistency of β_n and $\hat{\beta}_n$ follows from Assumptions MS1–MS4. To prove consistency of β_n, we verify the conditions of Theorem 2.1 of Newey and McFadden (1994): (i) $M(\beta)$ is uniquely maximized at β_0, (ii) $\mathcal{B}$ is compact, (iii) $M(\beta)$ is continuous, and (iv) $M(\beta)$ converges uniformly in probability to $M(\beta_0)$. Condition (ii) is true by Assumption MS4. For condition (i),

$$\begin{aligned} M(\beta) &= E\big[(2y-1)(\{\tilde{x}'\beta + x_d \geq 0\} - \{\tilde{x}'\beta_0 + x_d \geq 0\})\big] \\ &= E\big[\big(2F(\tilde{x}'\beta_0 + x_d|\tilde{x}, x_d) - 1\big) \\ &\qquad \times (\{\tilde{x}'\beta + x_d \geq 0\} - \{\tilde{x}'\beta_0 + x_d \geq 0\})\big] \\ &\leq 0 \end{aligned}$$

since $\operatorname{sign}(\tilde{x}'\beta_0 + x_d) = \operatorname{sign}(2F(\tilde{x}'\beta_0 + x_d|\tilde{x}, x_d) - 1)$ for almost every $\tilde{x}$ and x_d (Assumption MS3). Therefore, β_0 maximizes $M(\beta)$ since $M(\beta_0) = 0$. Assume there exists another β_* such that $M(\beta_*) = 0$. Then

$$E\big[\big(2F(\tilde{x}'\beta_0 + x_d|\tilde{x}, x_d) - 1\big)(\{\tilde{x}'\beta_* + x_d \geq 0\} - \{\tilde{x}'\beta_0 + x_d \geq 0\})\big] = 0$$

or, equivalently,

$$\begin{aligned} &E\big[\big(2F(\tilde{x}'\beta_0 + x_d|\tilde{x}, x_d) - 1\big) \\ &\quad \times \big(\{\tilde{x}'\beta_* + x_d \geq 0 > \tilde{x}'\beta_0 + x_d\} - \{\tilde{x}'\beta_0 + x_d \geq 0 > \tilde{x}'\beta_* + x_d\}\big)\big] \\ &\quad = 0. \end{aligned}$$

This equality, combined with Assumption MS3, implies $\{\tilde{x}'\beta_* + x_d \geq 0 > \tilde{x}'\beta_0 + x_d\} = 0$ for almost every $\tilde{x}$ and x_d. By Assumption MS2, this can only be true if $\beta_* = \beta_0$, which proves condition (i). Conditions (iii) and (iv) follow from the uniform law of large numbers (e.g., Lemma 2.4 of Newey and McFadden (1994)) since $|m(z,\beta)|$ is bounded by 1 and $m(z,\beta)$ is continuous at each β with probability 1 (by Assumption MS2(c)). The proof of consistency for $\hat{\beta}_n$ is similar and omitted. For $\eta > 0$, the class of functions $\mathcal{M}_\eta = \{m(z,\beta) : |\beta - \beta_0| \le \eta\}$ is VC subgraph with envelope function

$$\begin{aligned} M_\eta(z) &= \sup_{|\beta-\beta_0|\le\eta} |m(z,\beta)| \\ &= \sup_{|\beta-\beta_0|\le\eta} \left|\{\tilde{x}'\beta \ge -x_d > \tilde{x}'\beta_0\} + \{\tilde{x}'\beta_0 \ge -x_d > \tilde{x}'\beta\}\right| \\ &\le \{-\eta|\tilde{x}| \le \tilde{x}'\beta_0 + x_d \le \eta|\tilde{x}|\}. \end{aligned}$$

Thus, Assumption 6 is satisfied. By Assumptions MS2, MS5, and MS6, the envelope function satisfies $PM_\eta^2 \le O(\eta)$, so that we can take $\phi(\eta) = c\eta$ for some constant c (which leads to the rate $r_n = n^{1/3}$). Assumption 4 is satisfied trivially since the envelope function is bounded by 1 and $\phi(\alpha) = O(\alpha)$. To verify Assumption 5, notice that

$$\begin{aligned} &P(m(\cdot,\beta_1) - m(\cdot,\beta_2))^2 \\ &\quad = \int (2y-1)^2(\{\tilde{x}'\beta_1 + x_d \ge 0\} - \{\tilde{x}'\beta_2 + x_d \ge 0\})^2\, dG(y,\tilde{x},x_d) \\ &\quad = \int (\{\tilde{x}'\beta_1 + x_d \ge 0\} - \{\tilde{x}'\beta_2 + x_d \ge 0\})^2\, dG(\tilde{x},x_d) \\ &\quad \le 2\int \{\tilde{x}'\beta_1 \ge -x_d > \tilde{x}'\beta_2\}\, dG(\tilde{x},x_d) \\ &\qquad + 2\int \{\tilde{x}'\beta_2 \ge -x_d > \tilde{x}'\beta_1\}\, dG(\tilde{x},x_d). \end{aligned}$$

Both of the terms in the last line are $O(|\beta_1 - \beta_2|)$. For the first term (the second term is similar),

$$\begin{aligned} &2\int \{\tilde{x}'\beta_1 \ge -x_d > \tilde{x}'\beta_2\}\, dG(\tilde{x},x_d) \\ &\quad = 2\int \{\tilde{x}'\beta_1 \ge -x_d > \tilde{x}'\beta_2\} g(x_d|\tilde{x})\, dG(\tilde{x}) \\ &\quad = O(1)\int |\beta_1 - \beta_2||\tilde{x}|\, dG(\tilde{x}) \\ &\quad = O(|\beta_1 - \beta_2|), \end{aligned}$$

where the second equality follows from the boundedness of $g(x_d|\tilde{x})$ (Assumption MS6) and the third equality follows from Assumption MS5. Thus, Assumption 5 is satisfied since

$$P\big(m(\cdot,\beta_0+\alpha s)-m(\cdot,\beta_0+\alpha t)\big)^2=O\big(|\alpha(s-t)|\big),$$

which implies that

$$\frac{P(m(\cdot,\beta_0+\alpha s)-m(\cdot,\beta_0+\alpha t))^2}{\alpha}=O(|s-t|).$$

For any $u, v \in \mathbb{R}^{d-1}$, the covariance kernel $H(u,v)$ in Assumption 3 is

$$\begin{aligned}
H(u,v) &= \lim_{\alpha\to 0}\alpha^{-1}Pm(\cdot,\beta_0+\alpha u)m(\cdot,\beta_0+\alpha v)\\
&= \lim_{\alpha\to 0}\alpha^{-1}\int \big(\{\tilde{x}'(\beta_0+\alpha u)\ge -x_d>\tilde{x}'\beta_0\}\\
&\qquad - \{\tilde{x}'\beta_0\ge -x_d>\tilde{x}'(\beta_0+\alpha u)\}\big)\\
&\qquad \times \big(\{\tilde{x}'(\beta_0+\alpha v)\ge -x_d>\tilde{x}'\beta_0\}\\
&\qquad - \{\tilde{x}'\beta_0\ge -x_d>\tilde{x}'(\beta_0+\alpha v)\}\big)\,dG(\tilde{x},x_d)\\
&= \lim_{\alpha\to 0}\alpha^{-1}\int \big(\{\tilde{x}'u>0\}\{\tilde{x}'v>0\}\\
&\qquad \times \big\{\min(\tilde{x}'(\beta_0+\alpha u),\tilde{x}'(\beta_0+\alpha v))\\
&\qquad\quad \ge -x_d>\tilde{x}'\beta_0\big\}\\
&\qquad + \{\tilde{x}'u<0\}\{\tilde{x}'v<0\}\\
&\qquad \times \big\{\tilde{x}'\beta_0\ge -x_d\\
&\qquad\quad > \max(\tilde{x}'(\beta_0+\alpha u),\\
&\qquad\qquad \tilde{x}'(\beta_0+\alpha v))\big\}\big)\,dG(\tilde{x},x_d)\\
&= \int \big(\{\tilde{x}'u>0\}\{\tilde{x}'v>0\}\min(\tilde{x}'u,\tilde{x}'v)\\
&\qquad - \{\tilde{x}'u<0\}\{\tilde{x}'v<0\}\max(\tilde{x}'u,\tilde{x}'v)\big)g(-\tilde{x}'\beta_0|\tilde{x})\,dG(\tilde{x})\\
&= E\big[\{\mathrm{sign}(\tilde{x}'u)=\mathrm{sign}(\tilde{x}'v)\}\min(|\tilde{x}'u|,|\tilde{x}'v|)\,g(-\tilde{x}'\beta_0|\tilde{x})\big].
\end{aligned}$$

The boundedness of $g(\cdot|\tilde{x})$ in a neighborhood of zero (Assumption MS6) and Assumption MS5 ensure the existence of the integrals above and, therefore, the existence of $H(u,v)$. This satisfies Assumption 3. For Assumption 2, first note that the limiting objective function $M(\beta)$ can be written

$$(39)\qquad M(\beta)=\int\big(2F(\tilde{x}'\beta_0+x_d|\tilde{x},x_d)-1\big)\times(\{\tilde{x}'\beta+x_d\geq 0\}-\{\tilde{x}'\beta_0+x_d\geq 0\})\,dG(\tilde{x},x_d).$$

Differentiating with respect to β yields

$$\begin{aligned}\nabla_\beta M(\beta)&=\int\big(2F(\tilde{x}'\beta_0+x_d|\tilde{x},x_d)-1\big)\{\tilde{x}'\beta+x_d=0\}\tilde{x}\,dG(\tilde{x},x_d)\\&=\int\big(2F(\tilde{x}'\beta_0-\tilde{x}'\beta|\tilde{x},-x'\beta)-1\big)g(-x'\beta|\tilde{x})\tilde{x}\,dG(\tilde{x}).\end{aligned}$$

Assumptions MS5 and MS6 ensure that this integral is well defined for β in a neighborhood around β_0. Differentiating again with respect to β yields

$$\begin{aligned}\nabla_{\beta\beta}M(\beta)&=-\int\big(2F(\tilde{x}'\beta_0-\tilde{x}'\beta|\tilde{x},-x'\beta)-1\big)g'(-x'\beta|\tilde{x})\tilde{x}\tilde{x}'\,dG(\tilde{x})\\&\quad+\int 2\nabla_\beta F(\tilde{x}'\beta_0-\tilde{x}'\beta|\tilde{x},-x'\beta)g(-x'\beta|\tilde{x})\tilde{x}\,dG(\tilde{x})\\&=-\int\big(2F(\tilde{x}'\beta_0-\tilde{x}'\beta|\tilde{x},-x'\beta)-1\big)g'(-x'\beta|\tilde{x})\tilde{x}\tilde{x}'\,dG(\tilde{x})\\&\quad-\int 2f(\tilde{x}'\beta_0-\tilde{x}'\beta|\tilde{x},-x'\beta)g(-x'\beta|\tilde{x})\tilde{x}\tilde{x}'\,dG(\tilde{x})\\&\quad-\int 2\frac{\partial F(\tilde{x}'\beta_0-\tilde{x}'\beta|\tilde{x},-x'\beta)}{\partial x_d}g(-x'\beta|\tilde{x})\tilde{x}\tilde{x}'\,dG(\tilde{x}).\end{aligned}$$

Assumptions MS5, MS6, MS7, and MS8 ensure the existence of $\nabla_{\beta\beta}M(\beta)$ for β in a neighborhood around β_0. Since $F(0|\tilde{x},x_d)=0.5$ for almost every $\tilde{x}$ and x_d (Assumption MS3), both the first term and the third term drop out when $\nabla_{\beta\beta}M(\beta)$ is evaluated at β_0. Therefore,

$$V\equiv-\nabla_{\beta\beta}M(\beta_0)=E\big[2f(0|\tilde{x},-\tilde{x}'\beta_0)g(-\tilde{x}'\beta_0|\tilde{x})\tilde{x}\tilde{x}'\big]\equiv V_{\mathrm{ms}},$$

which completes the proof since V_{ms} is assumed to be positive definite (Assumption MS10). *Q.E.D.*

PROOF OF THEOREM 5: Since Van der Vaart and Wellner (1996, Example 3.2.15) derived the asymptotic distribution in (31), we summarize their approach but also incorporate the bootstrap so that Theorem 2 may be applied. Let $G_n(x)$ denote the empirical distribution function of the x_i's and define

$$V_n(x)\equiv n^{-1}\sum_{i=1}^{n}y_i\{x_i\leq x\}.$$

Then

$$F_n(x) \le a \quad \text{if and only if} \quad U_n(a) \ge x,$$

where

$$U_n(a) \equiv \sup_s \{V_n(s) - aG_n(s) \text{ is minimal}\}.$$

Then

$$\begin{aligned} P\big(n^{1/3}(F_n(x_0) - F(x_0)) \le s\big) &= P\big(U_n(F(x_0) + sn^{-1/3}) \ge x_0\big) \\ &= P(h_n \ge 0), \end{aligned}$$

where $h_n \equiv \sup_h \{V_n(x_0 + hn^{-1/3}) - (F(x_0) + sn^{-1/3})G_n(x_0 + hn^{-1/3}) \text{ is minimal}\}$. Letting $g_n \equiv n^{1/3}h_n$ and $z \equiv (y, x)$, the value g_n maximizes the objective function $M_n(g) \equiv P_n m(\cdot, g)$ with

$$m(z, g) \equiv (y - F(x_0) - sn^{1/3})(\{x \le x_0 + g\} - \{x \le x_0\}).$$

It is straightforward to show that g_n is consistent (i.e., converges to zero in probability) and that Assumptions 4–6 hold (see, for example, Van der Vaart and Wellner (1996, Example 3.2.14)). The bootstrap counterparts for all of the above quantities are denoted with hats. So, for instance, $\hat{U}_n(a) \equiv \sup_s\{\hat{V}_n(s) - a\hat{G}_n(s) \text{ is minimal}\}$, where $\hat{G}_n$ is the empirical distribution function of the bootstrap x sample and $\hat{V}_n$ is defined analogously to V_n except on the bootstrap sample. Then the bootstrap estimator $\hat{F}_n$ is related to $\hat{U}_n$ as

$$\hat{F}_n(x) \le a \quad \text{if and only if} \quad \hat{U}_n(a) \ge x.$$

The quantities $\hat{h}_n$ and $\hat{g}_n$ are defined as

$$\begin{aligned} \hat{h}_n \equiv \sup_h \big\{&\hat{V}_n(x_0 + hn^{-1/3}) \\ &- (F(x_0) + sn^{-1/3})\hat{G}_n(x_0 + hn^{-1/3}) \text{ is minimal}\big\}, \\ \hat{g}_n \equiv n^{1/3}&\hat{h}_n. \end{aligned}$$

The convergence of $\hat{g}_n$ to zero in probability follows from similar arguments to those for g_n. Theorem 1 can be applied almost directly, except that a linear term (linear in h) is included in deriving the limiting processes. In particular, $Z(t) = -t^2V/2 + W(t) - st$ and $\hat{Z}(t) = -t^2V/2 + W(t) + \hat{W}(t) - st$. It is easy to verify that $V = -f(x_0)g(x_0)$, which exists and is nonzero by the assumptions on $f(x_0)$ and $g(x_0)$, and $H(u, v) = \{\text{sign}(u) = \text{sign}(v)\}F(x_0)(1 - F(x_0))g(x_0)\min(|u|, |v|)$. Thus,

$$Z(t) = t^2 f(x_0)g(x_0)/2 + \sqrt{F(x_0)(1 - F(x_0))g(x_0)}B(t) - sg(x_0)t$$

and

$$\hat{Z}(t) = t^2 f(x_0)g(x_0)/2 + \sqrt{F(x_0)(1-F(x_0))g(x_0)}B(t) + \sqrt{F(x_0)(1-F(x_0))g(x_0)}\hat{B}(t) - sg(x_0)t,$$

where B and $\hat{B}$ are independent two-sided Brownian motions originating at zero. Applying Theorem 2 yields (in P_r probability)

$$n^{1/3}\begin{pmatrix} g_n \\ \hat{g}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \arg\max(Z(t)) \\ \arg\max(\hat{Z}(t)) \end{pmatrix}.$$

By properties of two-sided Brownian motion (see, e.g., Van der Vaart and Wellner (1996, p. 308)),

$$n^{1/3}\begin{pmatrix} g_n \\ \hat{g}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} c\arg\max(-t^2 + B(t)) - d \\ c\arg\max(-t^2 + B(t) + \hat{B}(t)) - d \end{pmatrix},$$

where the constants c and d are given by

$$c \equiv \left(\frac{2\sqrt{F(x_0)(1-F(x_0))}}{f(x_0)\sqrt{g(x_0)}}\right) \quad \text{and} \quad d \equiv -\frac{s}{f(x_0)}.$$

Then

$$\begin{aligned}
&P\big(n^{1/3}(F_n(x_0) - F(x_0)) \le s\big) \\
&\quad = P(g_n \ge 0) \\
&\quad \to P\big(c\arg\max(-t^2 + B(t)) - d \ge 0\big) \\
&\quad = P\big(cf(x_0)\arg\max(-t^2 + B(t)) \ge -s\big) \\
&\quad = P\Big(\big(4F(x_0)(1-F(x_0))f(x_0)/g(x_0)\big)^{1/3} \\
&\qquad\qquad \times \arg\max(-t^2 + B(t)) \ge -s\Big) \\
&\quad = P\Big(\big(4F(x_0)(1-F(x_0))f(x_0)/g(x_0)\big)^{1/3}\arg\max(-t^2 + B(t)) \le s\Big),
\end{aligned}$$

where the last equality follows from the symmetry of $\arg\max(-t^2 + B(t))$. Similarly, since $\hat{U}_n(F(x_0) + sn^{-1/3}) - x_0 = n^{1/3}\hat{h}_n$,

$$\begin{aligned}
&P_r\big\{n^{1/3}(\hat{F}_n(x_0) - F(x_0)) \le s\big\} \\
&\quad = P_r(\hat{g}_n \ge 0) \\
&\quad \to P\Big(\big(4F(x_0)(1-F(x_0))f(x_0)/g(x_0)\big)^{1/3} \\
&\qquad\qquad \times \arg\max(-t^2 + B(t) + \hat{B}(t)) \le s\Big).
\end{aligned}$$

Combining these limiting distributions yields (in P_r probability)

$$n^{1/3}\left(\frac{4F(x_0)(1-F(x_0))f(x_0)}{g(x_0)}\right)^{-1/3}\begin{pmatrix}F_n(x_0)-F(x_0)\\ \hat{F}_n(x_0)-F(x_0)\end{pmatrix}$$
$$\xrightarrow{d}\begin{pmatrix}\arg\max(-t^2+B(t))\\ \arg\max(-t^2+B(t)+\hat{B}(t))\end{pmatrix}.$$

The limit distributions in (31) and (32) then follow directly from the argument at the end of the proof of Theorem 2. *Q.E.D.*

REFERENCES

ABREVAYA, J. (2000): "Rank Estimation of a Generalized Fixed-Effects Regression Model," *Journal of Econometrics*, 95, 1–23.

ARCONES, M. A., AND E. GINÉ (1992): "On the Bootstrap of M-Estimators and Other Statistical Functionals," in *Exploring the Limits of Bootstrap*, ed. by R. LePage and L. Billard. New York: Wiley, 13–47.

BARTIK, T. J., J. S. BUTLER, AND J. LIU (1992): "Maximum Score Estimates of the Determinants of Residential Mobility: Implications for the Value of Residential Attachment and Neighborhood Amenities," *Journal of Urban Economics*, 32, 233–256.

DELGADO, M. A., J. M. RODRIGUEZ-POO, AND M. WOLF (2001): "Subsampling Inference in Cube Root Asymptotics with an Application to Manski's Maximum Score Estimator," *Economics Letters*, 73, 241–250.

DYKSTRA, R., AND C. CAROLAN (1999): "The Distribution of the Argmax of Two-Sided Brownian Motion with Quadratic Drift," *Journal of Statistical Computation and Simulation*, 63, 47–58.

EFRON, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

GOODWIN, B. K. (1996): "Semiparametric (Distribution-Free) Testing of the Expectations Hypothesis in a Parimutuel Gambling Market," *Journal of Business & Economic Statistics*, 14, 487–496.

GROENEBOOM, P. (1989): "Brownian Motion with a Parabolic Drift and Airy Functions," *Probability Theory and Related Fields*, 81, 79–109.

GROENEBOOM, P., AND J. A. WELLNER (1992): *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel: Birkhäuser.

——— (2001): "Computing Chernoff's Distribution," *Journal of Computational and Graphical Statistics*, 10, 388–400.

HAHN, J. (1995): "Bootstrapping Quantile Regression Estimators," *Econometric Theory*, 11, 105–121.

HAUSMAN, J. A., J. ABREVAYA, AND F. M. SCOTT-MORTON (1998): "Misclassification of the Dependent Variable in a Discrete-Response Setting," *Journal of Econometrics*, 87, 239–269.

HOROWITZ, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

——— (2002): "Bootstrap Critical Values for Tests Based on the Smoothed Maximum Score Estimator," *Journal of Econometrics*, 111, 141–167.

HUANG, J., AND J. A. WELLNER (1995): "Estimation of a Monotone Density or Monotone Hazard under Random Censoring," *Scandinavian Journal of Statistics*, 22, 3–33.

KIM, J., AND D. POLLARD (1990): "Cube Root Asymptotics," *The Annals of Statistics*, 18, 191–219.

MANSKI, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

——— (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333.

——— (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357–362.

MANSKI, C. F., AND T. S. THOMPSON (1986): "Operational Characteristics of Maximum Score Estimation," *Journal of Econometrics*, 32, 85–108.

MAYER, W. J., T. BLACKSTONE, B. SHANG, AND N. K. WOMER (2002): "Maximum Score Estimation of Binary Choice Panel Data Models with Sample Selection: An Application to Navy Officer Retention," Unpublished Manuscript, Department of Economics, University of Mississippi.

NEWEY, W. K., AND D. L. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. L. McFadden. Amsterdam: North-Holland, 2111–2245.

POLLARD, D. (1989): "Asymptotics via Empirical Processes" (with Discussion), *Statistical Science*, 4, 341–366.

——— (1990): *Empirical Processes: Theory and Applications. Regional Conference Series in Applied Mathematics*. Washington, DC: Conference Board of the Mathematical Sciences.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge, U.K.: Cambridge University Press.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

WELLNER, J. A., AND Y. ZHAN (1996): "Bootstrapping Z-Estimators," Unpublished Manuscript, Department of Statistics, University of Washington.