ESTIMATING FEATURES OF A DISTRIBUTION FROM BINOMIAL DATA*

Arthur Lewbel†
Boston College

Daniel McFadden‡
University of California, Berkeley

Oliver Linton§
London School of Economics

July 2010

Abstract

We propose estimators of features of the distribution of an unobserved random variable W. What is observed is a sample of Y, V, X where a binary Y equals one when W exceeds a threshold V determined by experimental design, and X are covariates. Potential applications include bioassay and destructive duration analysis. Our empirical application is referendum contingent valuation in resource economics, where one is interested in features of the distribution of values W (willingness to pay) placed by consumers on a public good such as endangered species. Sample consumers with characteristics X are asked whether they favor (with $Y = 1$ if yes and zero otherwise) a referendum that would provide the good at a cost V specified by experimental design. This paper provides estimators for quantiles and conditional on X moments of W under both nonparametric and semiparametric specifications.

*This research was supported in part by the National Science Foundation through grants SES-9905010 and SBR-9730282, by the E. Morris Cox Endowment, and by the ESRC. The authors would like to thank anonymous referees and the co-editor for many helpful suggestions.

†Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA. Phone: (617) 552-3678. E-mail address: lewbel@bc.edu

‡Department of Economics, University of California, Berkeley, CA 94720-3880, USA. E-mail address: mcfadden@econ.berkeley.edu

§Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. E-mail address: o.linton@lse.ac.uk. This paper was partly written while I was a Universidad Carlos III de Madrid-Banco Santander Chair of Excellence, and I thank them for financial support.

JEL Codes: C14, C25, C42, H41. Keywords: Willingness to Pay, Contingent Valuation, Discrete Choice, Binomial response, Bioassay, Destructive Duration Testing, Semiparametric, Nonparametric, Latent Variable Models.

1 Introduction

Consider an experiment where an individual is asked if he would be willing to pay more than V dollars for some product. Let unobserved W be the most the individual would be willing to pay, and let the individual's response be $Y = 1(W > V)$, so $Y = 1$ if his latent willingness to pay (WTP) is greater than the proposed bid price V, and zero otherwise. In a typical experiment like this V is a random draw from some distribution determined by the researcher. For example, in our empirical application V is chosen by the researcher to be one of fourteen different possible dollar values ranging from \$25 to \$375, and the question is whether the individual would be willing to pay more than V dollars to protect wetlands in California. Given data on Y and V, the goal of the analysis is estimation of features of the distribution of W across individuals, such as moments and quantiles.

We also observe covariates X in addition to Y and V, and we consider estimation of $\mu_r(x) = E(r(W, X) \mid X = x)$ for some function r specified by the researcher. For example, our data includes education level and gender, so we could let $r(W, X) = W$ to estimate the mean WTP $E(W \mid X = x)$ among college educated women x. Letting t denote a parameter, other examples are $r(W, X) = e^{tW}$, leading to the moment generating function; $r(W, X) = 1(W \leq t)$, leading to the probability of the event $W \leq t$; and $r(W, X) = 1(W \leq t) W$, leading to a trimmed mean, all conditioned on, say, likely voters x. Our analysis permits the experimental design to depend on x, e.g., wealthier individuals could be assigned relatively high bids.[1] Our estimators are readily extended to interval censored data with multiple (adaptive) test levels and multinomial status.

We have given the example of a contingent valuation study using a referendum format elicitation,

[1]Other experimental designs include follow up queries to gain more information about WTP, and open ended questions, where subjects are simply asked to state their WTP. Open ended questions often suffer from high rates of nonresponse (with possible selection bias), while referendum format follow up responses can be biased due to the framing effect of the first bid. This shadowing effect is common in unfolding bracket survey questions. See McFadden (1994) for references and experimental evidence regarding response biases. Other issues regarding the framing of questions also impact survey responses, particularly anchoring to test values, including the initial test value; see Green et al. (1998) and Hurd et al. (1998). The data generation process may then be a convolution of the target distribution and a distribution of psychometric errors. This paper will ignore these issues and treat the data generation process as if it is the target distribution. However, we do empirically apply our estimators separately to first round and follow up bids, and find differences in the results, which provides evidence that such biases are present. The difficult general problem of deconvoluting a target distribution in the presence of psychometric errors is left for future research.

but the same structure arises in many other contexts. The problem can be generally described as uncovering features of conditional survival curves from unbounded interval censored data. Let W denote a random failure time, and let $G(w \mid x) = \Pr(W > w \mid X = x)$ denote the survival curve conditioned on time invariant covariates X. For example, in bioassay, V is the time an animal exposed to an environmental hazard is sacrificed for testing, G is the distribution of survival times until the onset of abnormality, and Y is an indicator for an abnormality found by the test, termed a current status observation. Duration may also occur in dimensions other than time. In dose-response studies, V is the administered dose of a toxin, W is the lethal dose, and G is the dose-response curve, and Y is a mortality indicator. For materials testing, Y indicates that the material meets some requirement at treatment level V, e.g., G could be the distribution of speeds W at which a car safety device fails, with $Y = 1$ indicating failure at test speed V.

A common procedure is to completely parameterize W, e.g., to assume W equals $X^\top\theta_0 - \varepsilon$ with $\varepsilon \sim N(\alpha_0, \sigma^2)$. The model then takes the form of a standard probit $Y = I[X^\top\theta_0 - V > \varepsilon]$ and can be estimated using maximum likelihood. However, estimation of the features of the distribution of W differs from ordinary binomial response model estimation when the model is not fully parameterized, because the goal is estimation of moments or quantiles of W, rather than response or choice probabilities of Y. So, for example, in the above parameterized model $E(W \mid X = x) = X^\top\theta_0 - \alpha_0$, and therefore any binomial response model estimator that fails to estimate the location term α_0, such as the semiparametrically efficient estimator of Klein and Spady (1993), is inadequate for estimation of moments of W.

Another important difference is the role of the support of V. By construction $G(v \mid x) = E(Y \mid V = v, X = x)$, so G can be estimated using ordinary parametric, semiparametric, or nonparametric conditional mean estimation. But nonparametric estimation of moments of W then requires identification of $G(v \mid x)$ everywhere on the support of W, so nonparametric identification requires that the support of V contains the support of W. However, virtually all experiments only consider a small number of values for v. While the literature contains many estimators of moments of W,[2] virtually all of them are parametric or semiparametric, using functional form assumptions to obtain identification, without recognizing or acknowledging the resulting failure of nonparametric identification.[3]

Our nonparametric estimators obtain identification by assuming either that bids v are draws from a continuously distributed random variable V, or that the experimental design varies with the

[2]See, e.g., Kanninen (1993) and Crooker and Herriges (2004) for comparisons of various, mostly parametric, WTP estimators. Estimators that are not fully parameterized include Chen and Randall (1997), Creel and Loomis (1997), and An (2000) for WTP and Ramgopal, Laud, and Smith (1993), and Ho and Sen (2000) for bioassay.

[3]In supplementary materials to this paper, we show that, given a fixed discrete design for V, even assuming that $W = m(X) - \varepsilon$ with X and ε independent is still not sufficient for identification, though identification does become possible in this case if $m(X)$ is finitely parameterized.

sample size n, so for any fixed n there may be a finite number of values bids can take on, but this number of possible bid values becomes dense in the support of W as n goes to infinity.[4] We also show how this dependence of survey design on sample size affects the resulting limiting distributions, and we provide an alternative identifying assumption based on a semiparametric specification of W described below.[5]

With an estimate of the density $g(w \mid x) = -\partial G(w \mid x)/\partial w$ and sufficient identifying assumptions, features of the distribution of W such as moments and percentiles can be readily recovered. In particular, moments $\mu_r(x) = E(r(W,X) \mid X = x) = \int r(w,x)g(w \mid x)dw$ can be estimated in two steps, using second-step numerical integration after plugging in a first-step estimate of the conditional density. We provide alternative semiparametric and nonparametric estimators of $\mu_r(x)$ that do not require estimation of $g(w \mid x)$. These estimators utilize the feature that v is determined by experimental design, use integration by parts to obtain an expression for $\mu_r(x)$ that depends on $G(w \mid x)$ but not its derivative, and use numerical integration methods that work with first-step undersmoothed estimators of $G(w \mid x)$ at a limited number of convenient evaluation points. When the experimental design for v is known, we provide estimators for smooth moments of W that use only the indicator Y and do not require a first-step estimator of $G(w \mid x)$.

We consider estimation for two different information conditions on the conditional distribution of W given X. In the most general case, this distribution is completely unrestricted apart from smoothness, and is estimated nonparametrically. We may write this case as $W = m(X, \varepsilon)$ with m unknown and ε an unobserved disturbance that is independent of X. This includes as a more restrictive case the location model $W = m(X) - \varepsilon$. We also include here the special cases where we are interested in unconditional moments of W, or in conditional moments when X has finite support.

The second case we analyze is the semiparametric model $W = \Lambda[m(X, \theta_0) - \varepsilon]$ for known functions m and Λ, an unknown finite parameter vector θ_0, and a distribution for the disturbance ε that is known only to be independent of X. This model includes as special cases the probit model discussed earlier, similar logit models, and the Weibull proportional hazards model in which Λ is exponential and ε is extreme value. In this semiparametric model, identification requires that the support of $m(X, \theta_0) - \Lambda^{-1}(V)$ become dense in the support of ε; if X includes a continuously distributed component, this can be achieved even if the support of V is fixed and finite.

We also consider estimation for two information conditions on the asymptotic distribution of the

[4] Virtually all existing contingent valuation data sets draw bids from discrete distributions. However, large surveys typically have bid distributions with more mass points than small surveys, consistent with our assumption of an increasing number of bid values as sample size grows. See, e.g., Crooker and Herriges (2000) for a study of WTP bid designs, with explicit consideration of varying numbers of mass points.

[5] An approach that we do not pursue in this paper is to sacrifice point identification and instead estimate bounds on features of G, as in McFadden (1998). See also Manski and Tamer (2002).

Information Conditions	Nonparametric G	Semiparametric G
Density of v known	$\widehat{\mu}_{1r}(x)$	$\widehat{\mu}_{3r}(x)$
Density of v unknown	$\widehat{\mu}_{2r}(x)$	$\widehat{\mu}_{4r}(x)$

Table 1: Information Used to Construct Different Estimators

bid values V, the case where this is known to the researcher, and the case where it is unknown. We provide estimators, and associated limiting normal distributions, for these primary information conditions, a Monte Carlo analyses of the estimators, and an empirical application estimating conditional mean WTP to protect wetland habitats in California's San Joaquin Valley. The two-step estimator that uses a first-step estimator of $g(w \mid x)$ is denoted $\widehat{\mu}_{0r}(x)$. Table 1 gives the notation for the other estimators we offer for the various information conditions.

2 Estimators

2.1 The Data Generation Process and Estimands

Let $G(w \mid x) = \Pr(W > w \mid X = x)$, so G is the unknown complementary cumulative distribution function of a latent, continuously distributed unobserved random scalar W, conditioned on a vector of observed covariates X. Let $g(w \mid x)$ denote the conditional probability density function of W, so $g = -dG/dw$. A test value v (a realization of V) is set by an experimental design or natural experiment. Define $Y = 1(W > V)$ where $1(\cdot)$ is the indicator function. The observed data consist of a sample of realizations of covariates X, test values V, and outcomes Y. The framework is similar to random censored regressions (with censoring point V), except that for random censoring we would observe W for observations having $W > V$, whereas in the present context we only observe $Y = I(W > V)$.

Given a function $r(w, x)$, the goal is estimation of the conditional moment $\mu_r(x) = E[r(W, X) \mid X = x]$ for any chosen x in the support of X. Let $r'(w, x)$ denote $\partial r(w, x)/\partial w$ wherever it exists, and let $G^{-1}(\cdot \mid x)$ denote the inverse of the function $G(w \mid x)$ with respect to its first argument. We assume the conditional distribution of W given $X = x$ is not finitely parameterized, since otherwise ordinary maximum likelihood estimation would suffice.

Assumption A.1. *The covariate vector X is composed of a possibly empty discrete subvector Q that ranges over a finite number of configurations, and a possibly empty continuous subvector Z that ranges over a compact rectangle in $\mathbb{R}^d$, Q has a positive density $p_1(q)$, and Z has a positive Lipschitz-continuous density $p_2(z \mid q)$. The latent scalar W has an unknown conditional CDF $1 - G(w \mid x)$*

for $x = (q, z)$ with compact support $[\rho_0(x); \rho_1(x)]$, and $G(w \mid x)$ is continuously differentiable with Lipschitz-continuous derivatives and a positive density function $g(w \mid x)$. The variables W and V are conditionally independent, given X, and $Y = I(W > V)$.

ASSUMPTION A.2. *The function $r(w, x)$, chosen by the researcher, is continuous in (w, x) and is continuously differentiable in w, with a uniformly Lipschitz derivative, for each x.*

We term a function $r(w, x)$ satisfying Assumption 2 *regular*. From Assumption A.1, and in particular the conditional independence of W and V,

$$G(v \mid x) = E(Y \mid V = v, X = x) = \Pr(Y = 1 | V = v, X = x). \tag{1}$$

For a regular function $r(w, x)$, integration by parts yields

$$\mu_r(x) = \int_{\rho_0(x)}^{\rho_1(x)} r(w, x) g(w \mid x) dw = r(\rho_0(x), x) + \int_{\rho_0(x)}^{\rho_1(x)} r'(v, x) G(v \mid x) dv. \tag{2}$$

The regular class includes smooth functions such as $r(w, x) = w^t$ and $r(w, x) = e^{tw}$ for a parameter t that correspond to moments and to the moment generating function.

For any $\kappa(x)$ having $\rho_0(x) < \kappa(x) < \rho_1(x)$, we can rewrite equation (2) as

$$\mu_r(x) = r(\kappa(x), x) + \int_{\rho_0(x)}^{\rho_1(x)} r'(v, x) \left[G(v \mid x) - 1(v < \kappa(x)) \right] dv \tag{3}$$

The estimators we consider below are obtained by substituting first-step estimates or empirical analogs of $g(v \mid x)$ or $G(v \mid x)$ into (2) or (3). The parameter $\kappa(x)$ doesn't affect the estimand $\mu_r(x)$, but it can affect some estimators in finite samples even though it drops out asymptotically. For simplicity we later demean v and take $\kappa(x)$ to be zero, or otherwise choose some central value for $\kappa(x)$.

It is possible to extend the regular class to include functions with a finite number of breaks, with a corresponding extension of the integration by parts formula (2); this can be used to obtain analogs of the estimators in this paper for conditional percentiles and trimmed moments[6].

If $G(w \mid x)$ is not at least partly parameterized, then equation (1) implies that for identification of the distribution of W, the support of V should contain the support of W. As noted in the introduction, and by the identification analysis in the supplemental Appendix to this paper, the distribution of W is in general not identified when the asymptotic support of V has a finite number of elements. To identify features of the distribution of W with minimal restrictions on G, our

[6]If $r(w, x)$ has possible break points at $\rho_0(x) = w_1(x) < ... < w_K(x) = \rho_1(x)$, integrating by parts between the breakpoints gives $\mu_r(x) = r(\rho_0(x)^+, x) + \sum_{k=2}^{K-1} \left[r(w_k(x)^+, x) - r(w_k(x)^-, x) \right] G(w_k(x) \mid x) dv + \int_{\rho_0(x)}^{\rho_1(x)} r'(v, x) G(v \mid x) dv$ when the one-sided limits exist.

nonparametric estimators assume an experimental design in which the test values of V become dense in the support of W as the sample size grows to infinity. Let $H_n(v, x \mid n)$ denote the empirical distribution function for observations of (V, X) for a sample of size n. Realizations could be random draws from a CDF $H(v, x \mid n)$, but the data, particularly bids, could also be derived from some purposive sampling protocol. The requirement we place on the data generating process to assure nonparametric identification is the following:

ASSUMPTION A.3. *There exists a CDF $H(v, x)$ with the property that the corresponding conditional distribution of test values V given $X = x$, denoted $H(v \mid x)$, has a strictly positive continuous density $h(v \mid x)$ with a compact support $[\delta_0(x), \delta_1(x)]$ that contains the support of W. The empirical distribution function satisfies $\sup_n |H_n(v, x \mid n) - H(v, x)| \to 0$ almost surely, and $n^\tau [H_n(v, x \mid n) - H(v, x)]$ converges weakly to a Gaussian process for some τ with $\tau = 1/2$ for root-n asymptotics.*

Two examples illustrate this data generating process assumption:

1. Suppose for each sample observation $i = 1, \ldots, n$, X_i, V_i is drawn randomly from the CDF $H(v, x)$. Then the required sup norm convergence follows by the Glivenko-Cantelli theorem, and the convergence to a Gaussian process with $\tau = 1/2$ can be shown by, e.g., the Shorack and Wellner (1986 p. 108ff) treatment of triangular arrays of empirical processes.

2. For each sample size n, supppose x_i is drawn at random from a distribution, and that v_i is drawn with random or quota sampling from a distribution $H(v \mid x_i, n)$ that has a finite support containing J_n points, and let $\rho_0(x) = v_{0n}(x) < ... \; v_{J_n+1,n}(x) = \rho_1(x)$ denote these points plus the end points. Suppose that $J_n \leq n$, $n^{-1/2-\gamma} J_n \to \infty$ for some $\gamma \in (0, 1/2)$, the maximum spacing S_n between the points $v_{jn}(x)$ satisfies $plim_{n\to\infty} n^{1/2} S_n = 0$, and $H(\cdot \mid x_i, n)$ converges to a distribution with a positive density $h(v \mid x)$. Let M be a bound on $r'(v, x)$, $r''(v, x)$ and $g(v \mid x)$. Then, starting from equation (3),

$$
\begin{aligned}
& n^{1/2} \Big| \int_{\rho_0(x)}^{\rho_1(x)} r'(v, x) \left[G(v \mid x) - 1\left(v < \kappa\left(x\right)\right) \right] dv - \sum_{j=1}^{J_n} r'(v_{jn}, x) \left[G(v_{jn} \mid x) - 1\left(v_{jn} < \kappa\left(x\right)\right) \right] \Big| \\
\leq \;\; & M \left(1 + K + M + M^2\right) n^{1/2} S_n \to_p 0
\end{aligned}
$$

and the numerical integration error associated with this design process for test values is root-n asymptotically negligible. If the design points are drawn randomly from a density $h(v \mid x)$ that satisfies $\min_v h(v \mid x) \geq m > 0$, then from David (2003, p. 327) and the constraints $n^\gamma \leq n^{-1/2} J_n \leq n^{1/2}$ for n large, $\lim_n \Pr\left(n^{1/2} S_n < c\right) \geq \lim_n \exp\left(-\exp\left(-n^{-1/2} J_n m c + \ln J_n\right)\right) = 1$, and the condition

$plim_{n\to\infty} n^{1/2} S_n = 0$ holds. Then, the design process in this example with the constraints on J_n are sufficient to make its deviation from the previous random sampling example root-n asymptotically negligible.

This second example covers all current contingent valuation studies of WTP provided they are embedded in design processes with test values satisfying the limit conditions on J_n and on the distributions $H(v \mid x, n)$. Of course, the statement that these designs can be embedded in processes that lead to consistent, normal asymptotics does not guarantee that these asymptotics provide a good approximation to finite-sample behavior. In our simulation studies, we will examine the size of finite sample bias that results when our estimators are applied with both discrete and continuous designs for the test values V.

2.2 Nonparametric Moments

For estimation we suppose that a sample (Q_i, Z_i, V_i, Y_i) with $X_i = (Q_i, Z_i)$ generated in accordance with Assumption A.3 for $i = 1, ..., n$. First consider $G(v \mid x)$ and $g(v \mid x)$. Let E_n denote the sample empirical expectation over functions of the random variables (Q, Z, V, Y). For concreteness and ease of exposition, in this section we will just consider Nadaraya-Watson kernel estimators to show consistency and asymptotic normal convergence rates. Later we provide limiting distribution theory, including explicit variance formulas, for a more general class of estimators, including local polynomials, that may be numerically preferable in applications.

Let K_1 and K_2 be kernel functions that are symmetric continuously differentiable densities with compact support on $\mathbb{R}$ and $\mathbb{R}^d$ respectively, and let λ denote a bandwidth parameter. Our estimators for $\mu_r(x)$ make use of the following set of first-stage estimators:

Standard arguments for kernel estimation (Silverman, Section 4.3) show that under Assumption A.3 and the bandwidth restrictions given in Table 2, these estimators converge in probability to the given limits, and with bandwidths that shrink to zero at the optimal λ rate or faster, the deviations of the estimators from their limits, normalized by $\left(n\lambda^d\right)^{1/2}$, converge to Gaussian processes, with associated MLE converging at $\left(n\lambda^d\right)^{-1}$ rates. For example, the estimator $\widehat{A}$ with an optimal rate $\lambda \propto n^{-1/(d+5)}$ has a MSE converging to zero at the rate $n^{-4/(d+5)}$.

The estimators $\widehat{G}$ and $\widehat{g}$ are useful when the function h is unknown, while $\widetilde{G}$ can be used when h is known from the experimental design. The estimators $\widehat{G}$ and $\widetilde{G}$ are not guaranteed to be monotone non-increasing. They can be modified to satisfy this condition using either a "pool adjacent violators" algorithm (e.g., Dinse and Lagakos, 1982) or with probability weights on the terms in $\widehat{A}$ chosen to minimize distance from uniform weights, subject to the monotonicity constraint (Hall and Huang, 2001). These modifications will not alter the asymptotic behavior of $\widehat{G}$ and $\widetilde{G}$, or necessarily improve

Estimator	**Formula**	**Restrictions**	**Optimal λ rate**	**Limit**
$\widehat{D}(v,x)$	$E_n \frac{1}{\lambda^{d+1}} K_1\left(\frac{V-v}{\lambda}\right) K_2\left(\frac{Z-z}{\lambda}\right) 1(Q=q)$	$\lambda \to 0$ $n\lambda^{d+1} \to \infty$	$n^{-1/(d+5)}$	$h(v\|x)p_2(z\|q)p_1(q)$
$\widehat{C}(x)$	$E_n \frac{1}{\lambda^{d}} K_2\left(\frac{Z-z}{\lambda}\right) 1(Q=q)$	$\lambda \to 0$ $n\lambda^{d} \to \infty$	$n^{-1/(d+4)}$	$p_2(z\|q)p_1(q)$
$\widehat{A}(v,x)$	$E_n \frac{Y}{\lambda^{d+1}} K_1\left(\frac{V-v}{\lambda}\right) K_2\left(\frac{Z-z}{\lambda}\right) 1(Q=q)$	$\lambda \to 0$ $n\lambda^{d+1} \to \infty$	$n^{-1/(d+5)}$	$G(v\|x)h(v\|x)p_2(z\|q)p_1(q)$
$\widehat{B}(v,x)$	$E_n \frac{1}{\lambda^{d+2}} K_1'\left(\frac{V-v}{\lambda}\right) K_2\left(\frac{Z-z}{\lambda}\right) 1(Q=q)$	$\lambda \to 0$ $n\lambda^{d+2} \to \infty$	$n^{-1/(d+6)}$	$g(v\|x)h(v\|x)p_2(z\|q)p_1(q)$
$\widehat{G}(v\|x)$	$\widehat{A}(v,x)/\widehat{D}(v,x)$		$n^{-1/(d+5)}$	$G(v\|x)$
$\widetilde{G}(v\|x)$	$\widehat{A}(v,x)/\left(h(v\|x)\widehat{C}(x)\right)$		$n^{-1/(d+5)}$	$G(v\|x)$
$\widehat{g}(v\|x)$	$\widehat{B}(v,x)/\widehat{D}(v,x)$		$n^{-1/(d+6)}$	$g(v\|x)$

Table 2: First Stage Estimators

finite-sample properties of functionals of this estimator, but they do simplify computation of statistics such as conditional quantiles. Another approach, the generalization by Beran (1981) of the Kaplan-Meier product limit estimator to conditional distributions, achieves monotonicity, but has more complex asymptotic behavior and achieves no better rate than the kernel based estimators $\widehat{G}$ or $\widetilde{G}$.

Now consider estimation of $\mu_r(x)$ for a piecewise regular function $r(w,x)$. Plugging $\widehat{g}$ into the definition of this moment in equation (2) yields the second-step estimator

$$\widehat{\mu}_{0r}(x) = \int_{\rho_0(x)}^{\rho_1(x)} r(v,x)\widehat{g}(v \mid x)dv \tag{4}$$

with evaluation requiring numerical integration that contributes an additional error that can be made asymptotically negligible. This estimator will inherit the optimal MSE rate $n^{-4/(d+6)}$ of $\widehat{g}$. Relative to $\widehat{\mu}_{0r}$, we now define the estimators, suitable for varying information sets, listed in Table 1.

The estimator $\widehat{\mu}_{2r}$ is obtained by plugging $\widehat{G}$ into equation (3) for some researcher chosen continuous function $\kappa(x)$ having $\rho_0(x) < \kappa(x) < \rho_1(x)$. This gives

$$\widehat{\mu}_{2r}(x) = r(\kappa(x),x) + \int_{\rho_0(x)}^{\rho_1(x)} r'(v,x)\frac{\widehat{G}(v \mid x) - 1(v < \kappa(x))}{f(v \mid x)}F(dv \mid x,n) \tag{5}$$

where equation (5), compared to equation (3), makes the required numerical integration explicit by introducing a researcher-chosen positive density $f(v \mid x)$ and associated CDF $F(v \mid x)$ with a support that contains $[\rho_0(x), \rho_1(x)]$ and chosen CDF $F(v \mid x,n)$ with finite support for each n such

that $\sup_v n^{1/2}|F(v \mid x,n) - F(v \mid x)|$ is stochastically bounded. The functions $f(v \mid x)$ and $\kappa(x)$ can be chosen for computational convenience or to limit variation in the integrand. The estimator $\widehat{\mu}_{2r}$ is superior to the base case estimator $\widehat{\mu}_{0r}$ in the sense that $\widehat{\mu}_{2r}$ inherits the optimal rate $n^{-4/(d+5)}$ of $\widehat{G}$, which is better than the optimal rate of $\widehat{\mu}_{0r}$.

When the limiting design density $h(v \mid x)$ is known, plugging $\widetilde{G}$ into equation (3), reversing the order of integration and empirical expectation, and substituting asymptotic limits yields the estimator

$$\widehat{\mu}_{1r}(x) = r(\kappa(x), x) + \frac{1}{\widehat{C}(x)} E_n r'(V,X) \frac{Y - 1(V < \kappa(X))}{h(V \mid X)} \lambda^{-d} K_2\left(\frac{Z-z}{\lambda}\right) 1(Q = q) \tag{6}$$

which requires no numerical integration and converges with an optimal MSE rate of $n^{-4/(d+4)}$. This is a specific case of the general principle that when a kernel estimator is plugged into a smooth functional, a better asymptotic rate of convergence can be achieved by undersmoothing; see Goldstein and Messer (1992).

In addition, if $d = 0$, corresponding to moments that are unconditional or conditioned only on one of the finite configurations of Q, then equation (6) reduces to

$$\widehat{\mu}_{1r}(q) = r(\kappa(x), x) + \frac{1}{E_n 1(Q = q)} E_n r'(V,X) \frac{Y - 1(V < \kappa(X))}{h(V \mid X)} 1(Q = q) \tag{7}$$

which requires no kernel smoothing, and is root-n consistent and asymptotically normal. The estimation problem in this case is related to that of estimating unconditional survival curves from current status data; see Jewell and van der Laan (2002) and Gromeboom, Maathuisi, and Wellner (2008).

The properties of the estimators introduced above are summarized in Theorem 1 below. The derivation of explicit variance formulas is deferred to later, when we generalize these results to allow for a larger class of nonparametric smoothers.

THEOREM 1. *Suppose Assumptions A.1- A.3 hold. Suppose K_1 and K_2 are kernel functions on $\mathbb{R}$ and $\mathbb{R}^d$ respectively that are each symmetric, compactly supported continuously differentiable densities. Then, the second-step estimator $\widehat{\mu}_{0r}(x)$ in equation equation (4), using the first-step estimator $\widehat{g}(v \mid x)$ with a bandwidth λ proportional to $n^{-1/(d+6)}$ is consistent and $n^{2/(d+6)}[\widehat{\mu}_{0r}(x) - \mu_r(x)]$ is asymptotically normal. The second-step estimator $\widehat{\mu}_{2r}(x)$ in equation (5) using the first-step estimator $\widehat{G}(v \mid x)$ with a bandwidth λ proportional to $n^{-1/(d+5)}$ is consistent and $n^{2/(d+5)}[\widehat{\mu}_{2r}(x) - \mu_r(x)]$ is asymptotically normal. When $h(v \mid x)$ is known, the second step estimator $\widehat{\mu}_{1r}(x)$ in equation (6) using the first-step estimator $\widehat{C}$ with a bandwidth λ proportional to $n^{-1/(d+4)}$ is consistent and $n^{2/(d+4)}[\widehat{\mu}_{1r}(x) - \mu_r(x)]$ is asymptotically normal. When $h(v \mid x)$ is known and the moment $\mu_r(x)$ is unconditional or is conditioned only on the discrete configuration Q, then the estimator $\widehat{\mu}_{1r}(q)$ in equation (7) is consistent and root-n asymptotically normal.*

PROOF OF THEOREM 1. We first note that Assumptions A1 and A2 suffice to make equations (2) and (3) hold. Also, verification of the asymptotic properties of the kernel estimators in Table 2 given our assumptions is standard. Rewrite $\widehat{\mu}_{2r}(x)$ in equation (5) as

$$\begin{aligned}\widehat{\mu}_{2r}(x) &= r(\kappa(x), x) + \int_{\rho_0(x)}^{\rho_1(x)} r'(v,x)\left(\widehat{G}(v \mid x) - 1(v < \kappa(x))\right)\left(\frac{F(dv \mid x, n)}{f(v \mid x)} - dv\right) \\ &\quad + \int_{\rho_0(x)}^{\rho_1(x)} r'(v,x)\left(\widehat{G}(v \mid x) - G(v \mid x)\right) dv\end{aligned}$$

Consider the terms on the right-hand-side of this expression, normalized by $\left(n\lambda^{d+1}\right)^{1/2}$. The first integral converges to zero since $r'(v,x)$ is uniformly bounded, and by construction $1/f(v \mid x)$ is uniformly bounded and $\sup_v n^{1/2}|F(v \mid x, n) - F(v \mid x)|$ is stochastically bounded. The final integral converges to a normal variate as it is a smooth bounded linear functional of a gaussian process plus an asymptotically negligible process.

A similar decomposition establishes that $\widehat{\mu}_{0r}(x)$ in equation (4), normalized by $\left(n\lambda^{d+2}\right)^{1/2}$ and adapted computationally with a numerical integration procedure with the same properties as that employed in (11) also converges to a normal variate.

Next, define $S(Y,V,X) = r'(V,X)[Y - 1(V < \kappa(X))]/h(V \mid X)$ and let

$$\psi(X) = E_{Y,V|X}\ S(Y,V,X) = \int_{\rho_0(X)}^{\rho_1(X)} r'(V,X)\left(G(V \mid X) - 1(V < \kappa(X))\right) dV.$$

Define $\omega(X,x,\lambda) = \lambda^{-d} K_2\left(\frac{Z-z}{\lambda}\right) 1(Q = q)$. Then $\widehat{\mu}_{1r}(x)$ in equation (6) can be rewritten as

$$\begin{aligned}\widehat{\mu}_{1r}(x) - \mu_r(x) &= \frac{E_n S(Y,V,X)\omega(X,x,\lambda) - E_X\psi(X)\,\omega(X,x,\lambda) - \psi(x)\left(\omega(X,x,\lambda) - E_X\omega(X,x,\lambda)\right)}{\widehat{C}(x)} \\ &\quad + \frac{E_X\left(\psi(X) - \psi(x)\right)\omega(X,x,\lambda)}{\widehat{C}(x)}.\end{aligned}$$

The denominator $\widehat{C}(x)$ converges in probability to $p_2(z \mid q)p_1(q) > 0$. The numerator of the first term, normalized by $\left(n\lambda^d\right)^{1/2}$, is a sum of a triangular array of independent bounded, mean zero random variables, and the variance of the numerator converges to a positive constant. Then a central limit theorem for triangular arrays (e.g., Pollard, 1984, p. 170-174) establishes that the numerator converges to a normal variate. The numerator of the second term with the transformation $Z = z + \lambda t$, becomes

$$\begin{aligned}E_X\left(\psi(X) - \psi(x)\right)\omega(X,x,\lambda) &= \int_Z \left(\psi(Z,q) - \psi(z,q)\right)\lambda^{-d} K_2\left(\frac{Z-z}{\lambda}\right) p_2(z \mid q)p_1(q)dZ \\ &= \int_t \left(\psi(z+\lambda t, q) - \psi(z,q)\right) K_2(t)\, p_2(z+\lambda t \mid q)p_1(q)dt.\end{aligned}$$

Assumptions A.1 and A.2 imply that $\psi(z,q)$ and $p_2(z \mid q)$ are Lipschitz-continuous in z, so that the last expression scaled by $\left(n\lambda^d\right)^{1/2}$ converges to $\left(n\lambda^{d+4}\right)^{1/2} c \int_t t^2 K_2(t)\, dt$ for a constant c. At the optimal rate $\lambda \propto n^{-1/(d+4)}$ this numerator then converges to a constant. This establishes that the estimator $\widehat{\mu}_{1r}(x)$ in equation (6) converges with an optimal MSE rate of $n^{-4/(d+4)}$. When $d = 0$, this gives equation (7) with a conventional MSE rate of n^{-1}. ■

In the special case of the nonparametric location model $W = \Lambda[m(X) - \varepsilon]$ with $\varepsilon \perp X$, and Λ known and invertible, these $\mu_r(x)$ estimators can be used to estimate an unknown $m(x)$, since $m(x) = \mu_r(x) - E(\varepsilon)$ with $r(w,x) = \Lambda^{-1}(w)$. Chen and Randall (1997) and Crooker and Herriges (2004) consider this case. An (2000) considers the model where Λ is unknown but m and the distribution of ε are known; this also is a special case of our nonparametric model.

2.3 Semiparametric Moments

Corollary 1 below will be used in place of Theorem 1 to obtain faster convergence rates using a semiparametric model for W. To simplify the analysis we take $\kappa = 0$ (or equivalently, we absorb κ into the definition of Λ) so when applying these results one could first recenter (e.g., demean) V, and adjust the definition of Λ accordingly.

ASSUMPTION A.4. *The latent W satisfies $W = \Lambda[m(X,\theta_0) - \varepsilon]$, where m and Λ are known functions, Λ is invertible and differentiable with derivative denoted Λ', $\theta_0 \in \Theta$ is a vector of parameters, and ε is a disturbance that is distributed independently of V, X, with unknown, twice continuously differentiable CDF $F_\varepsilon(\varepsilon)$ and compact support $[a_0, a_1]$ that contains zero. Define $U = m(X,\theta_0) - \Lambda^{-1}(V)$. Let $\Psi_n(U \mid n)$ denote the empirical CDF of U at sample size n. $\sup_U |\Psi_n(U \mid n) - \Psi(U)| \to 0$ a.s., where $\Psi(U)$ is a CDF that has an associated PDF $\psi(U)$ that is continuous and strictly positive on the interval $[a_0, a_1]$.*

Define $s_r^*(x,u,y)$ and $t_r^*(x,u)$ by

$$s_r^*(x,u,y) = r[\Lambda(m(x,\theta_0)),x] + \frac{r'[\Lambda(m(x,\theta_0)-u),x]\Lambda'(m(x,\theta_0)-u)[y - 1(u>0)]}{\psi(u)}.$$

$$t_r^*(x,u) = \frac{r'[\Lambda(m(x,\theta_0)-u),x]\Lambda'(m(x,\theta_0)-u)[F_\varepsilon(u) - 1(u>0)]}{\psi(u)}.$$

If Λ is the identity function, then W equals a parameterized function of x plus an additive independent error. If Λ is the exponential function, then it is $\ln(W)$ that is modeled with an additive error. Other examples include: the Box-Cox, $\Lambda^{-1}(W) = (W^\lambda - 1)/\lambda$, the Zellner-Revankar $\Lambda^{-1}(W) = \ln W + \lambda W$, and the arcsinh $\Lambda^{-1}(W) = \sinh^{-1}(\lambda W)/\lambda$, where in each case λ is a free parameter.

COROLLARY 1. *Let Assumptions A.1, A.2, and A.4 hold. Then*

$$E(Y \mid U = u) = F_\varepsilon(u)$$

$$\mu_r(x) = r[\Lambda(m(x,\theta_0)), x] + \int_{a_0}^{a_1} t_r^*(x,u)\Psi(du),$$

$$\mu_r(x) = E\left[s_r^*(x,U,Y) \mid n\right] + \int_{a_0}^{a_1} t_r^*(x,u)[\Psi(du) - \Psi(du \mid n)] = \lim_{n\to\infty} E\left[s_r^*(x,U,Y) \mid n\right]$$

and, if Assumption A.3 also holds:

$$\Psi_n(u \mid n) = E(1 - H[\Lambda(m(X,\theta_0) - u) \mid X, n])$$

$$\psi_n(u) = E\left[h[\Lambda(m(X,\theta_0) - u) \mid X, n]\Lambda'(m(X,\theta_0) - u) \mid n\right] \to \psi(u),$$

where expectation is with respect to $H(v \mid x, n)$.

PROOF OF COROLLARY 1. Recall that $Y = I(W > V) = I(\varepsilon < U)$, so $E(Y \mid U = u) = F_\varepsilon(u)$.

Starting from the definition of $\mu_r(x)$,

$$\begin{aligned}
\mu_r(x) &= \int_{a_0}^{a_1} r[\Lambda(m(x,\theta_0) - \varepsilon), x]F_\varepsilon(d\varepsilon) \\
&= \int_{a_0}^{0} r[\Lambda(m(x,\theta_0) - u), x]\frac{dF_\varepsilon(u)}{du}du + \int_{0}^{a_1} r[\Lambda(m(x,\theta_0) - u), x]\frac{d[F_\varepsilon(u) - 1]}{du}du
\end{aligned}$$

and applying integration by parts to each of the above integrals yields

$$\begin{aligned}
\mu_r(x) &= r[\Lambda(m(x,\theta_0)), x] + \int_{a_0}^{a_1} r'[\Lambda(m(x,\theta_0) - u), x]\Lambda'(m(x,\theta_0) - u)[F_\varepsilon(u) - I(u > 0)]du \\
&= r[\Lambda(m(x,\theta_0)), x] + \int_{a_0}^{a_1} t_r^*(x,u)\Psi(du) \\
&= r[\Lambda(m(x,\theta_0)), x] + \int_{a_0}^{a_1} t_r^*(x,u)\Psi_n(du \mid n) + \int_{a_0}^{a_1} t_r^*(x,u)[\Psi(du) - \Psi_n(du \mid n)]
\end{aligned}$$

Next, apply the law of iterated expectations to obtain

$$\begin{aligned}
E\left[s_r^*(x,U,Y)\right] &= r[\Lambda(m(x,\theta_0)), x] \\
&\quad + E\left(\frac{r'[\Lambda(m(x,\theta_0) - U), x]\Lambda'(m(x,\theta_0) - U)[F_\varepsilon(u) - 1(U > 0)]}{\psi(u)}\right) \\
&= r[\Lambda(m(x,\theta_0)), x] + \int_{a_0}^{a_1} t_r^*(x,u)\Psi_n(du \mid n),
\end{aligned}$$

which gives the expressions for $\mu_r(x)$, and $\int_{a_0}^{a_1} t_r^*(x,u)[\Psi(du) - \Psi_n(du \mid n)] \to_p 0$ by the uniform convergence of Ψ_n.

Note that $\Psi_n(u \mid n)$ is the empirical probability that $U \leq u$, which is the same event as $V \geq \Lambda(m(X,\theta_0) - u)$. Conditioning on $X = x$ this probability would be $1 - H_n[\Lambda(m(x,\theta_0) - u) \mid x, n]$, and averaging over X gives $\Psi_n(u \mid n) = E(1 - H_n[\Lambda(m(X,\theta_0) - u) \mid X, n])$. This implies $\Psi(u) = \lim_{n\to\infty} E(1 - H_n[\Lambda(m(X,\theta_0) - u) \mid X, n])$, where the only role of the limit is to evaluate the expectation at the limiting distribution of X. Taking the derivative with respect to u gives $\psi(u) = \lim_{n\to\infty} E(h[\Lambda(m(X,\theta_0) - u \mid X)\Lambda'(m(X,\theta_0) - u)])$. Consistency of $\psi_n(u)$ then follows from the uniform convergence of the distribution of X to its limiting distribution in Assumption A.3. ■

Now consider rate root n estimation of arbitrary conditional moments based on Corollary 1. It will be convenient to first consider the case where θ_0 is known, implying that the conditional mean of W is known up to an arbitrary location (since ε is not required to have mean zero). A special case of known θ_0 is when x is empty, i.e., estimation of unconditional moments of W, since in that case we can without loss of generality take m to equal zero.

2.3.1 Estimation With Known θ

Suppose that θ_0 is known. Considering first the case where the limiting design density $h(v|x)$ is also known, for a given u define the sample average $\widehat{\psi}(u)$ by

$$\widehat{\psi}(u) = \frac{1}{n}\sum_{i=1}^{n} h[\Lambda(m(X_i,\theta_0) - u) \mid X_i]\Lambda'(m(X_i,\theta_0) - u).$$

Then, based on Corollary 1, we have consistency of the estimator

$$\widehat{\mu}^*_{3r}(x) = r[\Lambda(m(x,\theta_0)), x] + \frac{1}{n}\sum_{i=1}^{n} \frac{r'[\Lambda(m(x,\theta_0) - U_i), x]\Lambda'(m(x,\theta_0) - U_i)[Y_i - 1(U_i > 0)]}{\widehat{\psi}(U_i)}.$$

This estimator is computationally extremely simple, since it entails only sample averages. Special cases of this estimator were proposed by McFadden (1994) and by Lewbel (1997).

Let $\widetilde{\psi}(u)$ be an estimator of $\psi(u)$ that does not depend on knowledge of h. For example $\widetilde{\psi}(u)$ could be a (one dimensional) kernel density estimator of the density of U, based on the data $\widehat{U}_i = m(X_i,\theta_0) - \Lambda^{-1}(V_i)$ and evaluated at u. We then have the estimator

$$\widehat{\mu}^*_{4r}(x) = r[\Lambda(m(x,\theta_0)), x] + \frac{1}{n}\sum_{i=1}^{n} \frac{r'[\Lambda(m(x,\theta_0) - U_i), x]\Lambda'(m(x,\theta_0) - U_i)[Y_i - 1(U_i > 0)]}{\widetilde{\psi}(U_i)},$$

which may be used when h is unknown.

2.3.2 Estimation of θ

First, consider estimation of θ. By Assumption A.4,

$$E[\Lambda^{-1}(W) \mid X = x] = \alpha_0 + m(x,\theta_0)$$

for some arbitrary location constant α_0. This constant is unknown since no location constraint is imposed upon ε. Let $s_{\Lambda^{-1}}(X,V,Y)$ denote $s_r(X,V,Y)$ with $r(w,x)=\Lambda^{-1}(w)$. It then follows from Theorem 1 that

$$\lim_{n\to\infty} E\left[s_{\Lambda^{-1}}(X,V,Y) \mid X=x\right] = \lim_{n\to\infty} E(\Lambda^{-1}(W) \mid X=x).$$

Note that the limit as $n\to\infty$ means that the expectations are taken at the limiting distributions of the data. In other words the asymptotic conditional expectation of the known or estimable quantity $s_{\Lambda^{-1}}$ is equal to $\alpha_0 + m(x,\theta_0)$. Under some identification conditions this can be used for estimation of (α_0,θ_0). Specifically, we could estimate θ_0 by minimizing the least squares criterion

$$(\widehat{\theta},\widehat{\alpha}) = \arg\min_{\theta,\alpha} \frac{1}{n}\sum_{i=1}^{n}[s_{\Lambda^{-1}}(X_i,V_i,Y_i) - \alpha - m(X_i,\theta)]^2. \tag{8}$$

If m is linear in parameters, then a closed form expression results for both parameter estimates. If h is not known, one could replace $h(V \mid X)$ in the expression of $s_{\Lambda^{-1}}(X,V,Y)$ with an estimate $\widehat{h}(V \mid X)$. The resulting estimator would then take the form of a two step estimator with a nonparametric first step (the estimation of h). This estimator of θ and α is equivalent to the estimator for general binary choice models proposed by Lewbel (2000), though Lewbel provides other extensions, such as to estimation with endogenous regressors.

With Assumption A.4, the latent error ε is independent of X, and therefore the binary choice estimator of Klein and Spady (1993) may provide a semiparametrically efficient estimator of θ.[7]

2.3.3 Estimation with Unknown θ

Let $\widehat{\theta}$ denote a root n consistent, asymptotically normal estimator for θ_0. Replacing θ_0 with any $\theta\in\Theta$ we may rewrite the estimators of the previous section as $\widehat{\mu}^*_{\lambda r}(x;\theta)$ for $\lambda = 3$ or 4. In doing so, note that θ appears both directly in the equations for $\widehat{\mu}^*_{\lambda r}$, and also in the definition of $U_i = m(X_i,\theta) - \Lambda^{-1}(V_i)$. We later derive the root n consistent, asymptotically normal limiting distribution for each estimator $\widehat{\mu}_{\lambda r}(x) = \widehat{\mu}^*_{\lambda r}(x;\widehat{\theta})$, where we suppress the dependence on $\widehat{\theta}$ for simplicity. The estimators are not differentiable in U_i, which complicates the derivation of their limiting distribution, e.g., even with a fixed design, Theorem 6.1 of Newey and McFadden (1994) is not be directly applicable due to this nondifferentiability.

[7] The Klein and Spady estimator does not identify a location constant α, but that is not required for this step, since no location constraint is imposed upon ε. Also, for the present application, the limiting distribution theory for Klein and Spady would need to be extended to allow for data generating processes that vary with the sample size.

3 Estimation Details and Distribution Theory

In this section we provide more detail about the computation of the estimators $\widehat{\mu}_{0r}(x), \ldots, \widehat{\mu}_{4r}(x)$ and their distribution theory. Earlier we focused on ordinary kernel regression based estimators for ease of exposition, but in this section we allow for more general classes of estimators.

3.1 Nonparametric Estimators

There are many different nonparametric methods for estimating regression functions. For purely continuous variables with density bounded away from zero throughout their support the local linear kernel method is attractive. This method has been extensively analyzed and has some positive properties like being design adaptive, and best linear minimax under standard conditions; see Fan and Gijbels (1996) for further discussion. One issue we are particularly concerned about is how to handle discrete variables. Specifically, some elements of X could be discrete, either ordered discrete or unordered discrete, while V can be ordered discrete. When there is a single discrete variable that takes only a small number of values, the pure frequency estimator is the natural and indeed optimal estimator to take in the absence of additional structure. In fact, one obtains parametric rates of convergence in the pure discrete case [and in the mixed discrete/continuous case the rate of consistency is unaffected by how many such discrete covariates there are], see Delgado and Mora (1995) for discussion. When there are many discrete covariates, it may be desirable to use some 'discrete smoothing', as discussed in Li and Racine (2004), see also Wang and Van Ryzin (1981). Coppejans (2003) considers a case most similar to our own - he allows the distribution of the discrete data to change with sample size. One major difference is that his data have arrived from a very specific grouping scheme that introduces an extra bias problem.

We shall not outline all the possibilities for estimation here with regard to the covariates X, rather we assume that X is continuously distributed with density bounded away from zero. However, the estimators we define can be applied in all of the above situations [although they may not be optimal], and the estimators are still asymptotically normal with the rate determined by the number of continuous variables.

We will pay more attention to the potential discreteness in V, since this is key to our estimation problem. For clarity we will avoid excessive subscripts/superscripts. We suppose that V is asymptotically continuous in the sense that for each n, V_i is drawn from a distribution $H(v|X_i, n)$ that has finite support, increasing with n. The case where V_i is drawn from a continuous distribution $H(v|X_i)$ for all n is really a special case of our set-up.

Under our conditions there is a bias in the estimates of $\mu_r(x)$ of order J^{-1} in this discrete case. Therefore, for this term not to matter in the limiting distribution we require that $\delta_n J^{-1} \to 0$,

where δ_n is the rate of convergence of the estimator in question [$\delta_n = \sqrt{n}$ in the parametric case but $\delta_n = \sqrt{nb^d}$ for some bandwidth b in the nonparametric cases]. In the nonparametric case, the spacing of the discrete covariates is closer than the bandwidth of a standard kernel estimator, that is, we know that $b^2 J \to \infty$ so that J^{-1} is much smaller than the smoothing window of a kernel estimator. Therefore, the pure frequency estimator is dominated by a smoothing estimator, and we shall just construct smoothing-based estimators.

The estimator $\widehat{\mu}_{1r}(x)$ involves smoothing the data

$$s_r(Z_i) = r[\kappa(X_i), X_i] + \frac{r'(V_i, X_i)[Y_i - 1(V_i < \kappa(X_i))]}{h(V_i \mid X_i)}$$

against X_i, where $Z_i = (V_i, X_i, Y_i)$. Define the $p-1$-th order local polynomial regression of $s_r(Z_i)$ on X_i by minimizing

$$Q^s_{p-1,n}(\vartheta) = \frac{1}{n}\sum_{i=1}^{n} K_b(X_i - x)\left[s_r(Z_i) - \sum_{0 \le |\mathbf{j}| \le p-1} \vartheta_{\mathbf{j}} (X_i - x)^{\mathbf{j}}\right]^2 \tag{9}$$

with respect to the vector ϑ containing all the $\vartheta_{\mathbf{j}}$, where $K_b(t) = \prod_{j=1}^{d} k_b(t_j)$ with $k_b(u) = k(u/b)/b$, where k is a univariate kernel function and $b = b(n)$ is a bandwidth. Here, we are using the multi-dimensional index notation, for vectors $\mathbf{j} = (j_1, \ldots, j_d)^\top$ and $a = (a_1, \ldots, a_d)^\top$: $\mathbf{j}! = j_1! \times \cdots \times j_d!$, $|\mathbf{j}| = \sum_{k=1}^{d} j_k$, $a^{\mathbf{j}} = a_1^{j_1} \times \ldots \times a_d^{j_d}$, and $\sum_{0 \le |\mathbf{j}| \le p-1}$ denotes the sum over all $\mathbf{j}$ with $0 \le |\mathbf{j}| \le p-1$. Let $\widehat{\vartheta}_0$ denote the first element of the vector $\widehat{\vartheta}$ that minimizes (9). Then let

$$\widehat{\mu}_{1r}(x) = \widehat{\vartheta}_0. \tag{10}$$

This estimator is linear in the dependent variable and has an explicit form.

In computing the estimator $\widehat{\mu}_{2r}(x)$ we require an estimator of $G(v \mid x)$, which is given by the smooth of Y_i on X_i, V_i. Let $\widetilde{X}_i = \left(V_i, X_i^\top\right)^\top$ and $\widetilde{x} = (v, x)^\top$ and define the $p-1$-th order local polynomial regression of Y_i on $\widetilde{X}_i$ by minimizing

$$Q^Y_{p-1,n}(\vartheta) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{K}_b(\widetilde{X}_i - \widetilde{x})\left[Y_i - \sum_{0 \le |\mathbf{j}| \le p-1} \vartheta_{\mathbf{j}} \left(\widetilde{X}_i - \widetilde{x}\right)^{\mathbf{j}}\right]^2, \tag{11}$$

where $\widetilde{K}_b(\widetilde{X}_i - \widetilde{x}) = k_b(V_i - v) K_b(X_i - x)$. Let $\widehat{\vartheta}_0$ denote the first element of the vector $\widehat{\vartheta}$ that minimizes (11), and let $\widehat{G}(v \mid x) = \widehat{\vartheta}_0$. Then define

$$\widehat{\mu}_{2r}(x) = r(\kappa(x), x) + \int_a^{a_1} r'(v, x)[\widehat{G}(v \mid x) - 1(v < \kappa(x))]dv, \tag{12}$$

where the univariate integral is interpreted in the Lebesgue Stieltjes sense (actually under our conditions $\widehat{G}(v \mid x)$ is a continuous function and $1(v < \kappa(x))$ is a simple step function).

Finally, to compute $\widehat{\mu}_{0r}(x)$ we use one higher order of polynomial, i.e., minimize $Q_{p,n}^{Y}(\vartheta)$ with respect to ϑ, and let $\partial\widehat{G}(v \mid x)/\partial v = \widehat{\vartheta}_v$, where $\widehat{\vartheta}_v$ is the second element of the vector $\widehat{\vartheta}$. Then define

$$\widehat{\mu}_{0r}(x) = -\int_{a_0}^{a_1} r(v,x)\frac{\partial\widehat{G}(v \mid x)}{\partial v}dv. \tag{13}$$

The estimator (12) is in the class of marginal integration/partial mean estimators sometimes used for estimating additive nonparametric regression models, see Linton and Nielsen (1995), except that the integrand is not just a regression function and the integrating measure λ, where (asymptotically) $d\lambda(v) = r'(v,x)1(\rho_0(x) \leq v \leq \rho_1(x))dv$, is not necessarily a probability measure, i.e., it may not be positive or integrate to one. The distribution theory for the class of marginal integration estimators has already been worked out for a number of specific smoothing methods when the covariate distribution is absolutely continuous, see the above references.

We make the following assumptions.

ASSUMPTION B.1. *k is a symmetric probability density with bounded support, and is continuously differentiable on its support.*

ASSUMPTION B.2. *The random variables (V,X) are asymptotically continuously distributed, i.e., for some finite constant c_h*

$$\sup_{v\in[\rho_0(x),\rho_1(x)]} |H(v|x,n) - H(v|x)| \leq \frac{c_h}{J}, \tag{14}$$

where $H(v,x)$ possesses a Lebesgue density $h(v,x)$ along with conditionals $h(v|x)$ and marginal $h(x)$. Furthermore, $\inf_{\rho_0(x)\leq v\leq\rho_1(x)} h(v,x) > 0$. The conditional variance of the limiting continuous distribution is equal to the limiting conditional variance $\sigma^2(v,x) = \lim_{n\to\infty}\mathrm{var}(Y \mid V = v, X = x) = G(v \mid x)[1 - G(v \mid x)]$. Furthermore, $G(v \mid x)$ and $h(v,x)$ are p-times continuously differentiable for all v with $\rho_0(x) \leq v \leq \rho_1(x)$, letting $g(v \mid x) = \partial G(v|x)/\partial v$ denote the conditional density of $W|X$. The set $[\rho_0(x),\rho_1(x)] \times \{x\}$ is strictly contained in the support of (V,X) for large enough n.

The condition (14) is satisfied provided the associated frequency function $h(v \mid x,n)$ satisfies $\min_{v\in\mathcal{J}_n} h(v \mid x,n) \geq \underline{v}/J_n$ and $\max_{v\in\mathcal{J}_n} h(v \mid x,n) \leq \overline{v}/J_n$ for some bounds $\underline{v} > 0$ and $\overline{v} < \infty$, and provided the support $\mathcal{J}_n$ becomes dense in $[\rho_0(x),\rho_1(x)]$. The other conditions are standard regularity conditions for nonparametric estimation.

For a function $f : \mathbb{R}^s \to \mathbb{R}$, arrange the elements of its partial derivatives $\partial^{\sum_{j=1}^s \pi_j} f(t)/\partial t_1^{\pi_1} \cdots \partial t_s^{\pi_s}$ (for all vectors $\pi = (\pi_1,\ldots,\pi_s)$ such that $\sum_{j=1}^s \pi_j = p$) as a large column vector $f^{(p;s)}(t)$ of dimensions $(s+p-1)!/s!(p-1)!$. Let $a_{d;p}(k)$, $\overline{a}_{d+1;p}(k)$ and $a^*_{d+1;p+1}(k)$ be conformable vectors of constants depending only on the kernel k, and let $c_{d;p}(k)$, $\overline{c}_{d+1;p}(k)$ and $c^*_{d+1;p+1}(k)$ also be constants only

depending on the kernels. Define

$$\beta_0(x) = a^*_{d+1;p+1}(k)^\top \int_{\rho_0(x)}^{\rho_1(x)} r(v,x) G^{(p+1;d+1)}(v|x) dv$$

$$\beta_1(x) = a_{d;p}(k)^\top \mu_r^{(p;d)}(x) \; ; \; \beta_2(x) = \overline{a}_{d+1;p}(k)^\top \int_{\rho_0(x)}^{\rho_1(x)} r'(v,x) G^{(p;d+1)}(v|x) dv,$$

$$\omega_0(x) = c^*_{d+1;p+1}(k) \int_{\rho_0(x)}^{\rho_1(x)} \sigma^2(v,x) \left(\frac{r'(v,x)h(v,x) - r(v,x)h'(v,x)}{h^2(v,x)} \right)^2 h(v,x) dv$$

$$\omega_1(x) = c_{d;p}(k) \frac{\mathrm{var}[s_r(Z) \mid X = x]}{h(x)} \; ; \; \omega_2(x) = \overline{c}_{d+1;p}(k) \int_{\rho_0(x)}^{\rho_1(x)} \sigma^2(v,x) \left(\frac{r'(v,x)}{h(v,x)} \right)^2 h(v,x) dv.$$

THEOREM 2. *Suppose that assumptions A1-A3, B1 and B2 hold and that the bandwidth sequence* $b = b(n)$ *satisfies* $b \to 0$, $nb^{d+2}/\log n \to \infty$, *and* $Jb^2 \to \infty$. *Then, for* $j = 1, 2$,

$$\sqrt{nb^d} \left[\widehat{\mu}_{jr}(x) - \mu_r(x) - b^p \beta_j(x) \right] \Longrightarrow N(0, \omega_j(x)).$$

If G *is* $p+1$*-times continuously differentiable, then*

$$\sqrt{nb^d} \left[\widehat{\mu}_{0r}(x) - \mu_r(x) - b^p \beta_0(x) \right] \Longrightarrow N(0, \omega_0(x)).$$

REMARKS.

1. In the local linear case ($p = 2$) the kernel constants in $\omega_1(x)$ and $\omega_2(x)$ are identical, and equal to $||K||_2^2 = \int K(u)^2 du$. A simple argument then shows that $\omega_1(x) \geq \omega_2(x)$. By the law of iterated expectation

$$\mathrm{var}[s_r(Z) \mid X = x] = E[\mathrm{var}[s_r(Z) \mid V, X] \mid X = x] + \mathrm{var}[E[s_r(Z) \mid V, X] \mid X = x].$$

Furthermore,

$$\begin{aligned} E[\mathrm{var}[s_r(Z) \quad | \quad V, X] \mid X = x] &= \int_{\rho_0(x)}^{\rho_1(x)} \left(\frac{r'(v,x)}{h(v \mid x)} \right)^2 \sigma^2(v,x) h(v \mid x) dv \\ &= h(x) \int_{\rho_0(x)}^{\rho_1(x)} \left(\frac{r'(v,x)}{h(v,x)} \right)^2 \sigma^2(v,x) h(v,x) dv. \end{aligned}$$

It follows that $\omega_1(x) \geq \omega_2(x)$. In the special case that $h'(v,x) = 0$, which would be true if V were uniformly distributed, $\omega_0(x)$ is the same as $\omega_2(x)$ apart from the kernel constants.

2. Regarding the biases, in the special case of local linear estimation and supposing that $g(v|x)$ has two continuous derivatives and the support of $V|X$ does not depend on X, we have:

$$\begin{aligned} \beta_1(x) &\propto \sum_{j=1}^{d} \int_{\rho_0}^{\rho_1} \frac{\partial^2 \{ r(v,x) g(v \mid x) \}}{\partial x_j^2} dv \\ \beta_2(x) &\propto \int_{\rho_0}^{\rho_1} \left[\sum_{j=1}^{d} \frac{\partial^2 G(v \mid x)}{\partial x_j^2} + \frac{\partial^2 G(v \mid x)}{\partial v^2} \right] r'(v,x) dv. \end{aligned}$$

Applying integration by parts shows that these two biases are sometimes the same, depending upon boundary conditions.

3. If $r(v,x)$ is a vector of functions, then the results are as above with the square operation replaced by outer product of corresponding vectors. Suppose one wants to estimate $\text{var}(W|X = x) = \mu_{W^2}(x) - \mu_W^2(x)$, a nonlinear function of the vector $(E(W^2|X = x), E(W|X = x))$. In this case, one obtains the asymptotic distribution by the delta method applied to the joint limiting behaviour of the estimators of $\mu_{W^2}(x), \mu_W(x)$.

4. Standard errors can be constructed by plugging in estimators of unknown quantities in the asymptotic distributions. For the estimator $\widehat{\mu}_1(x)$, standard formulae can be applied as given in Härdle and Linton (1994) and more recently reviewed in Linton and Park (2009). For the integration based estimators $\widehat{\mu}_0(x)$ and $\widehat{\mu}_2(x)$ one can apply the methods described in Sperlich, Linton, and Härdle (1999). For example, $\omega_2(x)$ just requires consistent estimation of $\sigma^2(v,x)$ and $h(v,x)$, a regression function and density function, and this follows from our proofs below. Note that estimation of the bias term in all cases is hard. If desired one could as usual remove the asymptotic bias by using an undersmoothed bandwidth.

5. We have shown that generally $\widehat{\mu}_{2r}(x)$ has smaller mean squared error than $\widehat{\mu}_{1r}(x)$. However, there are other comparisons between the estimators that are also relevant. For example, the estimator $\widehat{\mu}_{1r}(x)$ requires prior knowledge of $h(v \mid x)$. On the other hand $\widehat{\mu}_{1r}(x)$ also uses a lower dimensional smoothing operation than $\widehat{\mu}_{2r}(x)$, which may be important in small samples. An advantage of the estimator $\widehat{\mu}_{1r}(x)$ is that it takes the form of a standard nonparametric regression estimator, so known regression bandwidth selection methods can be automatically applied. Sperlich, Linton, and Härdle (1999) present a comprehensive study of the finite sample performance of marginal integration estimators and discuss bandwidth selection.

6. The term $\kappa(x)$ drops out of these limiting distributions, showing that choice of $\kappa(x)$ is asymptotically irrelevant. For simplicity, $\kappa(x)$ can be taken to be some convenient value in the range of the v data such as its mean.

3.2 Semiparametric Estimators

In this section we assume the conditions of A4 prevail. In this case, discreteness of V_i is less of an issue - even if V_i is discrete, if there are continuous variables in X_i, then $U_i = m(X_i, \theta_0) - \Lambda^{-1}(V_i)$ can be continuously distributed. For simplicity we therefore assume a fixed design for our limiting distribution calculations. Similar asymptotics will result when the assumption that V_i is continuously distributed is replaced by an assumption like equation (14).

Let $\widehat{\theta}$ be some consistent estimator of θ_0. Define:

$$\widehat{\mu}_{3r}(x) = r[\Lambda(m(x,\widehat{\theta})),x] + \frac{1}{n}\sum_{i=1}^{n} \frac{r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\widehat{\psi}(\widehat{U}_i)}$$

$$\widehat{\mu}_{4r}(x) = r[\Lambda(m(x,\widehat{\theta})),x] + \frac{1}{n}\sum_{i=1}^{n} \frac{r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\widetilde{\psi}(\widehat{U}_i)},$$

where $\widehat{U}_i = m(X_i,\widehat{\theta}) - \Lambda^{-1}(V_i)$ and

$$\widehat{\psi}(\widehat{U}_i) = \frac{1}{n}\sum_{j=1}^{n} h[\Lambda(m(X_j,\widehat{\theta})-\widehat{U}_i)|X_j]\Lambda'(m(X_j,\widehat{\theta})-\widehat{U}_i) \quad ; \quad \widetilde{\psi}(\widehat{U}_i) = \frac{1}{nb}\sum_{j=1}^{n} k\left(\frac{\widehat{U}_i - \widehat{U}_j}{b}\right).$$

Define also the estimators $\widehat{\mu}^*_{3r}(x)$ and $\widehat{\mu}^*_{4r}(x)$ as the special cases of $\widehat{\mu}_{3r}(x)$ and $\widehat{\mu}_{4r}(x)$ in which θ is known, in which case $\widehat{U}_i$ is replaced by U_i.

We next state the asymptotic properties of the conditional moment estimators based on Corollary 1. We need some conditions on the estimator and on the regression functions and densities.

ASSUMPTION C.1. *Suppose that*

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varsigma(Z_i,\theta_0) + o_p(1)$$

for some function ς such that $E[\varsigma(Z_i,\theta_0)] = 0$ and $\Omega = E[\varsigma(Z_i,\theta_0)\varsigma(Z_i,\theta_0)^\top] < \infty$. Suppose also that θ_0 is an interior point of the parameter space.

ASSUMPTION C.2. *The function m is twice continuously differentiable in θ and for any $\delta_n \to 0$,*

$$\sup_{\|\theta-\theta_0\|\le\delta_n} \left\| \frac{\partial m}{\partial\theta}(x,\theta) \right\| \le d_1(x) \quad ; \quad \sup_{\|\theta-\theta_0\|\le\delta_n} \left\| \frac{\partial^2 m}{\partial\theta\partial\theta^\top}(x,\theta) \right\| \le d_2(x)$$

with $Ed_1^r(X_i) < \infty$ and $Ed_2^r(X_i) < \infty$ for some $r > 2$.

ASSUMPTION C.3. *The density function h is continuous and is strictly positive on its compact support and is twice continuously differentiable. The transformation Λ is three times continuously differentiable.*

ASSUMPTION C.4. *The kernel k is twice continuously differentiable on its support, and therefore $\sup_t |k''(t)| < \infty$. The bandwidth b satisfies $b \to 0$ and $nb^6 \to \infty$.*

The regularity conditions are quite standard. Assumption C4 is used for $\widehat{\mu}_{4r}(x)$, which is based on a one-dimensional kernel density estimator.

For each $\theta \in \Theta$ and $x \in \mathcal{X}$, define the stochastic processes:

$$f_0(Z_i,\theta) = \frac{r'[\Lambda(m(x,\theta)-U_i(\theta)),x]\Lambda'(m(x,\theta)-U_i(\theta))[Y_i - 1(U_i(\theta) > 0)]}{\psi(U_i)}$$

$$f_1(Z_i,\theta) = r[\Lambda(m(x,\theta)),x] + \frac{r'[\Lambda(m(x,\theta)-U_i(\theta)),x]\Lambda'(m(x,\theta)-U_i(\theta))[Y_i - 1(U_i(\theta) > 0)]}{\psi(U_i)}$$

where $U_i(\theta) = m(X_i, \theta) - \Lambda^{-1}(V_i)$. Then

$$\begin{aligned}
\Gamma_F &= \left(\frac{\partial}{\partial\theta} E\left[f_1(Z_i,\theta)\right]\right)\Big|_{\theta=\theta_0} \\
\Psi_F &= -E\left[f_0(Z_i,\theta_0)\frac{\psi'(U_i)}{\psi(U_i)}\widetilde{\gamma}_i\right] + E\left[\frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\widetilde{\zeta}_{ij}\widetilde{\gamma}_j\right] \\
\widetilde{\gamma}_i &= \frac{\partial m}{\partial\theta^\top}(X_i,\theta_0) - E\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0)\right]
\end{aligned}$$

and $\zeta_{ij} = \left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda''\right](m(X_j,\theta_0) - U_i)$, where $\widetilde{\zeta}_{ij} = \zeta_{ij} - E_i\zeta_{ij}$.

The above quantities may depend on x but we have suppressed this notationally. Note also that $Ef_1(Z_i,\theta_0) = \mu_r(x)$.

THEOREM 3. *Suppose that Assumptions A1-A4 and C1-C3 hold. Then, as* $n \to \infty$,

$$\sqrt{n}[\widehat{\mu}_{3r}(x) - \mu_r(x)] \Longrightarrow N(0, \sigma^2_\eta(x)), \tag{15}$$

where $0 < \sigma^2_\eta(x) = \mathrm{var}(\eta_j) < \infty$ *with* $\eta_j = \eta_{1j} + \eta_{2j} + \eta_{3j}$, *where:*

$$\eta_{1j} = f_0(Z_j,\theta_0) - Ef_0(Z_j,\theta_0)$$

$$\eta_{2j} = (\Gamma_F - \Psi_F)\,\varsigma(Z_j;\theta_0)$$

$$\eta_{3j} = -E\left[f_0(Z_i,\theta_0)\frac{h[\Lambda(m(X_j,\theta_0)-U_i)|X_j]\Lambda'(m(X_j,\theta_0)-U_i) - \psi(U_i)}{\psi(U_i)} \mid X_j\right].$$

The three terms η_{1j}, η_{2j}, and η_{3j} are all mean zero and have finite variance. They are generally mutually correlated. When θ_0 is known, the term $\eta_{2j} = 0$ and this term is missing from the asymptotic expansion. The term η_{3j} is due to the estimation of ψ even when θ_0 is known.

We next give the distribution theory for the semiparametric estimator $\widehat{\mu}_{4r}(x)$. Let

$$\begin{aligned}
\Psi^*_F &= E\left[\frac{\psi'(U_i)}{\psi(U_i)}\left\{f_0(Z_i,\theta_0) - E[f_0(Z_i,\theta_0)|U_i]\right\}\gamma^*_i\right] - E\left[E[f_0(Z_i,\theta_0)|U_i]\overline{m}'_{\theta_0}(U_i)\right] \\
\overline{m}_{\theta_0}(U_i) &= E\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0) \mid U_i\right] \\
\gamma^*_i &= \frac{\partial m}{\partial\theta^\top}(X_i,\theta_0) - E\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0) \mid U_i\right].
\end{aligned}$$

THEOREM 4. *Suppose that assumptions A1-A4, B1, B2 and C1-C4 hold. Then*

$$\sqrt{n}[\widehat{\mu}_{4r}(x) - \mu_r(x)] \Longrightarrow N(0, \sigma^{*2}_\eta(x)),$$

where $0 < \sigma_\eta^{*2}(x) = \text{var}(\eta_j^*) < \infty$, *with:* $\eta_j^* = \eta_{1j}^* + \eta_{2j}^* + \eta_{3j}^*$, *where* $\eta_{1j}^* = \eta_{1j}$, *while*

$$\eta_{2j}^* = (\Gamma_F - \Psi_F^*)\,\varsigma(Z_j;\theta_0)$$

$$\eta_{3j}^* = -\left(E[f_0(Z_i,\theta_0)|U_i] - E\left[E[f_0(Z_i,\theta_0)|U_i]\right]\right).$$

The three terms η_{1j}^*, η_{2j}^*, and η_{3j}^* are all mean zero and have finite variance. They are generally correlated. When θ_0 is known, the term $\eta_{2j}^* = 0$ and this term is missing from the asymptotic expansion. The term η_{3j}^* is due to the estimation of ψ.

REMARKS.

1. Consistent standard errors can be constructed by substituting population quantities by estimated ones along the lines discussed in Newey and McFadden (1994) for finite dimensional parameters. An alternative approach to inference here is based on the bootstrap. In our case a standard i.i.d. resample from the data set can be shown to work for the nonparametric and semiparametric cases even under our discrete/asymptotically continuous design at least as far as approximating the asymptotic variance (see , e.g., Horowitz (2001) and Mammen (1992) for the nonparametric case and Chen, Linton, and Van Keilegom (2003) for the semiparametric case). We have taken this approach to inference in the application due to its simplicity.

2. Regarding the semiparametric estimators, it is not possible to provide an efficiency ranking of the two estimators $\widehat{\mu}_{3r}(x)$ and $\widehat{\mu}_{4r}(x)$ uniformly throughout the 'parameter space'. This result partly depends on the choice of $\widehat{\theta}$. It may be possible to develop an efficiency bound for estimation of the function $\mu_r(.)$ by following the calculations of Bickel, Klaassen, Ritov and Wellner (1993, Chapter 5). Since there are no additional restrictions on μ_r, the plug-in estimator with efficient $\widehat{\theta}$ should be efficient. See, e.g., Brown and Newey (1998)

4 Numerical Results

4.1 Monte Carlo

We report the results of a small simulation experiment based on a design of Crooker and Herriges (2004). Let

$$W_i = \beta_1 + \beta_2 X_i + \sigma\varepsilon_i,$$

where X_i is uniformly distributed on $[-30, 30]$ and ε_i is standard normal. We take $\beta_1 = 100$ and $\beta_2 = 2$, which guarantees that the mean WTP is equal to 100. We vary the value of $\sigma \in \{5, 10, 25, 50\}$ and sample size $n \in \{100, 300, 500\}$. For our first set of experiments the bid values are five points in $[25, 175]$ if $n = 100$, ten points if $n = 300$, and 15 points if $n = 500$; these points are randomly

assigned to individuals i before drawing the other data and so are fixed in repeated experiments. We take $\kappa = 100$. This design was chosen because it permits direct comparison with the parametric and SNP estimators of WTP considered by Crooker and Herriges (2004), at least when $n = 100$ (they did not increase the number of bids with sample size).

In this case $G(v|x) = 1 - \Phi((v - \beta_1 - \beta_2 x)/\sigma)$ and $g(v|x) = \phi((v - \beta_1 - \beta_2 x)/\sigma)/\sigma$, where Φ, ϕ denote the standard normal c.d.f. and density functions respectively. We estimate the moments: $E[W \mid X = x]$, i.e., $r(w, x) = w$, and $\text{std}(W \mid X = x) = \sqrt{E[W^2 \mid X = x] - E^2[W \mid X = x]}$, which corresponds to taking $r(w, x) = (w^2, w)$ and then computing the square root of $r_{w^2} - r_w^2$. Then: $\mu_w(x) = \beta_1 + \beta_2 x$, $\mu_{w^2}(x) = (\beta_1 + \beta_2 x)^2 + \sigma^2$, and $\text{std}(W \mid X = x) = \sigma$.

We compute estimators $\widehat{\mu}_\lambda(.)$ for $\lambda = 1, 2, 3, 4$. We used a local linear estimator with product Gaussian kernel and Silverman's rule of thumb bandwidths, that is, $b = 1.06sn^{-1/5}$, where s is the sample standard deviation of the specific covariate. The kernel and bandwidth are not likely to be optimal choices for this problem, but they are automatic and convenient and hence are fairly widely used choices in practice. We take $h(v \mid n)$ to be uniform over the range of bid values.

In this design, the estimator $\widehat{\mu}_1(x)$ is predicted to be approximately unbiased while the predicted bias of $\widehat{\mu}_2(x)$ is small but non-zero.

In Tables 3 and 4 we report four different performance measures: root pointwise mean squared error (RPMSE), pointwise mean absolute error (PMAE), root integrated mean squared error (RIMSE), and integrated mean absolute error (IMAE). Crooker and Herriges (2004) only report pointwise results. Like Crooker and Herriges, our pointwise results are calculated at the central point $x = 0$. Thus, their Table 2a ($n = 100$) and Appendix Table 1a ($n = 300$) are directly comparable with a subset of our results. Our conclusions are:

(A1) The performance of our estimators improves as σ decreases and as sample size increases according to all measures: the pointwise measures improve at approximately our theoretical asymptotic rate, while the integrated measures improve much more slowly; the semi-parametric estimators improve more rapidly with sample size.

(A2) For the larger samples, estimator $\widehat{\mu}_4$ performs best according to nearly all measures although for large σ, the difference between $\widehat{\mu}_4$ and some other estimators is minimal. For smaller sample sizes the ranking is a bit more variable: only $\widehat{\mu}_3$ is never ranked first.

(A4) Our best estimators always perform better than the Crooker and Herriges SNP estimator.

(A5) The estimates of $\text{std}(W \mid X = x)$ are subject to much more variability and bias than the estimates of $E[W \mid X = x]$, particularly in the large σ case.

While our estimators seem to work reasonably well in this discrete bid case, we would expect to obtain better results when the bid distribution is actually continuous and with full support like W. We repeated the above experiments with bid distribution uniform on $[25, 175]$ and report the results

in Tables 5 and 6. Our conclusions are:

(B1) The performance in the continuous design is somewhat better than in the discrete design. For some designs the pointwise results in Table 3 are better, but the integrated results are always better in Table 5. Note that for the pointwise results the chosen point of evaluation $x = 0$ corresponds to $E[W \mid X = 0] = 100$ and in Table 3 there is a point mass in the distribution of the bids at this point.

(B2) The results for standard deviation estimation are in most cases better in Table 6 than in Table 4.

(B3) The ranking of the estimators is the same in Table 5 as Table 3. Once again $\widehat{\mu}_4$ performs the best in large samples.

We also computed $\widehat{\mu}_0$, but the performance was considerably worse than $\widehat{\mu}_1$ and $\widehat{\mu}_2$, especially in the discrete design case. This maybe as expected since with data as discrete as this, derivative estimates can be nonsensical.

We also considered a design with a heavier tailed asymmetric error, specifically $\varepsilon_i \sim (\chi^2 - 1)/\sqrt{2}$. The performance results corresponding to Tables 3-6 were very similar in terms of the level of performance and the comparison across estimators likewise, a little bit worse everywhere, so to save space these results are not presented here. Instead in Figure 1 we report a QQ plot comparing the finite sample distribution of the standardized (studentized) estimator $\widehat{\mu}_1(0)/se(\widehat{\mu}_1(0))$ with the predicted normal distribution.[8] Based on this plot the distributional approximation appears quite good.

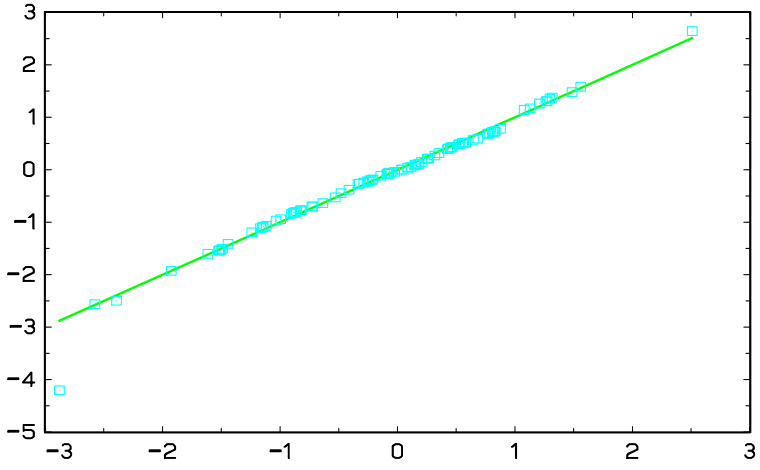

Figure 1. QQ plot of standardized studentized estimator $\widehat{\mu}_1(0)/se$ against standard normal distribution.

[8]The studentization is done with the estimated variance $\sum_j w_{ij}^2 \widehat{\varepsilon}_j^2$, where w_{ij} are the local linear smoothing weights and $\widehat{\varepsilon}_j$ are the nonparametric residuals.

This is from the case $n = 500$ and $\sigma = 5$.

4.2 Empirical Application

We examine a dataset used in An (2000), which is from a contingent valuation study conducted by Hanemann et al. (1991) to elicit the WTP for protecting wetland habitats and wildlife in California's San Joaquin Valley. Each respondent was assigned a bid value. They were then also given a second bid that was either higher or lower than the first, depending on their acceptance or rejection of the first bid. The total number of bid values in this unfolding bracket design is 14: {25, 30, 40, 55, 65, 75, 80, 110, 125, 140, 170, 210, 250, 375}. The dataset consists of bid responses and some personal characteristics of the respondents. The covariates X are age and number of years resident in California, education and income bracket, and binary indicators of sex, race, and membership in an environmental organization. The sample size, after excluding nonrespondents, incomplete responses, etc., is $n = 518$. The marginal distribution of Y across first bids was $\overline{Y}_1 = 0.396$ and across second bids was $\overline{Y}_2 = 0.581$, while $\overline{V}_1 = 132.4$ and $\overline{V}_2 = 153.9$. The second bid was more likely to receive a yes response, which is consistent with the larger mean value of the bid size. The contingency table is

	$Y_2 = 1$	$Y_2 = 0$
$Y_1 = 1$	131	74
$Y_1 = 0$	170	143

This gives a chi-squared statistic of 4.68, which is to be compared with $\chi^2_{0.05}(1) = 3.84$, so we reject the hypothesis of independence across bids, although not strongly.

The individuals for whom either $Y_1 = 0$ and $Y_2 = 1$ or $Y_1 = 1$ and $Y_2 = 0$ reveal a bound on their willingness to pay, because for these individuals we know their WTP lies between $\min\{V_1, V_2\}$ and $\max\{V_1, V_2\}$. By selecting these 244 individuals we obtain that $E(W)$ lies in the interval $[112.1, 187.1]$. This assumes that the first bids themselves do not influence the behaviour in the second round through, e.g., framing or anchoring effects. We provide some empirical evidence below that this assumption may not hold in our data.

We first consider semiparametric specifications for W, in particular:

$$W = X_i^\top \theta - \varepsilon \text{ and } \log(W) = X_i^\top \theta - \varepsilon,$$

so m is linear and Λ is the identity or the exponential function, respectively. With these specifications we estimate the quantity $\mu_w(x) = E(W \mid X = x)$ using our semiparametric estimators $\widehat{\mu}_j(x)$, $j = 3, 4$. To check for possible framing effects, we estimate this conditional mean WTP separately using first bid data and second bid data. Given that first bids were drawn with close to equal probabilities from

a discrete distribution of bids, we assumed that the limiting design density $h(V|X)$ is uniform on the interval $[V_{\min}, V_{\max}]$ (which is not a bad approximation).

In Table 7 we report the sample average of the estimates of $E(W \mid X = X_i)$, denoted $\overline{\widehat{\mu}_j}$, $j = 3, 4$, along with bootstrap confidence intervals. The computation of $\widehat{\mu}_j$ is exactly as described in the simulation section.

	Bid 1		Bid 2	
	Linear	Log-Linear	Linear	Log-linear
$\overline{\widehat{\mu}_3}$	110.480 [101.3,126.0]	112.676 [106.4,126.0]	172.838 [154.3,202.3]	356.771 [317.3,631.3]
$\overline{\widehat{\mu}_4}$	105.611 [97.8,118.1]	104.674 [99.7,115.0]	246.059 [196.8,294.5]	715.210 [380.8,1810.4]

Table 7: Estimates of WTP

Table 8 provides parameter estimates along with their 95% bootstrap confidence intervals, and asterisks indicating significant departure from zero at the 5% level.

	Bid 1		Bid 2	
	Linear	Log Linear	Linear	Log Linear
$YEARCA$	0.3823 [−0.118,0.935]	0.0051 [−0.0011,0.01]	0.5382 [−1.24,1.97]	0.0096 [−0.008,0.03]
$FEMALE$	0.5560 [−12.540,11.75]	0.0105 [−0.14,0.15]	28.290 [−10.5,70.8]	0.5033* [0.073,0.98]
$ln(AGE)$	−13.591 [−37.31,8.78]	−0.159 [−0.5,0.12]	−31.714 [−98.9,32.7]	−0.6081 [−1.73,0.33]
$EDUC$	−2.0237 [−4.98,0.72]	−0.0266 [−0.067,10.01]	1.2919 [−10.66,10.10]	0.0563 [−0.05,0.15]
$WHITE$	6.238 [−9.36,26.07]	0.0211 [−0.17,0.21]	60.2098* [3.0,112.0]	0.5206* [0.04,1.08]
$ENVORG$	1.968 [−15.07,16.49]	0.0423 [−0.17,0.20]	34.8597 [−23.9,88.7]	0.0931 [−0.56,0.66]
$ln(INCOME)$	2.378 [−9.70,12.21]	0.0459 [−0.09,0.17]	40.4140* [6.09,68.93]	0.2769 [−0.15,0.56]

Table 8

The estimated mean WTP based on only first bid data agree quite closely regardless of estimator or whether linear or loglinear model are assumed and the confidence intervals are quite narrow. Similar results were obtained for the sample median of $\{\widehat{\mu}_j(X_i)\}_{i=1}^n$ and for the estimates at the mean covariate value $\widehat{\mu}_j(\overline{X})$. The results for the second bid data are rather erratic and generally produce higher mean WTP values. This may be an indicator of framing, shadowing, or anchoring effects, in which hearing the first bid and replying to it affects responses to later bids. See, e.g.,

McFadden (1994), Green et al. (1998) and Hurd et al. (1998). These results may also be due to small sample problems associated with the survey design, in particular, the distribution of second bids differs markedly from the distribution of first bids, including some far larger bid values. An (2000), using a very different modeling methodology, tests and accepts the hypothesis of no framing effects in these data, though he does report some large differences in coefficient estimates based on data using both bids versus just first bid data. Using different estimators and combining both first and second bid data sets, An (2000) reports WTP at the mean ranging from 155 to 227 (plus one outlier estimate of 1341), which may be compared to our estimates of 99 to 113 for first bid data and 143 to 715 using only second bids.

Finally, we conducted a purely nonparametric analysis with each of the four continuous covariates, one at a time. In Figures 2 and 3 we provide the marginal smooths ($\widehat{\mu}_1(X_i)$) themselves along with a pointwise 95% confidence interval. These figures from the nonparametric estimator show some nonlinear effects. However, there are not likely to be statistically significant in view of the wide (pointwise) confidence intervals except for one case using second bid data. Recall that we are looking at marginal smooths and so it need not be the case that $E[W|X_j]$ is linear under either of our semiparametric specifications. Note also that the level of the effect is similar to that calculated in the full semiparametric specifications.

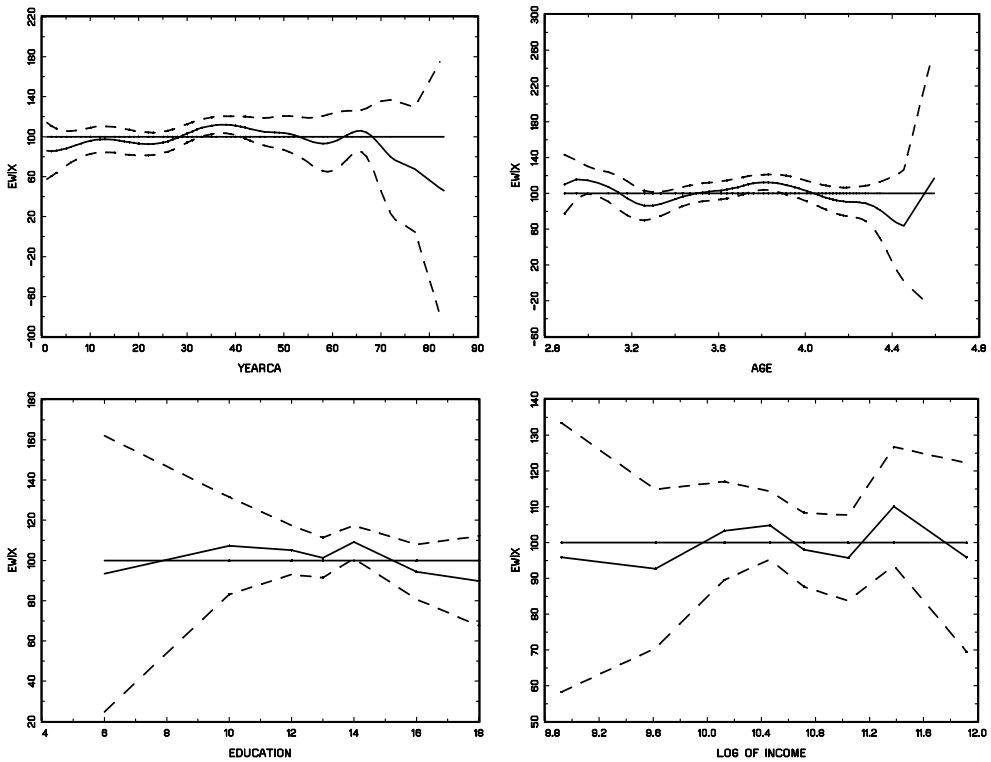

Figure 2. First bid data. Marginal smooths $\widehat{\mu}_1(X_i)$ with pointwise confidence intervals with estimated unconditional mean.

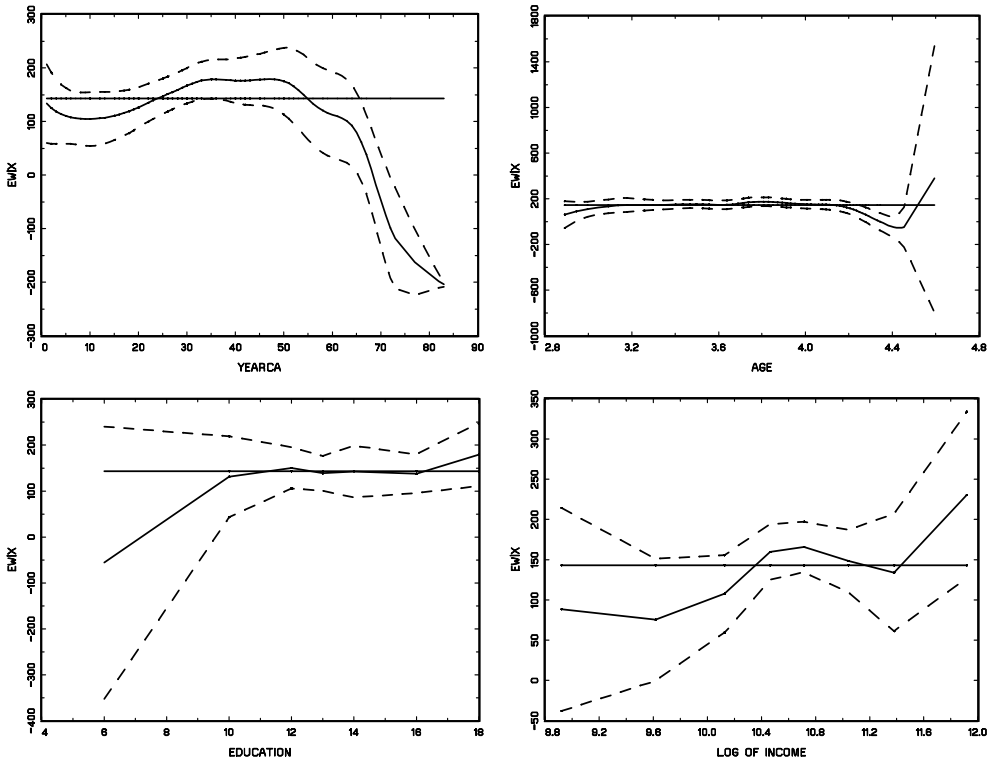

Figure 3. Second bid data. Marginal smooths $\widehat{\mu}_1(X_i)$ with pointwise confidence intervals with estimated unconditional mean.

5 Concluding Remarks

We have provided semiparametric and nonparametric estimators of conditional moments and quantiles of the latent W. The estimators appear to perform well with both simulated and actual data.

We have for convenience assumed throughout that the limiting support of V is bounded. Most of the results here should extend readily to the infinite support case, although some of the estimators may then require asymptotic trimming to deal with issues arising from division by a density estimate when the true density is not bounded away from zero.

The results here show the importance, for both identification and estimation, of experimental designs in which the distribution of bids or test values V possesses at least a fair number of mass points, and ideally is continuous. This should be taken as a recommendation to future designers of contingent valuation experiments. The precision of the estimators also depends in part on the distribution of test values. When designing experiments, one may wish to choose the limiting density h to maximize efficiency based on the variance estimators.

6 Appendix

This Appendix provides our main limiting distribution theorems. Some technical lemmas used by these theorems have dropped to save space. They appear in a supplemental appendix to this paper. The supplemental appendix also contains results regarding identification when the bid distribution is discrete.

6.1 Distribution Theory for Nonparametric Estimators

PROOF OF THEOREM 2. First consider the estimator $\widehat{\mu}_{1r}(x)$. Given our assumptions regarding the triangular array nature of the sampling scheme, replacing $H(v|x,n)$ with $H(v|x)$, and hence treating the estimator $\widehat{\mu}_{1r}(x)$ as if V's were drawn based on a continuous distribution $H(v|x)$, changes the limiting distribution of $\widehat{\mu}_{1r}(x)$ by an amount of smaller order than the leading term, and hence is first order asymptotically negligible. Moreover, when V's are draws based on a continuous distribution $H(v|x)$, the estimator $\widehat{\mu}_{1r}(x)$ just equals an ordinary nonparametric regression (noting that V_i is part of the variable being smoothed), and so follows the standard limiting distribution theory associated with nonparametric regression, which we do not spell out here to save space.

We now turn to $\widehat{\mu}_{2r}(x)$. First, we introduce some notation to define the local polynomial estimator $\widehat{G}(v \mid x)$. Following the notation of Masry (1996a,b), let $N_\ell = (\ell+d-1)!/\ell!(d-1)!$ be the number of distinct d-tuples j with $|j| = \ell$. Arrange these N_ℓ d-tuples as a sequence in a lexicographical order and let ϕ_ℓ^{-1} denote this one-to-one map. Define $\widetilde{X}_i = (V_i, X_i)$ and $\widetilde{x} = (v, x)$, and write $\widehat{G}(v \mid x) = \widehat{G}(\widetilde{x})$ and $G(v \mid x) = G(\widetilde{x})$ for short. We have $\widehat{G}(\widetilde{x}) = e_1^\top M_n^{-1} \Psi_n$, where $e_1 = (1, 0, \ldots, 0)^\top$ is the vector with the one in the first position, $M_n(\widetilde{x})$ and $\Psi_n(\widetilde{x})$ are symmetric $N \times N$ ($N = \sum_{\ell=0}^{p-1} N_\ell \times 1$) matrix and $N \times 1$ dimensional column vector respectively and are defined as

$$M_n(\widetilde{x}) = \begin{bmatrix} M_{n,0,0}(\widetilde{x}) & \cdots & M_{n,0,p-1}(\widetilde{x}) \\ \vdots & \ddots & \vdots \\ M_{n,p-1,0}(\widetilde{x}) & \cdots & M_{n,p-1,p-1}(\widetilde{x}) \end{bmatrix}, \ \Psi_n(\widetilde{x}) = \begin{bmatrix} \Psi_{n,0}(\widetilde{x}) \\ \vdots \\ \Psi_{n,p-1}(\widetilde{x}) \end{bmatrix},$$

where $M_{n,|j|,|k|}(\widetilde{x})$ is a $N_{|j|} \times N_{|k|}$ dimensional submatrix with the (l,r) element given by

$$\left[M_{n,|j|,|k|}\right]_{l,r} = \frac{1}{nb^d} \sum_{i=1}^{n} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b}\right)^{\phi_{|j|}(l)+\phi_{|k|}(r)} \widetilde{K}\left(\frac{\widetilde{x} - \widetilde{X}_i}{b}\right),$$

and $\Psi_{n,|j|}(\widetilde{x})$ is a $N_{|j|}$ dimensional subvector whose r-th element is given by

$$\left[\Psi_{n,|j|}\right]_r = \frac{1}{nb^d} \sum_{i=1}^{n} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b}\right)^{\phi_{|j|}(r)} \widetilde{K}\left(\frac{\widetilde{x} - \widetilde{X}_i}{b}\right) Y_i.$$

We can write

$$\widehat{G}(\widetilde{x}) - G(\widetilde{x}) = e_1^\top M_n^{-1}(\widetilde{x}) U_n(\widetilde{x}) + e_1^\top M_n^{-1}(\widetilde{x}) B_n(\widetilde{x}). \tag{16}$$

The stochastic term $U_n(\widetilde{x})$ and the bias term $B_n(\widetilde{x})$ are $N \times 1$ vectors

$$U_n(\widetilde{x}) = \begin{bmatrix} U_{n,0}(\widetilde{x}) \\ \vdots \\ U_{n,p-1}(\widetilde{x}) \end{bmatrix}, \; B_n(\widetilde{x}) = \begin{bmatrix} B_{n,0}(\widetilde{x}) \\ \vdots \\ B_{n,d}(\widetilde{x}) \end{bmatrix},$$

where $U_{n,l}(\widetilde{x})$ and $B_{n,l}(\widetilde{x})$ are defined similarly as $\Psi_{n,l}(\widetilde{x})$ so that $U_{n,|j|}(\widetilde{x})$ and $B_{n,|j|}(\widetilde{x})$ are a $N_{|j|}$ dimensional subvectors whose r-th elements are given by:

$$\left[U_{n,|j|}\right]_r = \frac{1}{nb^d} \sum_{i=1}^{n} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right)^{\phi_{|j|}(r)} \widetilde{K} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right) \varepsilon_i$$

$$\left[B_{n,|j|}\right]_r = \frac{1}{nb^d} \sum_{i=1}^{n} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right)^{\phi_{|j|}(r)} \widetilde{K} \left(\frac{\widetilde{x} - \widetilde{X}_i}{b} \right) \Delta_i(\widetilde{x}),$$

where $\Delta_i(\widetilde{x}) = G(\widetilde{X}_i) - \frac{1}{\mathbf{k}!} \sum_{0 \le |\mathbf{k}| \le p-1} (D^{\mathbf{k}} G)(\widetilde{x})(\widetilde{X}_i - \widetilde{x})^{\mathbf{k}}$, while $\varepsilon_i = Y_i - E(Y_i | \widetilde{X}_i)$ are independent random variables with conditional mean zero and uniformly bounded variances.

The argument is similar to Fan, Härdle, and Mammen (1998, Theorem 1); we just sketch out the extension to our quasi-discrete case. The first part of the argument is to derive a uniform approximation to the denominator in (16). We have

$$\sup_{v \in [\rho_0(x), \rho_1(x)]} |M_n(v,x) - E[M_n(v,x)]| = O_p(a_n), \tag{17}$$

where $a_n = \sqrt{\log n / nb^{d+1}}$. The justification for this comes from Masry (1996a, Theorem 2). Although he assumed a continuous covariate density, it is clear from the proofs that the argument goes through in our case. Discreteness of V_i only affects the bias calculation. We calculate $E[M_n(v,x)]$, for simplicity just the upper diagonal element

$$\begin{aligned} & E\widetilde{K}_b(\widetilde{x} - \widetilde{X}_i) \\ = & \int k_b(v - v') K_b(x - x') \, dH(v', x' | n) \\ = & \int k_b(v - v') K_b(x - x') \, dH(v', x') + \int k_b(v - v') K_b(x - x') \left[dH(v', x' | n) - dH(v', x') \right]. \end{aligned}$$

Then using integration by parts for Lebesgue integrals (Carter and van Brunt (2000, Theorem 6.2.2.), for large enough n we have

$$\int_{\rho_0(x)}^{\rho_1(x)} k_b(v - v') \left[dH(v' | x', n) - dH(v' | x') \right] = -\frac{1}{b^2} \int_{\rho_0(x)}^{\rho_1(x)} k' \left(\frac{v - v'}{b} \right) \left[H(v' | x', n) - H(v' | x') \right] dv',$$

since the function k is continuous everywhere and the boundary term

$$\begin{aligned}\mu_{kH}([\rho_0(x),\rho_1(x)]) &= k_b(v-\rho_1(x))\left[H(\rho_1(x)|x',n)-H(\rho_1(x)|x')\right] \\ &\quad -k_b(v-\rho_0(x))\left[H(\rho_0(x)|x',n)-H(\rho_0(x)|x')\right]=0\end{aligned}$$

for large enough n, where $\mu_{kH}(A)$ denotes the $H-$measure of the set A. Therefore, by the law of iterated expectation for some constant $C<\infty$,

$$\begin{aligned}&\left|\int k_b(v-v')K_b\left(x-x'\right)\left[dH(v',x'|n)-dH(v',x')\right]\right| \\ =\ &\left|\int k_b(v-v')K_b\left(x-x'\right)\left[dH(v'|x',n)-dH(v'|x')\right]dH(x')\right| \\ =\ &\left|\frac{1}{b^2}\int k'\left(\frac{v-v'}{b}\right)\left[H(v'|x',n)-H(v'|x')\right]dv'K_b\left(x-x'\right)dH(x')\right| \\ \leq\ &\sup_{v'}\sup_{|x'-x|\leq b}|H(v'|x',n)-H(v'|x')|\times\frac{1}{b^2}\int|k'\left(\frac{v-v'}{b}\right)|dv'\times\int\left|K_b\left(x-x'\right)\right|dH(x') \\ \leq\ &C\left(\frac{1}{b}\sup_{v'}\sup_{|x'-x|\leq b}|H(v'|x',n)-H(v'|x')|\right)\int|k'(t)|dt\int\left|K\left(u\right)\right|du\times\sup_{|x'-x|\leq b}h(x')=O_p(J^{-1}b^{-1}),\end{aligned}$$

by the integrability and smoothness on k. The right hand side does not depend on v so the bound is uniform.

For each j with $0\leq|j|\leq 2(p-1)$, let $\mu_j(\widetilde{K})=\int_{\mathbb{R}^{d+1}}u^j\widetilde{K}(u)du$, $\nu_j(\widetilde{K})=\int_{\mathbb{R}^{d+1}}u^j\widetilde{K}^2(u)du$, and define the $N\times N$ dimensional matrices M and Γ and $N\times 1$ vector B, where $N=\sum_{\ell=0}^{p-1}N_\ell\times 1$, by

$$M=\begin{bmatrix}M_{0,0} & M_{0,1} & \cdots & M_{0,p-1}\\ M_{1,0} & M_{1,1} & \cdots & M_{1,p-1}\\ \vdots & & & \vdots\\ M_{p-1,0} & M_{p-1,1} & \cdots & M_{p-1,p-1}\end{bmatrix},\ \Gamma=\begin{bmatrix}\Gamma_{0,0} & \Gamma_{0,1} & \cdots & \Gamma_{0,p-1}\\ \Gamma_{1,0} & \Gamma_{1,1} & \cdots & \Gamma_{1,p-1}\\ \vdots & & & \vdots\\ \Gamma_{p-1,0} & \Gamma_{p-1,1} & \cdots & \Gamma_{p-1,p-1}\end{bmatrix},\quad B=\begin{bmatrix}M_{0,p}\\ M_{1,p}\\ \vdots\\ M_{p-1,p}\end{bmatrix},$$

where $M_{i,j}$ and $\Gamma_{i,j}$ are $N_i\times N_j$ dimensional matrices whose (ℓ,m) element are, respectively, $\mu_{\phi_i(\ell)+\phi_j(m)}$ and $\nu_{\phi_i(\ell)+\phi_j(m)}$. Note that the elements of the matrices $M=M(\widetilde{K})$ and $\Gamma=\Gamma(\widetilde{K})$ are simply multivariate moments of the kernel $\widetilde{K}$ and $\widetilde{K}^2$, respectively.

Under the smoothness conditions on $h(v,x)$ we have for all j,k,l,r

$$\frac{1}{b^d}\int\left(\frac{\widetilde{x}-\widetilde{x}'}{b}\right)^{\phi_{|j|}(l)+\phi_{|k|}(r)}\widetilde{K}\left(\frac{\widetilde{x}-\widetilde{x}'}{b}\right)dH(v',x')=h(v,x)\left[M_{|j|,|k|}\right]_{l,r}+O(b)$$

uniformly over v. Therefore,

$$M_n(\widetilde{x})=h(\widetilde{x})M+O_p(c_n),\tag{18}$$

where $c_n = a_n + b + J^{-1}b^{-1}$, and the error is uniform over v in the support of $H(v|x,n)$. There is an additional term here of order $J^{-1}b^{-1}$ due to the discreteness. This term is of small order under our conditions.

Then $e_1^\top M_n^{-1}(\widetilde{x})U_n(\widetilde{x}) = e_1^\top M^{-1}U_n(\widetilde{x})/h(\widetilde{x}) + \text{rem}(\widetilde{x})$,where $\text{rem}(\widetilde{x})$ is a remainder term that is $o_p(n^{-1/2}b^{-(d+1)/2})$. By similar arguments we obtain $e_1^\top M_n^{-1}(\widetilde{x})B_n(\widetilde{x}) = b^p\beta(\widetilde{x}) + \text{rem}(\widetilde{x})$, where $\text{rem}(\widetilde{x})$ is a remainder term that is $o_p(b^p)$ and $\beta(\widetilde{x}) = e_1^\top M^{-1}BG^{(p+1)}(\widetilde{x})$. Therefore, we obtain

$$\widehat{G}(\widetilde{x}) - G(\widetilde{x}) = \frac{1}{h(\widetilde{x})}e_1^\top M^{-1}U_n(\widetilde{x}) + b^p\beta(v,x) + \text{rem}(\widetilde{x}),$$

where $\text{rem}(\widetilde{x})$ is a remainder term that is $o_p(n^{-1/2}b^{-(d+1)/2}) + o_p(b^p)$. We next substitute the leading terms into $\widehat{\mu}_{2r}(x)$, and recall that

$$\widehat{\mu}_{2r}(x) - \mu_r(x) = \int_{\rho_0(x)}^{\rho_1(x)} r'(v,x)[\widehat{G}(v \mid x) - G(v \mid x)]dv.$$

The standard integration argument along the lines of Fan, Härdle, and Mammen (1998) shows that the term $\text{rem}(\widetilde{x})$ can be ignored, and we obtain

$$\widehat{\mu}_{2r}(x) - \mu_r(x) = e_1^\top M^{-1}\overline{U}_n(x) + b^p\overline{\beta}(x) + o_p(n^{-1/2}b^{-d/2}),$$

where $\overline{\beta}(x) = \int \beta(v,x)d\lambda(v)$, while $\overline{U}_n(x) = [\overline{U}_{n,0}(x), \ldots, \overline{U}_{n,p}(x)]^\top$ is an $N \times 1$ vector, where $\overline{U}_{n,|j|}(x)$ is an $N_{|j|}$ dimensional subvector whose r-th elements are given by:

$$\left[\overline{U}_{n,|j|}\right]_r = \int u^{\phi^v_{|j|}(r)}k(u)du\frac{1}{nb^d}\sum_{i=1}^n \left(\frac{x - X_i}{b}\right)^{\phi^x_{|j|}(r)} K\left(\frac{x - X_i}{b}\right)\frac{r'(V_i,x)}{h(V_i,x)}\varepsilon_i.$$

We can write

$$e_1^\top M^{-1}\overline{U}_n(x) = \frac{1}{nb^d}\sum_{i=1}^n L_{d,p}\left(\frac{x - X_i}{b}\right)\frac{r'(V_i,x)}{h(V_i,x)}\varepsilon_i, \text{ where}$$

$$L_{d,p}\left(\frac{x - X_i}{b}\right) = e_1^\top M^{-1}\begin{bmatrix} \vdots \\ \int u^{\phi^v_{|j|}(r)}k(u)du\left(\frac{x-X_i}{b}\right)^{\phi^x_{|j|}(r)} K\left(\frac{x-X_i}{b}\right) \\ \vdots \end{bmatrix}.$$

Under our conditions

$$\begin{aligned}
& E\left[\left(\frac{r'(V_i,x)}{h(V_i,x)}\right)^2 \sigma^2(V_i,X_i)|X_i = x\right] \\
= & \int \left(\frac{r'(v,x)}{h(v,x)}\right)^2 \sigma^2(v,x)dH(v|x,n) \\
= & \int \left(\frac{r'(v,x)}{h(v,x)}\right)^2 \sigma^2(v,x)dH(v|x) + \int \left(\frac{r'(v,x)}{h(v,x)}\right)^2 \sigma^2(v,x)\left[dH(v|x,n) - dH(v|x)\right] \\
= & \int \left(\frac{r'(v,x)}{h(v,x)}\right)^2 \sigma^2(v,x)dH(v|x) + o(1).
\end{aligned}$$

It follows that the asymptotic variance of $\widehat{\mu}_{2r}(x)$ is

$$\begin{aligned}
& \frac{1}{nb^d} E\left[\frac{1}{b^d} L_{d,p}^2\left(\frac{x-X_i}{b}\right)\left(\frac{r'(V_i,x)}{h(V_i,x)}\right)^2 \sigma^2(V_i,X_i)\right] \\
= & \frac{1}{nb^d}\left[\int \frac{1}{b^d} L_{d,p}^2\left(\frac{x-X}{b}\right)\left(\frac{r'(V,x)}{h(V,x)}\right)^2 \sigma^2(V,X) dH(V|X) dH(X) + o(1)\right] \\
\simeq & \frac{1}{nb^d} \|L_{d,p}\|^2 \int \sigma^2(v,x)\left(\frac{r'(v,x)}{h(v,x)}\right)^2 h(v,x) dv,
\end{aligned}$$

by a change of variables and dominated convergence and taking account of the discreteness error. Furthermore, the central limit theorem holds by the arguments used in Gozalo and Linton (2000, Lemma CLT) and is not affected by the discreteness of V. The quantity $\|L_{d,p}\|^2$ can also be defined in terms of the basic kernel k.

The properties of

$$\widehat{\mu}_{0r}(x) = -\int_{a_0}^{a_1} r(v,x) \frac{\partial \widehat{G}(v \mid x)}{\partial v} dv \tag{19}$$

follow similarly. We have

$$\frac{\partial \widehat{G}(v \mid x)}{\partial v} - \frac{\partial G(v \mid x)}{\partial v} = e_v^\top M_{n*}^{-1}(\widetilde{x}) U_{n*}(\widetilde{x}) + e_1^\top M_{n*}^{-1}(\widetilde{x}) B_{n*}(\widetilde{x}),$$

where $e_v^\top = (0, 1, \ldots, 0)$ and $M_{n*}(\widetilde{x})$, $U_{n*}(\widetilde{x})$, and $B_{n*}(\widetilde{x})$ are like $M_n(\widetilde{x})$, $U_n(\widetilde{x})$, and $B_n(\widetilde{x})$ except that they are for one order higher polynomial. In particular, $e_v^\top M_{n*}^{-1}(\widetilde{x}) U_{n*}(\widetilde{x}) = e_v^\top M^{-1} U_{n*}(\widetilde{x})/h(\widetilde{x}) + rem(\widetilde{x})$, where $rem(\widetilde{x})$ is a small remainder term and M is the corresponding matrix of kernel weights. We apply the same integration argument except that we have to apply integration by parts to eliminate a bandwidth factor, and this is why the limiting variance involves the derivative of $r(v,x)/h(v,x)$. The bias term arguments are the same except that p is replaced by $p+1$. ■

6.2 Distribution Theory for Semiparametric Quantities

Let E_i denote expectation conditional on Z_i. In the proofs of Theorems 3 and 4 we make use of Lemmas 1 and 2 given in our supplemental appendix. Define

$$\rho_j(u,\theta) = h[\Lambda(m(X_j,\theta) - u)|X_j]\Lambda'(m(X_j,\theta) - u)$$

and $\psi_\theta(u) = E\rho_j(u,\theta)$ with $\psi(u) = \psi_{\theta_0}(u)$. Then, interchanging differentiation and integration (which is valid under our conditions) we have

$$\psi'(u) = E\frac{\partial \rho_j(u,\theta_0)}{\partial u} = -E\left(\left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda''\right](m(X_j,\theta_0) - u)\right). \tag{20}$$

PROOF OF THEOREM 3. Recall that

$$\widehat{\mu}_{3r}(x) = r[\Lambda(m(x,\widehat{\theta})),x] + \frac{1}{n}\sum_{i=1}^{n}\frac{r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)[Y_i - 1(\widehat{U}_i>0)]}{\widehat{\psi}(\widehat{U}_i)},$$

where $\widehat{U}_i = m(X_i,\widehat{\theta}) - \Lambda^{-1}(V_i)$ and

$$\widehat{\psi}(\widehat{U}_i) = \frac{1}{n}\sum_{j=1}^{n} h[\Lambda(m(X_j,\widehat{\theta})-\widehat{U}_i)|X_j]\Lambda'(m(X_j,\widehat{\theta})-\widehat{U}_i)) = \frac{1}{n}\sum_{j=1}^{n}\rho_j(\widehat{U}_i,\widehat{\theta}).$$

By a geometric series expansion of $1/\widehat{\psi}(\widehat{U}_i)$ about $1/\psi(U_i)$ we can write

$$\begin{aligned}
\widehat{\mu}_{3r}(x) &= \frac{1}{n}\sum_{i=1}^{n} f_1(Z_i,\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} f_2(Z_i,\theta_0)[\widehat{\psi}(\widehat{U}_i)-\psi(U_i)] && (21)\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}[f_2(Z_i,\widehat{\theta}) - f_2(Z_i,\theta_0)][\widehat{\psi}(\widehat{U}_i)-\psi(U_i)] && (22)\\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\frac{r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)[Y_i-1(\widehat{U}_i>0)]}{\psi^2(U_i)\widehat{\psi}(\widehat{U}_i)}[\widehat{\psi}(\widehat{U}_i)-\psi(U_i)]^2, && (23)
\end{aligned}$$

where

$$f_2(Z_i,\theta) = \frac{r'[\Lambda(m(x,\theta)-U_i(\theta)),x]\Lambda'(m(x,\theta)-U_i(\theta)))[Y_i - 1(U_i(\theta)>0)]}{\psi^2(U_i)}.$$

The leading terms are derived from (21), while (22) and (23) contain remainder terms. We just present the leading terms here and refer to the supplementary material for the full arguments.

LEADING TERMS. Lemma 1 implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[f_1(Z_i,\widehat{\theta}) - Ef_1(Z_i,\theta_0)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\Gamma_F\varsigma(Z_i,\theta_0) + [f_1(Z_i,\theta_0) - Ef_1(Z_i,\theta_0)]\} + o_p(1), \quad (24)$$

where $Ef_1(Z_i,\theta_0) = \mu_r(x)$, and $f_1(Z_i,\theta_0) - Ef_1(Z_i,\theta_0) = f_0(Z_i,\theta_0) - Ef_0(Z_i,\theta_0)$ due to the cancellation of the common term $r[\Lambda(m(x,\theta_0)),x]$. The stochastic equicontinuity condition of Lemma 1 is verified in a separate appendix, see below.

Let $L(Z_i,Z_j) = \xi_j(U_i) + \Gamma(Z_i)\varsigma(Z_j;\theta_0)$, and

$$\xi_j(u) = \rho_j(u,\theta_0) - E\rho_j(u,\theta_0)$$

$$\Gamma(Z_i) = E_i\left[\zeta_{ij}\frac{\partial m(X_j,\theta_0)}{\partial\theta}\right] - E_i[\zeta_{ij}]\frac{\partial m(X_i,\theta_0)}{\partial\theta}, \text{ where } \zeta_{ij} = -\frac{\partial\rho_j(U_i,\theta_0)}{\partial u}.$$

Note that $E_i[\xi_j(U_i)] = 0$ but $E_j[\xi_j(U_i)] \neq 0$. We first approximate $n^{-1}\sum_{i=1}^{n} f_2(Z_i,\theta_0)[\widehat{\psi}(\widehat{U}_i)-\psi(U_i)]$ by $n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} f_2(Z_i,\theta_0)L(Z_i,Z_j)$. Specifically, by Lemma 2 and the fact that $E|f_2(Z_i,\theta_0)| < \infty$,

we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} f_2(Z_i,\theta_0)[\widehat{\psi}(\widehat{U}_i)-\psi(U_i)-\frac{1}{n}\sum_{j=1}^{n} L(Z_i,Z_j)]\right| \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}|f_2(Z_i,\theta_0)|\times\max_{1\leq i\leq n}\left|[\widehat{\psi}(\widehat{U}_i)-\psi(U_i)-\frac{1}{n}\sum_{j=1}^{n} L(Z_i,Z_j)]\right| \\
=\ & o_p(n^{-1/2}).
\end{aligned}$$

Next, letting $\varphi_n(z_1,z_2)=n^{-2}f_2(z_1,\theta_0)L(z_1,z_2)$ we have

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f_2(Z_i,\theta_0)L(Z_i,Z_j)=\sum_{i=1}^{n}\sum_{j=1}^{n}\varphi_n(Z_i,Z_j),$$

which can be approximated by a second order U-statistic as follows. Letting $p_n(z_1,z_2)=n(n-1)[\varphi_n(z_1,z_2)+\varphi_n(z_2,z_1)]/2$ we have

$$\mathcal{Q}_n=\sum_{i=1}^{n}\sum_{j=1}^{n}\varphi_n(Z_i,Z_j)=\binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_n(Z_i,Z_j)+o_p(n^{-1/2}),$$

since $\sum_{i=1}^{n}\varphi_n(Z_i,Z_i)=o_p(n^{-1/2})$. Now p_n is a symmetric kernel, i.e., $p_n(z_1,z_2)=p_n(z_2,z_1)$ and we can apply Lemma 3.1 of Powell, Stock, and Stoker (1989). Letting

$$\widehat{\mathcal{Q}}_n=\frac{2}{n}\sum_{j=1}^{n}\omega_n(Z_j),\ \text{where}\ \omega_n(Z_i)=E_i\left[p_n(Z_i,Z_j)\right],$$

we have $\sqrt{n}(Q_n-\widehat{\mathcal{Q}}_n)=o_p(1)$. It remains to find $\omega_n(Z_i)$. We have $2\omega_n(Z_i)=E\left[f_2(Z_j,\theta_0)\Gamma(Z_j)\right]\varsigma(Z_i;\theta_0)+E_i\left[f_2(Z_j,\theta_0)\xi_i(U_j)\right]$, because $E_i[L(Z_i,Z_j)]=0$. Furthermore,

$$\begin{aligned}
E_j\left[f_2(Z_i,\theta_0)\xi_j(U_i)\right] &= E_j\left[f_2(Z_i,\theta_0)[\rho_j(U_i,\theta_0)-E_i\rho_j(U_i,\theta_0)]\right] \\
&= E_j\left[f_0(Z_i,\theta_0)\frac{[\rho_j(U_i,\theta_0)-\psi(U_i)]}{\psi(U_i)}\right].
\end{aligned}$$

$$\begin{aligned}
E\left[f_2(Z_i,\theta_0)\Gamma(Z_i)\right] &= E\left[f_2(Z_i,\theta_0)\left\{E_i\left[\zeta_{ij}\frac{\partial m(X_j,\theta_0)}{\partial\theta}\right]-E_i[\zeta_{ij}]\frac{\partial m(X_i,\theta_0)}{\partial\theta}\right\}\right] \\
&= E\left[\frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\zeta_{ij}\left\{\frac{\partial m(X_j,\theta_0)}{\partial\theta}-\frac{\partial m(X_i,\theta_0)}{\partial\theta}\right\}\right].
\end{aligned}$$

Writing $\zeta_{ij} = E_i\zeta_{ij} + \zeta_{ij} - E_i\zeta_{ij}$, where $E_i\zeta_{ij} = -\psi'(U_i)$, we have

$$E\left[\frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\zeta_{ij}\left\{\frac{\partial m(X_j,\theta_0)}{\partial\theta} - \frac{\partial m(X_i,\theta_0)}{\partial\theta}\right\}\right]$$

$$\begin{aligned} = \; & E\left[-f_0(Z_i,\theta_0)\frac{\psi'(U_i)}{\psi(U_i)}\left\{\frac{\partial m(X_i,\theta_0)}{\partial\theta} - E\left(\frac{\partial m(X_i,\theta_0)}{\partial\theta}\right)\right\}\right] \\ & + E\left[\frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\left\{\zeta_{ij} - E_i\zeta_{ij}\right\}\left\{\frac{\partial m(X_j,\theta_0)}{\partial\theta} - E\left(\frac{\partial m(X_j,\theta_0)}{\partial\theta}\right)\right\}\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} \widehat{\mathcal{Q}}_n \; = \; & E\left[-f_0(Z_i,\theta_0)\frac{\psi'(U_i)}{\psi(U_i)}\widetilde{\gamma}_i + \frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\widetilde{\zeta}_{ij}\widetilde{\gamma}_j\right]\frac{1}{n}\sum_{j=1}^{n}\varsigma(Z_j;\theta_0) \\ & + \frac{1}{n}\sum_{j=1}^{n}E_j\left[f_0(Z_i,\theta_0)\frac{[\rho_j(U_i,\theta_0) - \psi(U_i)]}{\psi(U_i)}\right]. \end{aligned} \tag{25}$$

We have shown that $n^{-1}\sum_{i=1}^{n} f_2(Z_i,\theta_0)[\widehat{\psi}(\widehat{U}_i) - \psi(U_i)] = \widehat{\mathcal{Q}}_n + o_p(n^{-1/2})$, where $\widehat{\mathcal{Q}}_n$ is given in (25). This concludes the analysis of the leading terms.

In conclusion, $\sqrt{n}[\widehat{\mu}_{3r}(x) - \mu_r(x;\theta_0)] = n^{-1/2}\sum_{i=1}^{n}\eta_i + o_p(1)$, as required. The asymptotic distribution of $\sqrt{n}[\widehat{\mu}_{3r}(x) - \mu_r(x)]$ follows from the central limit theorem for independent random variables with finite variance. ■

Proof of Theorem 4. By a geometric series expansion we can write

$$\widehat{\mu}^*_{4r}(x;\widehat{\theta}) \; = \; \frac{1}{n}\sum_{i=1}^{n} f_1(Z_i,\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} f_2(Z_i,\theta_0)[\widetilde{\psi}(\widehat{U}_i) - \psi(U_i)] \tag{26}$$

$$-\frac{1}{n}\sum_{i=1}^{n}[f_2(Z_i,\widehat{\theta}) - f_2(Z_i,\theta_0)] \times [\widetilde{\psi}(\widehat{U}_i) - \psi(U_i)] \tag{27}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\frac{r'[\Lambda(m(x,\widehat{\theta}) - \widehat{U}_i), x]\Lambda'(m(x,\widehat{\theta}) - \widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\psi^2(U_i)\widetilde{\psi}(\widehat{U}_i)}[\widetilde{\psi}(\widehat{U}_i) - \psi(U_i)]^2. \tag{28}$$

The leading terms in this expansion are derived from (26), while (27) and (28) contain remainder terms.

Leading Terms. We make use of Lemma 3 given in our supplemental appendix. The term $n^{-1}\sum_{i=1}^{n} f_1(Z_i,\widehat{\theta})$ has already been analyzed above. By Lemma 3 we have with probability tending to one for some function d(.) with finite r moments

$$\begin{aligned} \left|\frac{1}{n}\sum_{i=1}^{n} f_2(Z_i,\theta_0)\left[\widetilde{\psi}(\widehat{U}_i) - \psi(U_i) - \frac{1}{n}\sum_{j=1}^{n}L^*(Z_i,Z_j)\right]\right| \; & \leq \; \frac{1}{nb^3}\left(\frac{1}{n}\sum_{i=1}^{n}|f_2(Z_i,\theta_0)|d(X_i)\right) \\ & = \; O_p(n^{-1}b^{-3}). \end{aligned} \tag{29}$$

where $L^*(Z_i, Z_j) = b^{-1}k((U_i - U_j)/b) - \psi(U_i) + \Gamma^*(Z_i) \cdot \varsigma(Z_j, \theta_0)$ and

$$\Gamma^*(Z_i) = \psi(U_i) \left\{ \frac{\psi'(U_i)}{\psi(U_i)} \left[\frac{\partial m}{\partial \theta^\top}(X_i, \theta_0) - E\left(\frac{\partial m}{\partial \theta^\top}(X_i, \theta_0) \mid U_i \right) \right] - \overline{m}'_{\theta_0}(U_i) \right\}.$$

Here, $\overline{m}_{\theta_0}(U_i) = E[\frac{\partial m}{\partial \theta^\top}(X_i, \theta_0) \mid U_i]$. Under our bandwidth conditions, the right hand side of (29) is $o_p(n^{-1/2})$.

Next,

$$\frac{1}{n}\sum_{i=1}^n f_2(Z_i, \theta_0) \frac{1}{n}\sum_{j=1}^n L^*(Z_i, Z_j) = \sum_{i=1}^n \sum_{j=1}^n \varphi_n(Z_i, Z_j)$$

where

$$\varphi_n(Z_i, Z_j) = \frac{1}{n^2} f_2(Z_i, \theta_0) \left[\frac{1}{b} k\left(\frac{U_i - U_j}{b} \right) - \psi(U_i) + \Gamma^*(Z_i) \cdot \varsigma(Z_j, \theta_0) \right].$$

Note that

$$\begin{aligned} E_i \varphi_n(Z_i, Z_j) &= \frac{1}{n^2} f_2(Z_i, \theta_0) \left[\int \frac{1}{b} k\left(\frac{U_i - u}{b} \right) \psi(u) du - \psi(U_i) \right] \\ &= \frac{1}{n^2} f_2(Z_i, \theta_0) \left[\int k(t) \psi(U_i + tb) dt - \psi(U_i) \right] = O_p(n^{-2} b^2) \end{aligned}$$

uniformly in i. Define $\overline{f}_2(U_i) = E[f_2(Z_i, \theta_0) | U_i]$. Then by iterated expectation

$$n^2 E_j \varphi_n(Z_i, Z_j) = E\left[\overline{f}_2(U_i) \frac{1}{b} k\left(\frac{U_i - U_j}{b} \right) \right] - E\left[\overline{f}_2(U_i) \psi(U_i) \right] + E\left[f_2(Z_i, \theta_0) \Gamma^*(Z_i) \right] \cdot \varsigma(Z_j, \theta_0),$$

where, using integration by parts, a change of variable, and dominated convergence,

$$E\left[\overline{f}_2(U_i) \frac{1}{b} k\left(\frac{U_i - U_j}{b} \right) \right] = \int \overline{f}_2(u) \frac{1}{b} k\left(\frac{u - U_j}{b} \right) \psi(u) du = \overline{f}_2(U_j) \psi(U_j) + O_p(b^2)$$

uniformly in i. Note that $\overline{f}_2(U_j)\psi(U_j) = \overline{f}_0(U_j) = E[f_0(Z_j, \theta_0) | U_j]$. Furthermore,

$$\begin{aligned} E\left[f_2(Z_i, \theta_0) \Gamma^*(Z_i) \right] &= E\left[f_0(Z_i, \theta_0) \left\{ \frac{\psi'(U_i)\gamma_i^*}{\psi(U_i)} - \frac{\overline{m}'_\theta(U_i)}{\psi(U_i)} \right\} \right] \\ &= E\left[\frac{\psi'(U_i)}{\psi(U_i)} \left\{ f_0(Z_i, \theta_0) - \overline{f}_0(U_i) \right\} \gamma_i^* \right] - E\left[\overline{f}_0(U_i) \overline{m}'_\theta(U_i) \right] \end{aligned}$$

by substituting in for f_2 and decomposing $f_0(Z_i, \theta_0) = \overline{f}_0(U_i) + f_0(Z_i, \theta_0) - \overline{f}_0(U_i)$ and using that $E[\gamma_i^* | U_i] = 0$. Using the same U-statistic argument as in the proof of Theorem 3 we obtain

$$\frac{1}{n^2}\sum_{i=1}^n f_2(Z_i, \theta_0) \sum_{j=1}^n L^*(Z_i, Z_j) = \frac{1}{n}\sum_{j=1}^n \omega_n(Z_j) + o_p(n^{-1/2}),$$

where $\omega_n(Z_j) = \overline{f}_0(U_j) - E[\overline{f}_0(U_j)] + E\left[f_2(Z_i, \theta_0) \Gamma^*(Z_i) \right] \varsigma(Z_j)$. ■

References

[1] An, M.Y. (2000). "A Semiparametric Distribution for Willingness to Pay and Statistical Inference with Dichotomous Choice CV Data," *American Journal of Agricultural Economics* 82, 487-500.

[2] Andrews, D.W.K. (1994): "Asymptotics for Semiparametric Econometric Models by Stochastic Equicontinuity." *Econometrica* 62, 43-72.

[3] Bickel, P.J., C.A.J. Klaassen, J. Ritov, and J. Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*. Springer: Berlin.

[4] Brown, B.W. and W.K. Newey (1998): "Efficient Semiparametric Estimation of Expectations," *Econometrica,* 66, 453-464.

[5] Carter, N., and B. van Brunt (2000): *The Lebesgue Stieltjes Integral*. Springer, Berlin.

[6] Chaudhuri, P. (1991). "Nonparametric estimates of regression quantiles and their local Bahadur representation," *Annals of Statistics,* 19, 760-777.

[7] Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion is not Smooth." *Econometrica* 71, 1591-1608.

[8] Chen, H. and A. Randall (1997): "Semi-nonparametric Estimation of Binary Response Models With an Application to Natural Resource Valuation, *Journal of Econometrics*, 76, 323-340.

[9] Coppejans, M. (2003): "Effective Nonparametric Estimation in the Case of Severely Discretized Data," *Journal of Econometrics* 117, 331-367.

[10] Creel, M., and J. Loomis (1997): "Semi-nonparametric Distribution-free Dichotomous Choice Contingent Valuation," *Journal of Environmental Economics and Management*, 32, 341-358.

[11] Crooker, J.R., and J.A. Herriges (2004): "Parametric and Semi-Nonparametric Estimation of Willingness-to-Pay in the Dichotomous Choice Contingent Valuation Framework," *Environmental and Resource Economics* 27, 451-480.

[12] Das, M. (2002), "Minimum Distance Estimators for Nonparametric Models With Grouped Dependent Variables," Columbia University Discussion paper no. 0102-41.

[13] Delgado, M., and J. Mora (1995): "Nonparametric and Semiparametric Estimation with Discrete Regressors," *Econometrica,* 63, 1477-1484.

[14] Fan, J, E., Härdle, W. and E. Mammen, (1998). Direct estimation of low dimensional components in additive models. *Annals of Statistics*, 26, 943-971.

[15] Gozalo, P., and O.B. Linton (2000): "Local Nonlinear Least Squares: Using Parametric Information in Nonparametric Regression." *Journal of Econometrics,* 99, 63-106.

[16] Hanemann, WM., J. Loomis, and B. Kanninen (1991) "Statistical Efficiency of Double-Bounded Dichotomous Choice Contingent Valuation." *American Journal of Agricultural Economics* 73, 1255-63.

[17] Härdle, W., and O.B. Linton (1994), "Applied Nonparametric Methods," *The Handbook of Econometrics*, vol. IV*,* eds. D.F. McFadden and R.F. Engle III. North Holland.

[18] Ho, K. and P.K. Sen (2000): "Robust Procedures For Bioassays and Bioequivalence Studies," *Sankhya, Ser. B,* 62, 119-133.

[19] Horowitz, J.L. (2001): "The Bootstrap," in Handbook of Econometrics, vol. V, J.J. Heckman and E.E. Leamer, eds., Elsevier Science B.V., Ch. 52, 3159-3228.

[20] Kanninen, B. (1993), "Dichotomous Choice Contingent Valuation," *Land Economics*, 69, 138-146.

[21] Klein, R. and R. H. Spady (1993), "An efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61, 387-421.

[22] Lewbel, A. (1997), "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory,* 13, 32-51.

[23] Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics* 97, 145-177.

[24] Li, Q. and Racine, J. (2004): "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics* 119, 99-130.

[25] Linton, O. and J.P. Nielsen (1995): "A kernel method of estimating structured nonparametric regression based on marginal integration," *Biometrika*, 82, 93-100.

[26] Linton, O. and J. Park (2009) A Comparison of Analytic Standard Errors for Nonparametric Regression. In progress.

[27] Mammen, E. (1992): "When Does Bootstrap Work? Asymptotic Results and Simulations." Lecture Notes in Statistics, 77, Springer: New York.

[28] Manski, C. and E. Tamer, (2000), "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519-546.

[29] Matzkin, R., (1992), "Non-parametric and Distribution-free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica* 60, 239-270.

[30] Masry, E. (1996a), "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *Journal of Time Series Analysis* 17, 571-599.

[31] Masry, E., (1996b), "Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and their Applications* 65, 81-101.

[32] McFadden, D. (1994), "Contingent Valuation and Social Choice," *American Journal of Agricultural Economics,* 76, 4.

[33] McFadden, D. (1998), "Measuring Willingness-to-Pay for Transportation Improvements, in T. Gärling, T. Laitila, and K. Westin (eds.),Theoretical Foundations of Travel Choice Modeling, 339-364, Elsevier Science: Amsterdam.

[34] Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[35] Pakes, A. and D. Pollard. (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-57.

[36] Ramgopal, P., P.W. Laud, and A.F.M. Smith. (1993) "Nonparametric Bayesian Bioassay With Prior Constraints on the Shape of the Potency Curve," *Biometrika*, 80, 489-498.

[37] Sherman, R. P. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123-37.

[38] Shorack, G. R. and J. A. Wellner, (1986): *Empirical Processes With Applications To Statistics*, John Wiley and sons.

[39] Silverman, B. (1986): *Density estimation for statistics and data analysis.* London, Chapman and Hall.

[40] Sperlich, S, O.B. Linton, and W. Härdle, (1999): "A Simulation comparison between the Backfitting and Integration methods of estimating Separable Nonparametric Models," *TEST,* 8, 419-458.

[41] Stone, C.J. (1982): "Optimal global rates of convergence for nonparametric regression." *Annals of Statistics* 8, 1040-1053.

[42] Wang, M-C, and J. Van Ryzin (1981): "A class of smooth estimators for discrete distributions," *Biometrika* 68, 301-309.

		$\sigma = 5$			$\sigma = 10$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
RPMSE	$\widehat{\mu}_1$	4.04	3.86	2.39	4.20	4.54	3.10	10.54	7.11	5.88
	$\widehat{\mu}_2$	4.52	3.66	2.16	4.49	4.30	2.69	9.78	6.37	5.02
	$\widehat{\mu}_3$	5.65	2.20	1.64	5.64	2.69	1.97	10.69	4.60	3.64
	$\widehat{\mu}_4$	4.80	1.77	1.25	4.75	2.15	1.59	8.42	4.53	3.32
	$\widehat{\mu}_6$	4.77	3.89	2.07	5.03	3.69	2.05	8.37	4.32	3.28
PMAE	$\widehat{\mu}_1$	3.20	3.20	1.89	3.32	3.70	2.47	8.45	5.69	4.61
	$\widehat{\mu}_2$	3.59	3.07	1.73	3.60	3.57	2.13	7.86	5.09	3.92
	$\widehat{\mu}_3$	4.50	1.77	1.31	4.54	2.12	1.59	8.65	3.67	2.91
	$\widehat{\mu}_4$	3.75	1.44	0.99	3.76	1.73	1.28	6.74	3.59	2.62
	$\widehat{\mu}_6$	3.82	3.29	1.63	4.06	3.07	1.60	6.98	3.54	2.67
RIMSE	$\widehat{\mu}_1$	14.39	7.89	5.81	14.29	7.88	5.76	17.29	11.72	9.59
	$\widehat{\mu}_2$	12.85	5.29	2.50	13.22	5.60	3.06	16.60	10.90	8.94
	$\widehat{\mu}_3$	12.28	4.76	3.35	12.21	5.16	3.46	16.12	9.83	7.81
	$\widehat{\mu}_4$	11.90	4.58	3.18	11.80	4.90	3.27	14.65	9.79	7.66
	$\widehat{\mu}_6$	11.83	5.72	3.58	11.80	5.75	3.50	14.62	9.67	7.65
IMAE	$\widehat{\mu}_1$	10.88	5.71	4.10	10.77	5.90	4.17	13.41	8.98	7.33
	$\widehat{\mu}_2$	10.54	4.17	1.92	10.76	4.47	2.34	13.05	8.56	6.94
	$\widehat{\mu}_3$	9.44	3.59	2.49	9.52	3.88	2.62	12.71	7.68	6.13
	$\widehat{\mu}_4$	9.12	3.36	2.33	9.16	3.59	2.45	11.53	7.69	5.98
	$\widehat{\mu}_6$	9.32	4.50	2.73	9.39	4.47	2.65	11.74	7.64	6.04

Table 3. Estimation of conditional mean in discrete bid design; 500 replications;

		$\sigma = 5$			$\sigma = 10$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
RPMSE	$\widehat{\mu}_1$	8.95	7.43	7.12	7.40	5.03	5.35	9.89	8.17	7.32
	$\widehat{\mu}_2$	16.39	14.51	12.89	13.00	10.93	9.97	8.73	7.29	6.64
	$\widehat{\mu}_3$	6.75	2.65	2.08	5.43	2.53	2.52	7.68	5.29	5.12
	$\widehat{\mu}_4$	6.60	2.13	1.39	5.15	1.98	1.49	8.27	5.33	4.87
	$\widehat{\mu}_6$	27.26	13.86	15.33	24.38	14.16	13.71	19.92	15.00	10.80
PMAE	$\widehat{\mu}_1$	8.02	7.03	6.95	6.29	4.36	4.91	7.86	6.86	6.34
	$\widehat{\mu}_2$	16.20	14.43	12.85	12.69	10.77	9.86	7.04	6.10	5.69
	$\widehat{\mu}_3$	5.64	2.04	1.60	4.51	1.97	2.08	6.11	4.45	4.60
	$\widehat{\mu}_4$	5.41	1.59	1.05	4.19	1.54	1.18	6.60	4.42	4.09
	$\widehat{\mu}_6$	23.28	10.77	12.42	21.01	12.25	12.08	14.67	10.98	8.19
RIMSE	$\widehat{\mu}_1$	17.56	11.35	8.11	14.23	9.73	7.68	13.02	13.08	13.61
	$\widehat{\mu}_2$	22.07	14.37	9.93	18.01	11.22	7.69	10.82	11.03	11.78
	$\widehat{\mu}_3$	6.75	2.65	2.08	5.43	2.53	2.52	7.68	5.29	5.12
	$\widehat{\mu}_4$	6.60	2.13	1.39	5.15	1.98	1.49	8.27	5.33	4.87
	$\widehat{\mu}_6$	21.78	14.62	11.68	19.31	13.37	11.36	19.93	17.30	17.30
IMAE	$\widehat{\mu}_1$	15.98	9.62	6.99	12.74	8.23	6.55	9.75	10.18	10.96
	$\widehat{\mu}_2$	21.59	13.37	9.19	17.40	10.24	7.03	8.43	8.94	9.92
	$\widehat{\mu}_3$	5.64	2.04	1.60	4.51	1.97	2.08	6.11	4.45	4.60
	$\widehat{\mu}_4$	5.41	1.59	1.05	4.19	1.54	1.18	6.60	4.42	4.09
	$\widehat{\mu}_6$	17.95	11.24	9.09	16.41	11.33	9.84	14.24	12.64	12.72

Table 4. Estimation of conditional standard deviation in discrete bid design; 500 replications;

		$\sigma = 5$			$\sigma = 10$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
RPMSE	$\widehat{\mu}_1$	5.64	3.30	2.58	6.34	3.79	3.07	11.39	7.25	5.90
	$\widehat{\mu}_2$	5.02	2.90	2.32	5.60	3.33	2.68	10.20	6.39	5.24
	$\widehat{\mu}_3$	4.40	2.24	1.68	5.10	2.70	2.08	8.01	4.55	3.44
	$\widehat{\mu}_4$	3.51	1.66	1.23	4.10	2.10	1.62	7.89	4.42	3.34
	$\widehat{\mu}_6$	5.85	3.36	2.59	5.96	3.39	2.65	7.57	4.34	3.31
PMAE	$\widehat{\mu}_1$	4.41	2.61	2.06	5.00	3.01	2.44	9.07	5.74	4.70
	$\widehat{\mu}_2$	4.00	2.31	1.84	4.43	2.66	2.13	8.16	5.12	4.18
	$\widehat{\mu}_3$	3.43	1.75	1.32	4.07	2.14	1.66	6.37	3.63	2.74
	$\widehat{\mu}_4$	2.76	1.32	0.98	3.26	1.68	1.29	6.22	3.51	2.65
	$\widehat{\mu}_6$	4.66	2.67	2.07	4.76	2.71	2.11	6.05	3.45	2.65
RIMSE	$\widehat{\mu}_1$	12.41	7.59	6.02	12.37	7.56	6.06	15.88	11.22	9.75
	$\widehat{\mu}_2$	7.85	4.42	3.42	8.30	4.80	3.77	14.36	10.33	9.16
	$\widehat{\mu}_3$	8.14	4.57	3.51	8.44	4.69	3.70	12.68	9.06	8.07
	$\widehat{\mu}_4$	7.70	4.32	3.32	7.89	4.38	3.47	12.61	9.00	8.03
	$\widehat{\mu}_6$	9.02	5.22	4.03	9.00	5.12	4.05	12.41	8.96	8.01
IMAE	$\widehat{\mu}_1$	8.88	5.44	4.33	8.97	5.48	4.40	12.11	8.56	7.50
	$\widehat{\mu}_2$	6.01	3.40	2.64	6.35	3.69	2.90	11.01	8.01	7.19
	$\widehat{\mu}_3$	6.04	3.39	2.61	6.39	3.56	2.81	9.80	7.08	6.37
	$\widehat{\mu}_4$	5.62	3.14	2.42	5.87	3.26	2.59	9.76	7.02	6.34
	$\widehat{\mu}_6$	6.95	4.03	3.11	6.99	3.97	3.14	9.61	6.98	6.33

Table 5. Estimation of conditional mean in continuous design; 10,000 replications;

		$\sigma = 5$			$\sigma = 10$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
RPMSE	$\widehat{\mu}_1$	9.65	7.43	6.46	7.51	5.33	4.55	10.70	7.80	7.18
	$\widehat{\mu}_2$	16.71	14.02	12.48	13.05	10.66	9.37	9.60	7.08	6.62
	$\widehat{\mu}_3$	5.20	2.90	2.12	5.20	2.81	2.36	7.95	5.62	5.09
	$\widehat{\mu}_4$	4.31	2.16	1.46	4.14	1.95	1.42	8.67	5.71	4.89
	$\widehat{\mu}_6$	20.62	15.39	13.24	19.29	14.55	12.71	24.09	15.72	12.25
PMAE	$\widehat{\mu}_1$	8.90	7.10	6.24	6.42	4.71	4.05	8.69	6.51	6.17
	$\widehat{\mu}_2$	16.52	13.94	12.43	12.72	10.50	9.25	7.92	5.93	5.67
	$\widehat{\mu}_3$	4.02	2.19	1.64	4.03	2.24	1.94	6.51	4.78	4.51
	$\widehat{\mu}_4$	3.32	1.59	1.10	3.19	1.54	1.12	7.08	4.71	4.11
	$\widehat{\mu}_6$	15.34	11.92	10.46	16.04	12.58	11.20	18.27	11.34	9.23
RIMSE	$\widehat{\mu}_1$	12.18	9.98	9.02	11.03	8.99	8.09	18.70	14.29	12.92
	$\widehat{\mu}_2$	14.69	12.28	11.12	12.07	9.69	8.58	16.49	12.26	11.16
	$\widehat{\mu}_3$	5.20	2.90	2.12	5.20	2.81	2.36	7.95	5.62	5.09
	$\widehat{\mu}_4$	4.31	2.16	1.46	4.14	1.95	1.42	8.67	5.71	4.89
	$\widehat{\mu}_6$	18.69	13.95	12.18	17.50	13.44	11.73	24.26	18.15	15.80
IMAE	$\widehat{\mu}_1$	10.07	8.42	7.63	9.38	7.59	6.77	14.00	10.99	10.34
	$\widehat{\mu}_2$	13.48	11.51	10.49	11.16	9.05	8.02	12.55	9.82	9.34
	$\widehat{\mu}_3$	4.02	2.19	1.64	4.03	2.24	1.94	6.51	4.78	4.51
	$\widehat{\mu}_4$	3.32	1.59	1.10	3.19	1.54	1.12	7.08	4.71	4.11
	$\widehat{\mu}_6$	13.47	10.51	9.37	14.18	11.31	10.03	18.34	13.35	11.88

Table 6. Estimation of conditional standard deviation in continuous design; 10,000 replications;

ESTIMATING FEATURES OF A DISTRIBUTION FROM BINOMIAL DATA - SUPPLEMENT*

Arthur Lewbel† Boston College

Daniel McFadden‡ University of California, Berkeley

Oliver Linton§ London School of Economics

July 2010

Abstract

This supplement contains additional results that, because of space constraints, could not be included in the main paper. These include an analysis of identification with discrete bids, and technical results required for deriving the limiting distributions provided in the main paper.

*This research was supported in part by the National Science Foundation through grants SES-9905010 and SBR-9730282, by the E. Morris Cox Endowment, and by the ESRC.

†Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA. Phone: (617) 552-3678. E-mail address: lewbel@bc.edu

‡Department of Economics, University of California, Berkeley, CA 94720-3880, USA. E-mail address: mcfadden@econ.berkeley.edu

§Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. E-mail address: o.linton@lse.ac.uk

1 Supplementary Appendix

1.1 Identification With Discrete Bids

The consistency of our estimators shows that moments $\mu_r(x) = E[r(W,X) \mid X = x]$ are nonparametrically identified, given our assumption that as $n \to \infty$, the distribution of V becomes dense in the support of W. As discussed in the introduction, nonparametric identification fails when the limiting support of V is a finite number of mass points, because the conditional distribution of $Y = I(W > V)$ given $X = x, V = v$ only identifies the distribution of $W|X = x$ at each support point v in the support of V, while $E[r(W,X) \mid X = x]$ depends on the distribution of $W|X = x$ at almost every support point w having a nonzero value of $r(w,x)$.

To further motivate our choice of nonparametric identifying assumptions, we show now that if the limiting support of V is a finite number of mass points, then nonparametric identification still fails even given an additive independent error model for W, that is, $W = m(X) - \varepsilon$ with $\varepsilon \perp X$. For simplicity in the proof it is assumed that X is a scalar, m is increasing in X, and V only takes on two values, but the basic logic can be extended to more general cases.

THEOREM. *Assume supp(X) is some open or closed interval on the real line, supp$(V) = \{-\delta, 0\}$ for some $\delta > 0$, and $W = m(X) - \varepsilon$ with ε having an unknown, strictly monotonic CDF $F_\varepsilon(\varepsilon)$ and m strictly monotonically increasing in X. Assume V, X, ε are mutually independent. Let $Y = I(W > V)$. The functions $m(x)$ and $F_\varepsilon(\varepsilon)$ are not identified given the distribution of Y conditional on V, X.*

PROOF OF THEOREM. Since Y is binary, the distribution of Y given X and V is $G(v \mid x) = E[Y \mid X = x, V = v] = F_\varepsilon[m(x) - v]$. Let $\zeta_0 = \inf[\text{supp}(X)]$, $m_0 = m(\zeta_0)$, and $\zeta_j = m^{-1}(m_0 + j\delta)$ for integers j. Let $\widetilde{m}(x)$ be any strictly monotonic function on $x \in [\zeta_0, \zeta_1]$ such that $\widetilde{m}(\zeta_0) = m_0$ and $\widetilde{m}(\zeta_1) = m_0 + \delta$. Define $\widetilde{F}_\varepsilon(\varepsilon)$ on $\varepsilon \in [m_0, m_0 + \delta]$ by $\widetilde{F}_\varepsilon(\varepsilon) = G[0 \mid \widetilde{m}^{-1}(\varepsilon)]$. Next, define $\widetilde{F}_\varepsilon(\varepsilon)$ on $\varepsilon \in (m_0 + \delta, m_0 + 2\delta]$ by $\widetilde{F}_\varepsilon(\varepsilon) = G[\delta \mid \widetilde{m}^{-1}(\varepsilon - \delta)]$, and define $\widetilde{m}(x)$ on $x \in (\zeta_1, \zeta_2]$ by $\widetilde{m}(x) = \widetilde{F}_\varepsilon^{-1}[G(0 \mid x)]$. Now define $\widetilde{F}_\varepsilon(\varepsilon)$ on $\varepsilon \in (m_0 + 2\delta, m_0 + 3\delta]$ by $\widetilde{F}_\varepsilon(\varepsilon) = G[\delta \mid \widetilde{m}^{-1}(\varepsilon - \delta)]$, and define $\widetilde{m}(x)$ on $x \in (\zeta_2, \zeta_3]$ by $\widetilde{m}(x) = \widetilde{F}_\varepsilon^{-1}[G(0 \mid x)]$. Continue on in this way until the support of x is exhausted By construction, the functions $\widetilde{m}$ and $\widetilde{F}_\varepsilon$ satisfy $G(v \mid x) = \widetilde{F}_\varepsilon[\widetilde{m}(x) - v]$ for all x and v on their support, and hence are observationally equivalent to $m(x)$ and $F_\varepsilon(\varepsilon)$. ■

Notes.

In this theorem, nothing can be identified about the function $m(x)$ (except possibly its endpoints) over the interval $x \in [\zeta_0, \zeta_1]$, since the observable data are consistent with $m(x)$ equalling any regular function over that interval, and the value of $m(x)$ in any other interval is identified only as a function of its unknown values in $[\zeta_0, \zeta_1]$.

The same proof could have been started by letting $\widetilde{F}_\varepsilon(\varepsilon)$ be any regular function with the correct endpoints on $\varepsilon \in [m_0, m_0 + \delta]$, then recovering the corresponding $\widetilde{m}$ on that interval, and proceeding as before. Therefore, the function $\widetilde{F}_\varepsilon$ is also completely unknown (except possibly at endpoints) over an initial interval, and it's values elsewhere are only recoverable as functions of its values in that interval.

The nonidentification here is not just an issue of location or scale. The proof assumes $m(x)$ may be known at two points, $m(\zeta_0)$ and $m(\zeta_1)$, which is equivalent to knowing (or choosing) a location and scale for $m(x)$. Similarly, the proof may be started by assuming $\widetilde{F}_\varepsilon(\varepsilon)$ is known at the two points and $\varepsilon = m_0$ and $\varepsilon = m_0 + \delta$, which is equivalent to knowing (or choosing) a location and scale for $\widetilde{F}$. These functions are therefore not identified up to location and scale.

Here $E[W \mid X = x] = m(x) - E(\varepsilon)$, so the nonidentification of $m(x)$ up to any location shows nonidentification of mean WTP. Other moments are likewise not identified.

This theorem can be applied to show nonidentification of other closely related models. In particular, it implies nonidentification of the nonparametric ordered choice model $Y = jI(\alpha_j < m(x) - \varepsilon \leq a_{j+1})$ for a set of integers j and threshold constants α_j (two of which can be normalized to zero and one to pin down the location of ε and the scaling of both ε and m) It also shows nonidentification of the model considered by Das (2002), in which $W = m(x) - \varepsilon$ and one only observes which of a few different fixed intervals each observation W lies in. With a partial parameterization, this model is what An (2000) and others call a double bounded dichotomous choice.

It follows from the consistency of our estimator $\widehat{\mu}_{4r}(x)$ (with, e.g., θ estimated using Klein and Spady 1993) that this model can be identified with a fixed discrete design V if $m(x)$ above is parameterized as $m(x, \theta)$ with a known function m and finite parameter vector θ. In this semiparametric specification, continuity of X takes the place of continuity of V.

The implications of this Theorem for bid design differ markedly from results on optimal bid design in parametric or semiparametric models. Summarizing Kanninen (1993), Crooker and Herriges (2004) say, in referring to parametric or semiparametric models "estimates of the mean WTP are best with relatively few bid levels."

Some existing estimators implicitly assume identification, such as the sieve estimators proposed

by Chen and Randall (1997) and Das (2002), which they apply to data in which v can only take on a finite number of values. This Theorem shows that such models are generally not identified.[1]

1.2 Subsidiary Results

Define $F_n(\theta) = n^{-1}\sum_{i=1}^n f(Z_i,\theta)$ for some function f, and let $F(\theta) = EF_n(\theta)$ and $\Gamma_F = \partial F(\theta_0)/\partial\theta$.

LEMMA 1. *Assume:*

(i) For some vector ς

$$\sqrt{n}(\widehat{\theta}-\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varsigma(Z_i,\theta_0) + o_p(1)$$

where $E[\varsigma(Z_i,\theta_0)] = 0$ *and* $\Omega = E[\varsigma(Z_i,\theta_0)\varsigma(Z_i,\theta_0)^\top] < \infty$.

(ii) There exists a finite matrix Γ_F *of full (column) rank such that*

$$\lim_{\|\theta-\theta_0\|\to 0} \frac{\|F(\theta)-\Gamma_F(\theta-\theta_0)\|}{\|\theta-\theta_0\|} = 0.$$

(iii) For every sequence of positive numbers $\{\delta_n\}$ *such that* $\delta_n \to 0$,

$$\sup_{\|\theta-\theta_0\|\le\delta_n} \left\|\sqrt{n}[F_n(\theta)-F(\theta)] - \sqrt{n}[F_n(\theta_0)-F(\theta_0)]\right\| = o_p(1).$$

Then

$$\sqrt{n}[F_n(\widehat{\theta})-F(\theta_0)] \Longrightarrow N(0,V), \text{ where}$$

$$V = \mathrm{var}[\Gamma_F\varsigma(Z_i,\theta_0) + f(Z_i,\theta_0)] = \Gamma_F\Omega{\Gamma_F}^\top + \mathrm{var}[f(Z_i,\theta_0)] + 2\Gamma_F E\varsigma(Z_i,\theta_0)f(Z_i,\theta_0).$$

See below for a discussion on the verification of (iii).

PROOF. Since $\widehat{\theta}$ is root-n consistent, there exists a sequence $\delta_n \to 0$ such that $\Pr[||(\widehat{\theta}-\theta_0)|| > \delta_n] \to 0$ as $n\to\infty$. We can therefore suppose that $||(\widehat{\theta}-\theta_0)|| \le \delta_n$ with probability tending to one.

[1] Their estimators essentially smooth between the different available test values v to obtain results with uncertain limiting values. Our nonparametric estimators also smooth between test values in an analogous way, but consistency is obtained by having the available bids become dense in the support of W. Crooker and Herriges' (2000) monte carlo design, which we also use, employs this feature of an increasingly fine grid of test values. An (2000) provides a semiparametric model that identifies and estimates the W distribution only at the available bid levels, and explicitly interpolates these estimates to obtain a generally inconsistent estimate of W at the mean.

We have

$$\begin{aligned}
\sqrt{n}[F_n(\widehat{\theta}) - F(\theta_0)] &= \sqrt{n}[F(\widehat{\theta}) - F(\theta_0)] + \sqrt{n}[F_n(\widehat{\theta}) - F(\widehat{\theta})] \\
&= \Gamma_F \sqrt{n}(\widehat{\theta} - \theta_0) + \sqrt{n}[F_n(\theta_0) - F(\theta_0)] + o(\|[\sqrt{n}(\widehat{\theta} - \theta_0)\|) \\
&\quad + \sqrt{n}\{[F_n(\widehat{\theta}) - F(\widehat{\theta})] - [F_n(\theta_o) - F(\theta_0)]\} \\
&= \Gamma_F \sqrt{n}(\widehat{\theta} - \theta_0) + \sqrt{n}[F_n(\theta_0) - F(\theta_0)] + o_p(1)[\text{by (ii) and (iii)}] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\Gamma_F \varsigma(Z_i, \theta_0) + [f(Z_i, \theta_0) - Ef(Z_i, \theta_0)]\} + o_p(1),
\end{aligned}$$

and the result now follows from standard CLT for independent random variables. ■

LEMMA 2. *Suppose that assumptions C1-C3 hold. Then, as* $n \to \infty$

$$\max_{1 \le i \le n} \left| \widehat{\psi}(\widehat{U}_i) - \psi(U_i) - \frac{1}{n} \sum_{j=1}^{n} L(Z_i, Z_j) \right| = o_p(n^{-1/2}) \tag{1}$$

$$\max_{1 \le i \le n} \left| \frac{1}{n} \sum_{j=1}^{n} L(Z_i, Z_j) \right| = O_p(n^{-1/2}) \tag{2}$$

$$\min_{1 \le i \le n} \widehat{\psi}(\widehat{U}_i) \ge \inf_{u \in \mathcal{U}} \psi(u) + o_p(1) \tag{3}$$

where $\mathcal{U}$ *is the support of* U_i, $L(Z_i, Z_j) = \xi_j(U_i) + \Gamma(Z_i)\varsigma(Z_j; \theta_0)$, *and:*

$$\begin{aligned}
\Gamma(Z_i) &= E_i \left[\zeta_{ij} \frac{\partial m(X_j, \theta_0)}{\partial \theta} \right] - E_i[\zeta_{ij}] \frac{\partial m(X_i, \theta_0)}{\partial \theta} \\
\xi_j(U_i) &= h(\Lambda|X_j)\Lambda'(m(X_j, \theta_0) - U_i) - E_i \left[h(\Lambda|X_j)\Lambda'(m(X_j, \theta_0) - U_i) \right] \\
\zeta_{ij} &= \left(\left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda'' \right] (m(X_j, \theta_0) - U_i) \right).
\end{aligned}$$

PROOF. Regarding (2), the pointwise rate follows by standard central limit theorem for each $Z_i = z$: we have $EL(z, Z_j) = 0$ for each z and $\sup_z \text{var} L(z, Z_j) < \infty$. Then because the function $L(z, Z_j)$ is bounded Lipschitz, the uniformity over z follows from FCLT.

Result (3) follows by an application of the triangle inequality $\min_{1 \le i \le n} \psi(U_i) \le \min_{1 \le i \le n} \widehat{\psi}(\widehat{U}_i) + \max_{1 \le i \le n} |\widehat{\psi}(\widehat{U}_i) - \psi(U_i)|$, and the fact that $\max_{1 \le i \le n} |\widehat{\psi}(\widehat{U}_i) - \psi(U_i)| = o_p(1)$ as a consequence of (1) and (2).

Before showing (1) we show that:

$$\max_{1\leq i\leq n}\widehat{U}_i \leq \max_{1\leq i\leq n} U_i + o_p(1) \tag{4}$$

$$\min_{1\leq i\leq n}\widehat{U}_i \geq \min_{1\leq i\leq n} U_i + o_p(1), \tag{5}$$

from which it follows that we can ignore the possibility that $\widehat{U}_i$ lies outside of the support of U_i, i.e., for any event A

$$\begin{aligned}\Pr[\mathcal{A}] &\leq \Pr\left[\mathcal{A} \text{ and } \{\widehat{U}_1,\ldots,\widehat{U}_n\}\subset\mathcal{U}\right] + \Pr\left[\widehat{U}_j\notin\mathcal{U} \text{ for some } j\right]\\ &\leq \Pr\left[\mathcal{A} \text{ and } \{\widehat{U}_1,\ldots,\widehat{U}_n\}\subset\mathcal{U}\right] + o(1) = o(1).\end{aligned} \tag{6}$$

Proof of (4). We have

$$\widehat{U}_i = U_i + \frac{\partial m}{\partial\theta}(X_i,\overline{\theta})(\widehat{\theta}-\theta_0)$$

by the mean value theorem, where $\overline{\theta}$ are intermediate values between $\widehat{\theta}$ and θ_0. Since $\widehat{\theta}$ is root-n consistent, there exists a sequence $\delta_n\to 0$ such that $\Pr[||\widehat{\theta}-\theta_0||\geq\delta_n]\to 0$. Therefore, with probability tending to one

$$|\frac{\partial m}{\partial\theta}(X_i,\overline{\theta})| \leq \sup_{||\theta-\theta_0||\leq\delta_n}|\frac{\partial m}{\partial\theta}(X_i,\theta)| \leq d_1(X_i).$$

Furthermore, applying the Bonferroni and Markov inequalities

$$\Pr\left[\max_{1\leq i\leq n} d_1(X_i) > \epsilon\sqrt{n}\right] \leq n\Pr\left[d_1(X_i)>\epsilon\sqrt{n}\right] \leq n\frac{Ed_1^r(X_i)}{(\epsilon\sqrt{n})^r} = o(1)$$

for any $\epsilon>0$ when $r>2$. This yields (4); (5) follows similarly.

We next prove (1). Define the stochastic process in θ

$$\widehat{\psi}(U_i(\theta)) = \frac{1}{n}\sum_{j=1}^n \rho_j(U_i(\theta),\theta).$$

Then by Taylor expansion

$$\widehat{\psi}(\widehat{U}_i) - \widehat{\psi}(U_i) = \frac{1}{n}\sum_{j=1}^n \frac{\partial\rho_j(U_i(\theta_0),\theta_0)}{\partial\theta^\top}(\widehat{\theta}-\theta_0) + R_{ni}, \tag{7}$$

where the derivative inside the summation is a total derivative defined below, while

$$R_{ni} = \frac{1}{2n}\sum_{j=1}^n (\widehat{\theta}-\theta_0)^\top \frac{\partial^2\rho_j(U_i(\overline{\theta}),\overline{\theta})}{\partial\theta\partial\theta^\top}(\widehat{\theta}-\theta_0),$$

where $\overline{\theta}$ are intermediate values between $\widehat{\theta}$ and θ_0, while:

$$\begin{aligned}
\frac{\partial \rho_j(U_i(\theta),\theta)}{\partial \theta} &= \left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda''\right](m(X_j,\theta) - U_i(\theta))\left[\frac{\partial m(X_j,\theta)}{\partial \theta} - \frac{\partial m(X_i,\theta)}{\partial \theta}\right] \\
\frac{\partial^2 \rho_j(U_i(\theta),\theta)}{\partial \theta \partial \theta^\top} &= \left[h''(\Lambda|X_j)(\Lambda')^3 + 3h'(\Lambda|X_j)\Lambda'\Lambda'' + h(\Lambda|X_j)\Lambda'''\right](m(X_j,\theta) - U_i(\theta)) \\
&\quad \times \left[\frac{\partial m(X_j,\theta)}{\partial \theta} - \frac{\partial m(X_i,\theta)}{\partial \theta}\right]\left[\frac{\partial m(X_j,\theta)}{\partial \theta} - \frac{\partial m(X_i,\theta)}{\partial \theta}\right]^\top \\
&\quad + \left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda''\right](m(X_j,\theta) - U_i(\theta))\left[\frac{\partial^2 m(X_j,\theta)}{\partial \theta \partial \theta^\top} - \frac{\partial^2 m(X_i,\theta)}{\partial \theta \partial \theta^\top}\right].
\end{aligned}$$

Applying (6) we have in (7) that with probability tending to one

$$|R_{ni}| \leq ||\widehat{\theta} - \theta_0||^2 \times \frac{1}{n}\sum_{j=1}^n \sup_{||\theta-\theta_0||\leq \delta_n} \left\| \frac{\partial^2 \rho_j(U_i(\theta),\theta)}{\partial \theta \partial \theta^\top} \right\| \leq O_p(n^{-1}) \times \frac{1}{n}\sum_{j=1}^n D(X_i, X_j)$$

for some measurable function D with finite mean. Therefore, $\max_{1\leq i\leq n}|R_{ni}| = o_p(n^{-1/2})$. We then show that

$$\max_{1\leq i\leq n}\left|\frac{1}{n}\sum_{j=1}^n \frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta^\top} - E_i\left[\frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta^\top}\right]\right| = o_p(1).$$

The pointwise limit follows by the law of large numbers, and the uniformity is obtained by another application of the Bonferroni and Markov inequalities. Therefore, uniformly in i

$$\widehat{\psi}(\widehat{U}_i) - \widehat{\psi}(U_i) = E_i\left[\frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta^\top}\right](\widehat{\theta} - \theta_0) + o_p(n^{-1/2}).$$

We have

$$E_i\left[\frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta^\top}\right] = E_i\left[\zeta_{ij}\frac{\partial m(X_j,\theta_0)}{\partial \theta}\right] - E_i[\zeta_{ij}]\frac{\partial m(X_i,\theta_0)}{\partial \theta}.$$

This is because by the chain rule

$$\begin{aligned}
\frac{\partial \rho_j(U_i(\theta_0),\theta_0)}{\partial \theta} &= \frac{\partial \rho_j(u,\theta)}{\partial \theta}\Big|_{\theta=\theta_0,u=U_i(\theta_0)} + \frac{\partial \rho_j(u,\theta_0)}{\partial u}\Big|_{\theta=\theta_0,u=U_i(\theta_0)}\frac{\partial U_i(\theta)}{\partial \theta}\Big|_{\theta=\theta_0} \\
&= -\frac{\partial \rho_j(u,\theta_0)}{\partial u}\Big|_{\theta=\theta_0,u=U_i(\theta_0)}\left[\frac{\partial m(X_j,\theta_0)}{\partial \theta} - \frac{\partial m(X_i,\theta_0)}{\partial \theta}\right],
\end{aligned}$$

where $\partial \rho_j(u,\theta_0)/\partial u$ was defined in (**??**).

■

LEMMA 3. *Suppose that assumptions C1-C4 hold. Then with probability tending to one for some function d with finite r moments:*

$$\max_{1\leq i\leq n}\left|\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)-\frac{1}{n}\sum_{j=1}^{n}L^*(Z_i,Z_j)\right| \leq \frac{k}{nb^3}d(X_i) \tag{8}$$

$$\max_{1\leq i\leq n}\left|\frac{1}{n}\sum_{j=1}^{n}L^*(Z_i,Z_j)\right| = O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2}\right\}+O_p(b^2) \tag{9}$$

$$\min_{1\leq i\leq n}\widetilde{\psi}(\widehat{U}_i) \geq \inf_{u\in\mathcal{U}}\psi(u)+o_p(1) \tag{10}$$

where

$$L^*(Z_i,Z_j)=\frac{1}{b}k\left(\frac{U_i-U_j}{b}\right)-\psi(U_i)+\Gamma^*(Z_i)\cdot\varsigma(Z_j,\theta_0)$$

$$\Gamma^*(Z_i)=\psi'(U_i)\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0)-E\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0)\mid U_i\right]\right]-\psi(U_i)\overline{m}'_\theta(U_i).$$

PROOF. Define

$$\overline{\psi}(U_i)=\frac{1}{nb}\sum_{j=1}^{n}k\left(\frac{U_i-U_j}{b}\right).$$

Making a second order Taylor series expansion we have

$$\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)=T_{ni}+R_{ni}, \tag{11}$$

where

$$T_{ni}=\overline{\psi}(U_i)-\psi(U_i)+\frac{1}{nb^2}\sum_{j=1}^{n}k'\left(\frac{U_i-U_j}{b}\right)\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0)-\frac{\partial m}{\partial\theta^\top}(X_j,\theta_0)\right](\widehat{\theta}-\theta_0)$$

$$\begin{aligned}R_{ni} &= \frac{1}{2nb^3}\sum_{j=1}^{n}k''\left(\frac{U_i^*-U_j^*}{b}\right)\left[\frac{\partial m}{\partial\theta}(X_i,\theta_0)-\frac{\partial m}{\partial\theta}(X_j,\theta_0)\right](\widehat{\theta}-\theta_0)(\widehat{\theta}-\theta_0)^\top\\ &\quad\times\left[\frac{\partial m}{\partial\theta}(X_i,\theta_0)-\frac{\partial m}{\partial\theta}(X_j,\theta_0)\right]^\top\\ &\quad+\frac{1}{nb^2}\sum_{j=1}^{n}k'\left(\frac{U_i-U_j}{b}\right)(\widehat{\theta}-\theta_0)^\top\left[\frac{\partial^2 m}{\partial\theta\partial\theta^\top}(X_i,\theta^*)-\frac{\partial^2 m}{\partial\theta\partial\theta^\top}(X_j,\theta^*)\right](\widehat{\theta}-\theta_0),\end{aligned}$$

where θ^* are intermediate values between $\widehat{\theta}$ and θ_0, and $U_i^*=U_i(\theta^*)$.

We first show that the remainder terms are of smaller order. We have with probability tending to one

$$\begin{aligned}|R_{ni}| &\leq b^{-3}\sup_u |k''(u)| \cdot ||\widehat{\theta}-\theta_0||^2 \cdot \left(\left\| \frac{\partial m}{\partial \theta}(X_i,\theta_0) \right\|^2 + \frac{1}{n}\sum_{j=1}^n \left\| \frac{\partial m}{\partial \theta}(X_j,\theta_0) \right\|^2 \right) \\ &\quad + b^{-1}||\widehat{\theta}-\theta_0||^2 \cdot \frac{1}{nb}\sum_{j=1}^n \left| k'\left(\frac{U_i-U_j}{b}\right)\right| (d_1(X_i)+d_2(X_j))\end{aligned}$$

by the Cauchy-Schwarz inequality. Since the function $|k'(u)|$ is Lipschitz continuous, we can apply the uniform convergence results of Masry (1996a):

$$\begin{aligned}\max_{1\leq i\leq n}\left| \frac{1}{nb}\sum_{j=1}^n \left| k'\left(\frac{U_i-U_j}{b}\right)\right| - E_i\left[\left| k'\left(\frac{U_i-U_j}{b}\right)\right|\right]\right| &= O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2}\right\} \\ \max_{1\leq i\leq n}\left| \frac{1}{nb}\sum_{j=1}^n \left| k'\left(\frac{U_i-U_j}{b}\right)\right| d_2(X_j) - E_i\left| k'\left(\frac{U_i-U_j}{b}\right)\right| d_2(X_j)\right| &= O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2}\right\},\end{aligned}$$

since $E[d_2^r(X_j)] < \infty$. Furthermore,

$$\begin{aligned}E_i\left[\frac{1}{b}\left| k'\left(\frac{U_i-U_j}{b}\right)\right|\right] &= \int |k'(t)|\,\psi(U_i+tb)dt \\ E_i\left[\left| k'\left(\frac{U_i-U_j}{b}\right)\right| d_2(X_j)\right] &= \int |k'(t)|\,\overline{d}_2(U_i+tb)\psi(U_i+tb)dt\end{aligned}$$

are uniformly bounded, where $\overline{d}_2(U_i) = E[d_2(X_i)|U_i]$. Therefore, for suitable constants and dominating functions

$$|R_{ni}| \leq \frac{k_1}{nb^3}(d_3(X_i)+k_2) + \frac{k_3}{nb}(d_1(X_i)+k_4)$$

with probability tending to one. This gives the result. Furthermore, we have $\max_{1\leq i\leq n} d_l(X_i) = O_p(n^{1/r})$, so that $\max_{1\leq i\leq n}|R_{ni}| = O_p(n^{-1}b^{-3}n^{1/r})$. Provided $n^{(r-2)/r}b^6 \to \infty$, this term is $o_p(n^{-1/2})$. With additional smoothness conditions on k and m, this condition can be substantially weakened.

We now turn to the leading term T_{ni}. By the Masry (1996a) results

$$\max_{1\leq i\leq n}\left| \frac{1}{nb^2}\sum_{j=1}^n k'\left(\frac{U_i-U_j}{b}\right) d(X_j) - E_i\left[\frac{1}{b^2}k'\left(\frac{U_i-U_j}{b}\right)\overline{d}(U_j)\right]\right| = O_p\left\{\left(\frac{\log n}{nb^3}\right)^{1/2}\right\}, \qquad (12)$$

for any function d with finite moments, where $\overline{d}(U_j) = E[d(X_j)|U_j]$. Under our bandwidth conditions

this term is $o_p(1)$. Furthermore, for any twice continuously differentiable function $\overline{d}(u)$ we have

$$\left| E\left[\frac{1}{b^2}k'\left(\frac{U_i-U_j}{b}\right)\overline{d}(U_j)\mid U_i\right] - [\overline{d}(U_i)\psi(U_i)]'\right| \tag{13}$$

$$= \left|\int \frac{1}{b^2}k'\left(\frac{U_i-u}{b}\right)\overline{d}(u)\psi(u)du - [\overline{d}(U_i)\psi(U_i)]'\right|$$

$$= \left|\int \frac{1}{b}k\left(\frac{U_i-u}{b}\right)[\overline{d}(u)\psi(u)]'du - [\overline{d}(U_i)\psi(U_i)]'\right|$$

$$= \left|\int k(t)\left([\overline{d}(U_i+tb)\psi(U_i+tb)]' - [\overline{d}(U_i)\psi(U_i)]'\right)dt\right|$$

$$= O_p(b^2)$$

by integration by parts, change of variables and dominated convergence using the symmetry of k. This order is uniform in i by virtue of the boundedness and continuity of the relevant functions. In (12) and (13) take $\overline{d}(u)=1$ and $\overline{d}(u)=\overline{m}_\theta(u)$, and note that $[\overline{d}(U_i)\psi(U_i)]' = \overline{d}'(U_i)\psi(U_i)+\overline{d}(U_i)\psi'(U_i)$. Therefore,

$$\frac{1}{nb^2}\sum_{j=1}^n k'\left(\frac{U_i-U_j}{b}\right)\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0) = \frac{\partial m}{\partial\theta^\top}(X_i,\theta_0)\psi'(U_i)+o_p(1)$$

$$\frac{1}{nb^2}\sum_{j=1}^n k'\left(\frac{U_i-U_j}{b}\right)\frac{\partial m}{\partial\theta^\top}(X_j,\theta_0) = \overline{m}'_{\theta_0}(U_i)\psi(U_i)+\overline{m}_{\theta_0}(U_i)\psi'(U_i)+o_p(1)$$

uniformly in i.

In conclusion,

$$\max_{1\le i\le n}|T_{ni}-\frac{1}{n}\sum_{j=1}^n L^*(Z_i,Z_j)| = o_p(n^{-1/2})\ ;\ \max_{1\le i\le n}|R_{ni}| = o_p(n^{-1/2}),$$

which gives the first part of the lemma. Also, we have

$$\max_{1\le i\le n}|\frac{1}{n}\sum_{j=1}^n L^*(Z_i,Z_j)| = O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2}\right\}+O_p(b^2),$$

by the Masry results.

The proof of (10) follows as for (3).

■

1.3 Stochastic Equicontinuity Results

We now show that condition (iii) of Lemma 1 is satisfied in our case. Let $\Theta_n(c) = \{\theta\text{: } \sqrt{n}|\theta - \theta^0| \leq c\}$. Since $\sqrt{n}(\widehat{\theta} - \theta^0) = O_p(1)$, for all $\epsilon > 0$ there exists a c_ϵ and an integer n_0 such that for all $n \geq n_0$, $\Pr[\widehat{\theta} \in \Theta_n(c_\epsilon)] \geq 1 - \epsilon$ so without loss of generality we can assume that c is fixed at some large number and suppress dependence on c in the sequel. Define the stochastic process

$$\nu_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(Z_i, \theta) - E[f(Z_i, \theta)], \quad \theta \in \Theta,$$

where

$$f(Z_i, \theta) = r[\Lambda(m(x, \theta)), x] + \frac{r'[\Lambda(m(x, \theta) - U_i(\theta)), x]\Lambda'(m(x, \theta) - U_i(\theta))[Y_i - 1(U_i(\theta) > 0)]}{\psi(U_i)}$$

and define the pseudo-metric $\rho(\theta, \theta') = E([f(Z_i, \theta) - f(Z_i, \theta')]^2)$, on Θ. Under this metric, the parameter space Θ is totally bounded. We are only interested in the behaviour of this process as θ varies in the small set Θ_n. By writing $\theta = \theta^0 + \gamma n^{-1/2}$, we shall make a reparameterization to $\nu_n(\gamma)$, where $\gamma \in \Gamma \subset \mathbb{R}^p$ with $\Gamma = \{\gamma : |\gamma| \leq c\}$. We establish the following result:

$$\sup_{\gamma \in \Gamma} |\nu_n(\gamma) - \nu_n(0)| = o_p(1) \tag{14}$$

To prove (14) it is sufficient to show a pointwise law of large numbers, e.g., $\nu_n(\gamma) - \nu_n(0) = o_p(1)$ for any $\gamma \in \Gamma$, and stochastic equicontinuity of the process ν_n at $\gamma = 0$. The pointwise result is immediate because the random variables are sums of i.i.d. random variables with finite absolute moment and zero mean; the probability limit of $\nu_n(\gamma)$ is the same for all $\gamma \in \Gamma$ by the smoothness of the expected value in γ. To complete the proof of (14) we shall use the following lemma, proved below, which states that ν_n is stochastically equicontinuous in θ. The difficulty in establishing the required equicontinuity arises solely because the function m inside U is nonlinear in θ.

LEMMA SE. *Under the above assumptions, the process $\nu_n(\gamma)$ is stochastically equicontinuous, i.e., for all $\epsilon > 0$ and $\eta > 0$, there exists $\delta > 0$ such that*

$$\limsup_{n \to \infty} \Pr\left[\sup_{\rho(t_1, t_2) < \delta} |\nu_n(t_1) - \nu_n(t_2)| > \eta\right] < \epsilon.$$

PROOF OF LEMMA SE. By a second order Taylor series expansion of $m(Z_i, \theta)$ around $m(Z_i, \theta^0)$:

$$m(Z_i, \theta^0 + \gamma n^{-1/2}) = m(Z_i, \theta^0) + \frac{1}{\sqrt{n}} \sum_{k=1}^{p} \frac{\partial m}{\partial \theta_k}(Z_i, \theta^0)\gamma_k + \frac{1}{n} \sum_{k=1}^{p} \sum_{r=1}^{p} \frac{\partial^2 m}{\partial \theta_k \partial \theta_r}(Z_i; \overline{\theta})\gamma_k \gamma_r \qquad (15)$$

for some intermediate points $\overline{\theta}$. Define the linear approximation to $m(Z_i, \theta^0 + \gamma n^{-1/2})$,

$$T(Z_i, \gamma) = m(Z_i, \theta^0) + \sum_{k=1}^{p} \frac{\partial m}{\partial \theta_k}(Z_i, \theta^0)\gamma_k$$

for any γ. By assumption C2, for all k, r, $\sup_{\theta \in \Theta} |\partial^2 m(Z_i, \theta)/\partial \theta_k \partial \theta_r|^2 \leq d(Z_i)$ with $Ed(Z_i) < \infty$. Therefore, for all $\delta > 0$ there exists an $\varepsilon > 0$ such that

$$\begin{aligned} \Pr\left[\frac{1}{\sqrt{n}} \max_{i,k,r} \sup_{\theta \in \Theta_n} \left|\frac{\partial^2 m}{\partial \theta_k \partial \theta_r}(Z_i, \theta)\right| > \varepsilon\right] &\leq n \sum_{k,r} \Pr\left[\frac{1}{\sqrt{n}} \sup_{\theta \in \Theta_n} \left|\frac{\partial^2 m}{\partial \theta_k \partial \theta_r}(Z_i, \theta)\right| > \varepsilon\right] \\ &\leq \frac{\sum_{k,r} E[d(Z_i)]}{\varepsilon^2} \leq \delta, \end{aligned}$$

by the Bonferroni and Chebychev inequalities. Therefore, with probability tending to one

$$\max_{1 \leq i \leq n} \left|\frac{1}{n} \sum_{k=1}^{p} \sum_{r=1}^{p} \frac{\partial^2 m}{\partial \theta_k \partial \theta_r}(Z_i; \overline{\theta})\gamma_k \gamma_r\right| \leq \frac{\overline{\pi}}{\sqrt{n}}$$

for some $\overline{\pi} < \infty$. Define the stochastic process

$$\nu_{n1}(\gamma, \pi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \overline{f}(Z_i, \theta_0 + \gamma n^{-1/2}, \pi n^{-1/2}) - E\overline{f}(Z_i, \theta_0 + \gamma n^{-1/2}, \pi n^{-1/2})$$

on $\gamma \in \Gamma$ and $\pi \in \Pi = [0, \overline{\pi}]$, where

$$\begin{aligned} & \overline{f}(Z_i, \theta_0 + \gamma n^{-1/2}, \pi n^{-1/2}) \\ = \; & r[\Lambda(m(x, \theta_0 + \gamma n^{-1/2})), x] \\ & + \frac{r'[\Lambda(m(x, \theta_0 + \gamma n^{-1/2}) - U_i(\theta_0 + \gamma n^{-1/2})), x]\Lambda'(m(x, \theta_0 + \gamma n^{-1/2}) - U_i(\theta_0 + \gamma n^{-1/2}))}{\psi(U_i)} \\ & \times [Y_i - 1(T(Z_i, \gamma n^{-1/2}) + \frac{\pi}{\sqrt{n}} - \Lambda^{-1}(V_i) > 0)] \end{aligned}$$

It suffices to show that $\nu_{n1}(\gamma, \pi)$ is stochastically equicontinuous in γ, π, and the deterministic centering term is of smaller order. The latter argument is a standard Taylor expansion. The argument for $\nu_{n1}(\gamma, \pi)$ is very similar to that contained in Sherman (1993) because we have a linear index structure in this part. One can apply Lemma 2.12 in Pakes and Pollard (1989). ■