ESTIMATING FEATURES OF A DISTRIBUTION FROM BINOMIAL DATA*

Arthur Lewbel† Oliver Linton‡
Boston College London School of Economics

Daniel McFadden§
University of California, Berkeley

Revised September, 2003

Abstract

A statistical problem that arises in several fields is that of estimating the features of an unknown distribution, which may be conditioned on covariates, using a sample of binomial observations on whether draws from this distribution exceed threshold levels set by experimental design. Applications include bioassay and destructive duration analysis. The empirical application we consider is referendum contingent valuation in resource economics, where one is interested in features of the distribution of values (willingness to pay) placed by consumers on a public good such as endangered species. Sample consumers are asked whether they favor a referendum that would provide the good at a cost specified by experimental design. This paper provides estimators for moments and quantiles of the unknown distribution in this problem under both nonparametric and semiparametric specifications.

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.

*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, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. E-mail address: lintono@lse.ac.uk

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

1 Introduction

A statistical problem that arises in several fields is that of estimating the features of an unknown distribution, which may be conditioned on covariates, using a sample of binomial observations on whether draws from this distribution exceed threshold levels set by experimental design. Consider estimating features of the distribution of some household economic variable W such as wealth, or the willingness to pay (WTP) for a good or resource such as a change in environmental quality. To minimize response bias, each subject i is asked if their W_i exceeds a test value V_i chosen by experimental design.[1] An observation consists of the test value or bid V_i that is posed to subject i, covariates X_i (such as the subject's age, income level, geographic location, or political party affiliation) and a binary indicator Y_i which equals one in the event that W_i exceeds V_i, and zero otherwise, so $Y_i = I(W_i > V_i)$, where $I(\cdot)$ is the indicator function. Objects of interest might include the moments of the distribution of wealth among individuals with certain observable characteristics such as demographics and education level, or the mean, variance and (for median voter models) median willingness-to-pay for a resource among individuals with characteristics like income level, party affiliation, and geographic location, that make them likely voters. Other statistical problems that have the same structure include some forms of bioassay[2] and destructive testing.[3]

Many parametric and semiparametric estimators of the distribution of W exist. See, e.g., Kanninen (1993) and Crooker and Herriges (2000) for comparisons of various, mostly parametric, WTP

[1] In many studies, follow up queries are used to gain more information about W, however, we will not consider the use of follow up data, because follow up responses may be shadowed by the framing effect of the first bid. This shadowing effect is common in unfolding bracket survey questions on economic variables, and on stated WTP for economic goods. McFadden (1994) provides references and experimental evidence that responses to follow up test values can be biased. There are additional issues of the impact of framing of questions on 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. The difficult problem of deconvoluting a target distribution in the presence of psychometric errors is left for future research.

[2] In bioassay the goal is estimation of features of the distribution of survival time W until the onset of an abnormality in laboratory animals exposed to an environmental hazard. The animals are sacrificed at times determined by experimental design, and tested for the abnormality. An observation consists of a vector of covariates X such as attributes of the animal and the exposure, a time V at which the animal is sacrificed for testing, and an indicator Y for whether the test reveals the presence of the abnormality at time V.

[3] An example of destructive testing would be estimation of features of the distribution of speeds W at which car safety device fails. At speeds selected by experimental design, drive cars into a barrier and determine whether a dummy occupant is injured. An observation consists of covariates X (attributes of the car, device, and dummy) a speed V at which the car is tested, and an indicator Y for injury to the test dummy.

estimators. WTP estimators that are not fully parameterized include Chen and Randall (1997), Creel and Loomis (1997), and An (2000), and for bioassay, Ramgopal, Laud, and Smith (1993), and Ho and Sen (2000). We propose new nonparametric and semiparametric estimators for conditional (on covariates) moments of the unknown W distribution, and we provide estimators of conditional quantiles of the unknown distribution.

A common estimation method 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 in a variety of ways, especially when the model is not fully parameterized. One difference is that the primary goals of our estimators are moments or quantiles of W, rather than the 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 Klein and Spady (1993), is inadequate for estimation of moments of W. Another important difference is the presence of a covariate V that is determined by experimental design. We exploit this feature of the data in the construction of our estimators.

Experimental design may depend on sample size. Our estimators explicitly allow for this dependence, which turns out to be crucial for nonparametric identification. Given $Y = I(W > V)$, the distribution function of Y equals $E(Y \mid V = v, X = x)$, which in turn equals the conditional distribution of $-W$, evaluated at V, conditioned on $X = x$. It follows that without additional modeling assumptions, the distribution of W can only be identifed on the support of V, and therefore moments of W are not identified when the support of V is limited. We show in an appendix that, given a fixed discrete design for V, 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.

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.[4] To obtain nonparametric identification, we therefore assume that if the bid or test value V distribution is discrete, then the number of mass points of this distribution grows with the sample size, eventually becoming dense in the support of W.[5] We also show how this dependence of survey design on sample

[4]See, e.g., Crooker and Herriges (2000) for a study of WTP bid designs, with explicit consideration of varying numbers of mass points.

[5]We also provide an alternative identifying assumption based on a semiparametric specification of W. Other possible identifying assumptions might include homogeneity as in Matzkin's (1992) threshold crossing model, or An's (2000) model which assumes W is an unknown monotonic transformation of $X^\top\theta_0 + \varepsilon$ with the distribution of ε known. See also Manski and Tamer (2002) and Das (2002) for related results, since V can be interpreted as providing

size affects the resulting limiting distributions. In Monte Carlos and in an empirical application with discrete bid distributions, we find that the estimators we propose perform reasonably, as long as the number of mass points is not too small.

We consider estimation for a few different information sets. In the most general case, the distribution of $W|X$ is completely unspecified apart from smoothness, and is nonparametrically estimated. We may write this case as $W = m(X, \varepsilon)$ for an unobserved ε. This includes as a special case, and is strictly weaker than, the location model $W = m(X) - \varepsilon$, where the function m and the distribution of ε are unknown. The second case we analyze is the semiparametric model $W = \Lambda[m(X, \theta_0) - \varepsilon]$ for known functions m and Λ, where the parameters θ_0 and the distribution of $\varepsilon \perp X$ are unknown. In this semiparametric model, identification requires that the support of $m(X, \theta_0) - \Lambda^{-1}(V)$ become dense in the support of ε, so in this semiparametric case identification is possible with a fixed, discrete design for V, given the presence of a continuously distributed element of X.

In either of these two cases (nonparametric or semiparametric W), the asymptotic design distribution of the test value V may either be known or unknown to the researcher, which yields a total of four different estimation scenarios. We provide estimators, and associated limiting normal distributions, for each of these four situations, since each is relevant for some applications. We also provide Monte Carlo analyses of the estimators, and an empirical application estimating conditional mean WTP to protect wetland habitats in California's San Joaquin Valley.

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 to equal one in the event that W exceeds V, and zero otherwise, so $Y = I(W > V)$ where $I(\cdot)$ 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 $\mu_r(x) = E[r(W, X) \mid$

(unbounded) interval observations of W.

$X = x]$ for any chosen x in the support of X. Of particular interest are the moments based on $r(W, X) = W^k$ for integers k. In addition to moments we also consider estimation of quantiles.

If the conditional distribution of W given $X = x$ is finitely parameterized, then those parameters can generally be efficiently estimated by maximum likelihood (corresponding to ordinary binary choice model estimation, e.g., logit or probit models), thereby yielding efficient estimates for moments and quantiles defined in terms of those parameters. We assume this distribution is not finitely parameterized.

Assumption A.1. The covariate vector X has support $\mathcal{X} \subseteq \mathsf{R}^d$. The latent scalar W has an unknown, twice continuously differentiable, strictly monotonic, conditional CDF $1 - G(w \mid x)$ with probability density function $g(w \mid x)$ and a compact support $[\rho_0(x), \rho_1(x)]$. The variables W and V are conditionally independent, given X. Let $Y = I(W > V)$. Let G^{-1} be the inverse of the function G with respect to its first element.

Assumption A.2. The function $r(w, x)$, chosen by the researcher, is regular, meaning that it is continuous in (w, x) for all w and x on their supports, and for each x is twice continuously differentiable in w. Define $r'(w, x) = \partial r(w, x)/\partial w$. Let $\kappa(x)$ be a function or constant in $[\rho_0(x), \rho_1(x)]$. The moment $\mu_r(x)$ exists, defined by $\mu_r(x) = E[r(W, X) \mid X = x]$.

It follows immediately from Assumption A.1, in particular the conditional independence of W and V, that

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

and if $G(v \mid x)$ can be estimated for all $v \in \text{supp}(W)$, then conditional moments $\mu_r(x)$ could be estimated using

$$\mu_r(x) = \int_{\text{supp}(W)} r(v, x) \frac{d[1 - G(v \mid x)]}{dv} dv.$$

The disadvantage of this expression is that it involves the derivative of a high dimensional function $G(v|x)$. We apply an integration by parts to this expression to obtain the basis for more direct estimators of $\mu_r(x)$.

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 Theorem 5 in the Appendix, the distribution of W is in general not identified when the support of V has a finite number of elements. To identify features of the distribution of W with minimal restrictions on G, our nonparametric estimators assume an experimental design in which the number of mass points may grow to infinity with the sample size, as follows.

Let $H_n(v, x \mid n)$ denote the realization of the observed sample of size n, which includes both nature's selection of X and the experimental design that selects V given X. 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 is the following.

Assumption A.3 Let $H_n(v, x \mid n)$ denote the empirical CDF of V, X, for sample size n. $\sup_v |H_n(v, x \mid n) - H(v, x)| \to 0$ *a.s.*, where $H(v, x)$ is a CDF having the property that the corresponding conditional distribution of V given $X = x$, denoted $H(v \mid x)$, has a strictly positive continuous probability density function $h(v \mid x)$ with compact support $[\delta_0(x), \delta_1(x)]$ such that $\delta_0(x) \leq \rho_0(x)$ and $\delta_1(x) \geq \rho_1(x)$.

Assumption A.3 is used to obtain nonparametric identification. For obtaining limiting distributions it will also be assumed that $n^\tau [H_n(v, x \mid n) - H(v, x)]$ 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. Suppose at sample size n, a fixed design with J_n possible values of V is selected. Suppose this design has the property that the maximum distance between a point in the support of W and a design point is of order $1/J_n$, and that $n^{\tau-1} J_n \to \infty$. Suppose X_i is drawn randomly from a distribution, and V_i is drawn randomly from a density $h(v \mid X_i, n)$ whose support is the fixed design. Suppose further that $h(v \mid X_i, n)$ is obtained by approximating a positive limiting density $h(v \mid x)$ on the finite support specified by the design. For example, one might define $h(v \mid X_i, n)$ so that its conditional CDF and the conditional CDF of $h(v \mid x)$ coincide at each design point. Then the weak convergence condition $n^{1/2} [H_n(v, x \mid n) - H(v, x)] \to 0$ is satisfied. This case covers [or would cover when the design sequence is spelled out satisfying the condition on J_n and the convergence properties of $h(v \mid X_i, n)$] all current studies, at least up to the quality of the asymptotic approximation of the design.

In our simulation studies, we will examine the size of finite sample bias that results when our estimators are applied both with discrete V and continuous V.

For estimation we suppose that a sample (X_i, V_i, Y_i) for $i = 1, \ldots, n$ is observed, generated in accordance with Assumption A.3, where V_i is a realization of V, Y_i is a realization of Y, and X_i is a realization of X. Using this data, we will provide five different estimators for $\mu_r(x)$, denoted $\widehat{\mu}_{jr}(x)$ for $j = 1, 2, 3, 4, 5$.

The estimator $\widehat{\mu}_{1r}(x)$ is for nonparametric estimation when the limiting experimental design

density $h(v \mid x)$ is known, and $\widehat{\mu}_{2r}(x)$ is for nonparametric estimation when $h(v \mid x)$ is unknown. Similarly, $\widehat{\mu}_{3r}(x)$ and $\widehat{\mu}_{4r}(x)$ cover the cases of semiparametric estimators where W is parameterized up to an unknown error term, with $h(v \mid x)$ known and unknown, respectively. An additional semiparametric estimator $\widehat{\mu}_{5r}(x)$ is provided that is simpler than $\widehat{\mu}_{3r}$ or $\widehat{\mu}_{4r}$, but may only be used for certain choices of r.

2.2 Nonparametric Moments

Theorem 1. *Let Assumptions A.1 and A.2 hold. Let $h(v \mid x)$ be a strictly positive conditional probability density function, and $H(v \mid x)$ be the associated CDF having compact support $[\delta_0(x), \delta_1(x)]$ such that $\delta_0(x) \leq \rho_0(x)$ and $\delta_1(x) \geq \rho_1(x)$. Define*

$$s_r(x, v, y) = r[\kappa(x), x] + \frac{r'(v, x)[y - 1(v < \kappa(x))]}{h(v \mid x)}$$

$$t_r(x, v) = \frac{r'(v, x)[G(v \mid x) - 1(v < \kappa(x))]}{h(v \mid x)}.$$

Then

$$\mu_r(x) = r[\kappa(x), x] + \int_{\rho_0(x)}^{\rho_1(x)} t_r(x, v) H(dv \mid x). \tag{2}$$

Also, if V is drawn from a conditional CDF $H(v \mid x, n)$ at sample size n, then

$$\mu_r(x) = E\left[s_r(X, V, Y) \mid X = x\right] + \int_{\rho_0(x)}^{\rho_1(x)} t_r(x, v)[H(dv \mid x) - H(dv \mid x, n)] \tag{3}$$

and, if Assumption A.3 also holds, as $n \to \infty$,

$$\mu_r(x) - E\left[s_r(X, V, Y) \mid X = x\right] \to 0. \tag{4}$$

Proof of Theorem 1. Starting from the definition of $\mu_r(x)$,

$$\begin{aligned} \mu_r(x) &= \int_{\rho_0(x)}^{\rho_1(x)} r(v, x) g(v \mid x) dv \\ &= \int_{\rho_0(x)}^{\kappa(x)} r(v, x) \frac{d[1 - G(v \mid x)]}{dv} dv + \int_{\kappa(x)}^{\rho_1(x)} r(v, x) \frac{-dG(v \mid x)}{dv} dv \end{aligned}$$

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

$$\begin{aligned} \mu_r(x) &= r[\kappa(x), x] + \int_{\rho_0(x)}^{\rho_1(x)} r'(v, x)[G(v \mid x) - 1(v < \kappa(x))] dv \\ &= r[\kappa(x), x] + \int_{\rho_0(x)}^{\rho_1(x)} \frac{r'(v, x)[G(v \mid x) - 1(v < \kappa(x))]}{h(v \mid x)} H(dv \mid x), \end{aligned}$$

which is equation (2). Adding and subtracting $\int_{\rho_0(x)}^{\rho_1(x)} t_r(x,v)H(dv \mid x,n)$ gives

$$\mu_r(x) = r[\kappa(x),x] + \int_{\rho_0(x)}^{\rho_1(x)} \frac{r'(v,x)[G(v \mid x) - 1(v < \kappa(x))]}{h(v \mid x)} H(dv \mid x,n) + \int_{\rho_0(x)}^{\rho_1(x)} t_r(x,v)[H(dv \mid x) - H(dv \mid x,n)],$$

which yields equation (3) after applying the law of iterated expectations. Equation (4) then follows from the convergence in Assumption A.3 and the bounded continuity of t_r. ■

We can use equation (3) to compute an estimator of $\mu_r(x)$ by the analogy principle substituting in estimators of the unknown quantities. Let $\widehat{\mu}_{1r}(x)$ denote this estimator, details supplied below. The estimator $\widehat{\mu}_{1r}(x)$ is numerically simple (and in particular does not require kernel or other smoothers if X is discrete), but requires the researcher to know, or be able to estimate, the limiting design density $h(v \mid X)$.[6] An estimator that does not entail knowing or estimating the limiting density h can be constructed as follows. First observe that equation (2) in Theorem 1 does not require Assumption A.3, so the CDF $H(v \mid x)$ and associated density $h(v \mid x)$ need not describe the limiting data generating process for V, but may simply be chosen for convenience or efficiency. In particular, letting $H(v \mid x)$ be a uniform distribution reduces equation (2) to

$$\mu_r(x) = r[\kappa(x),x] + \int_{\rho_0(x)}^{\rho_1(x)} r'(v,x)[G(v \mid x) - 1(v \geq \kappa(x))]dv. \tag{5}$$

Let a_0 and a_1 be known or estimated constants such that $a_0 \leq \rho_0(x)$ and $a_1 \geq \rho_1(x)$. Then, by equations (5) and (1), a consistent estimator of $\mu_r(x)$ is given by

$$\widehat{\mu}_{2r}(x) = r[\kappa(x),x] + \int_{a_0}^{a_1} r'(v,x)[\widehat{E}[Y \mid V = v, X = x] - 1(v < \kappa)]dv, \tag{6}$$

where $\widehat{E}[Y \mid V = v, X = x]$ is an estimate of $E[Y \mid V = v, X = x]$. One could construct additional analogous estimators based on (2) instead of (5), using other choices of H, but for simplicity, we apply Theorem 1 only in the form of equations (3) and (5).

Consistency and potential effects of finite sample design on limit distributions for $\widehat{\mu}_{2r}(x)$ are analogous to the above discussion of $\widehat{\mu}_{1r}(x)$. In applications, the choice between using $\widehat{\mu}_{1r}(x)$ or $\widehat{\mu}_{2r}(x)$ would be based at least in part on the information set of the researcher regarding the limiting design density. We provide more details later on the construction and limiting distributions of these estimators.

[6] If h is unknown, then based on $\widehat{\mu}_{1r}$ an estimator of $\mu_r(x)$ could be constructed by first estimating h. Specifically, one could replace $h(v \mid x)$ with an estimate $\widehat{h}(v \mid x)$ (using, e.g., kernel density estimation) in the definition of $s_r(x,v,y)$. Call the result $\widehat{s}_r(x,v,y)$. The estimator of $\mu_r(x)$ would then be $\widehat{\mu}^*_{1r}(x) = \widehat{E}[\widehat{s}_r(X,V,Y) \mid X = x]$

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 $r(w, x) = \Lambda^{-1}(w)$.[7]

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.

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. $\sup_v |\Psi_n(U \mid n) - \Psi(U)| \to 0$ *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.

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

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

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

$$\mu_r(x) - E\left[s_r^*(x, U, Y)\right] = \int_{a_0}^{a_1} t_r^*(x, u)[\Psi(du) - \Psi(du \mid n)] \to 0$$

and, if Assumption A.3 also holds,

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

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

[7] In this special case of a location model, many other functions r provide additional information about m. For example, taking $r(w, x) = [\Lambda^{-1}(w)]^2$, makes $\mu_r(x) = m^2(x) - 2m(x)E(\varepsilon) + \sigma_\varepsilon^2$ for some constant σ_ε^2.

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

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

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 $\Psi(u) = \lim_{n\to\infty} E(1 - H[\Lambda(m(X,\theta_0) - u) \mid X])$, 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 $\widehat{U}_i$ 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 with Unknown 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 $s_w(X,V,Y)$ denote $s_r(X,V,Y)$ with $r(w,x) = \Lambda^{-1}(w)$. It then follows from Theorem 1 that

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

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_w 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

$$(\widehat{\theta}, \widehat{\alpha}) = \arg\min_{\theta,\alpha} \frac{1}{n}\sum_{i=1}^{n} [s_w(X_i,V_i,Y_i) - \alpha - m(X_i,\theta)]^2. \tag{7}$$

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 $s_w(X,V,Y)$ 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.[8]

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) would not be directly applicable due to this nondifferentiability.

2.3.3 A Special Case

In this section we suppose that

$$r(w,x) = [\Lambda^{-1}(w)]^k.$$

This, when Λ is the identity function, would be the typical choice of function r in applications. Let $s_{w^k}(X,V,Y)$ denote $s_r(X,V,Y)$ with $r(w,x) = [\Lambda^{-1}(w)]^k$. For any k we then have

$$E[(\Lambda^{-1}(W))^k \mid X = x] = E\left[(m(X,\theta_0) - \varepsilon)^k \mid X = x\right] = \sum_{\ell=0}^{k} m(x,\theta_0)^\ell (-1)^{k-\ell} \binom{k}{\ell} E(\varepsilon^{k-\ell})$$

by the binomial expansion. Therefore,

$$E[[\Lambda^{-1}(W)]^k \mid X = x] = \sum_{\ell=0}^{k} m(x,\theta_0)^\ell \alpha_{k\ell},$$

where $\alpha_{k\ell}$, $\ell = 0,\ldots,k$ are unknown parameters depending on the moments of the error distribution and on the binomial coefficients. It also follows from Theorem 1 that

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

We may estimate the nuisance parameters $\alpha_{k\ell}$ by solving the least squares problem

$$(\widehat{\alpha}_{k0},\ldots,\widehat{\alpha}_{kk}) = \arg\min_{\alpha_{k0},\ldots,\alpha_{kk}} \frac{1}{n}\sum_{i=1}^{n}\left(s_{w^k}(X_i,V_i,Y_i) - \sum_{\ell=0}^{k} m(X_i,\widehat{\theta})^\ell \alpha_{k\ell}\right)^2,$$

[8] 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.

where $\widehat{\theta}$ is any root-n consistent estimator such as defined in (7). Then let

$$\widehat{\mu}_{5w^k}(x) = \sum_{\ell=0}^{k} m(x,\widehat{\theta})^{\ell}\widehat{\alpha}_{k\ell}.$$

to estimate $\mu_{w^k}(x)$.[9] For identification we require that the matrix $(M_{jl})_{j,l=0}^{k}$ be of full rank, where

$$M_{jl} = E\left[m(X_i,\theta_0)^{j+l}\right].$$

This estimator should work well when k is small, but otherwise a large number of auxiliary parameters $\alpha_{k\ell}$ have to be estimated and this may result in the estimate of $\mu_r(x)$ having a large variance. It is also sensitive to the existence of moments.

2.4 Quantiles

Let $w_q(x)$ denote the q'th conditional quantile of W given $X = x$. It follows immediately from Assumption A.1, in particular equation (1), that

$$w_q(x) = G^{-1}(1-q \mid x), \tag{8}$$

where $G(v \mid x) = E(Y \mid V = v, X = x)$, so we may invert a nonparametric estimator of this expectation to obtain an estimate of $w_q(x)$, for any q such that $1-q \in \text{supp}(V)$, and so will be identified for all quantiles given Assumption A.3. The rate of convergence of $\widehat{w}_q(x) = \widehat{G}^{-1}(1-q \mid x)$ will be slow, because of the high dimension of $\widehat{G}$.

For semiparametric quantile estimation, if Assumptions A.1 and A.4 hold then

$$\begin{aligned} q &= \Pr[\Lambda[m(X,\theta_0)-\varepsilon] \leq w_q(X) \mid X = x)] \\ &= 1 - F_\varepsilon[m(X,\theta_0) - \Lambda^{-1}(w_q(x))] \end{aligned}$$

so

$$w_q(x) = \Lambda[m(X,\theta_0) - F_\varepsilon^{-1}(1-q)]$$

and from Corollary 1, F_ε is obtained by $F_\varepsilon(u) = E(Y \mid U = u)$. Therefore, let $\widehat{U}_i = m(X_i,\widehat{\theta}) - \Lambda^{-1}(V_i)$ and estimate the conditional quantile $w_q(x)$ by

$$\begin{aligned} \widehat{F_\varepsilon}(u) &= \widehat{E}(Y \mid \widehat{U} = u) \\ \widetilde{w}_q(x) &= \Lambda[m(x,\widehat{\theta}) - \widehat{F_\varepsilon}^{-1}(1-q)], \end{aligned}$$

[9] This method could also be extended to more general classs of r functions. Suppose $r(w,x) = \sum_{j=1}^{\infty}\psi_j(x)w^j$ for some known coefficients $\{\psi_j(x)\}_{j=1}^{\infty}$. This is true for a large class of r functions of interest like the exponential and logarithm. Then in practice, we approximate $r(w,x)$ by $\sum_{j=1}^{\tau}\psi_j(x)w^j$, where $\tau = \tau(n)$ is some truncation parameter, and then apply our method for estimating $\mu_{w^k}(x)$.

where the function $\widehat{F_\varepsilon}$ is obtained by nonparametrically regressing Y on $\widehat{U}$, and is then numerically inverted to get $\widehat{F_\varepsilon}^{-1}$. This estimator $\widetilde{w}_q(x)$ will converge at a faster rate than the nonparametric quantile estimator $\widehat{w}_q(x)$, because estimation of the quantiles $w_q(x)$ given θ only requires estimation of the one dimensional regression $F_\varepsilon(u) = E(Y \mid U = u)$, instead of the high dimensional $G(v \mid x)$.

3 Estimation Details and Distribution Theory

In this section we provide more detail about the computation of the estimators $\widehat{\mu}_{1r}(x), \ldots, \widehat{\mu}_{5r}(x)$ and their distribution theory.

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.[10] 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 (2002), 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

[10] If there is a continuous density but with some points in the support of zero density, the rate of convergence may be slower but Hengartner and Linton (1996) have shown that the local linear estimator can still achieve the optimal rate in this case. There are other non-standard cases: Lu (2002) considers the case where the covariate process has fractal dimension [e.g., in the multivariate case where the covariates lie on a nonlinear manifold of lower local dimension].

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 superscript V by n, so V^n is the population random variable for each n and V_i^n is a draw from it, and use V to denote a 'limiting' version of V^n. We shall suppose that V^n is asymptotically continuous in the sense that for each n, V_i^n is drawn from a distribution $H(v|X_i, n)$ that has finite support, increasing with n.[11] In this case, the pure frequency estimator of $h(v|x)$ is asymptotically normal at rate $\sqrt{n/J}$.

As already discussed in Theorem 1 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 $s_r(Z_i^n)$ against X_i, where $Z_i^n = (V_i^n, X_i, Y_i)$. Let $(\widehat{\vartheta}_0, \widehat{\vartheta}_x)$ minimize the following localized least squares criterion

$$\sum_{i=1}^{n} K_b\left(x - X_i\right)\left[s_r(Z_i^n) - \vartheta_0 - \vartheta_x'(X_i - x)\right]^2,$$

where $K_b\left(t\right) = \prod_{j=1}^{d} k_b(t_j)$ with $k_b(u) = k(u/b)/b$, where k is a univariate kernel function and $b = b(n)$ is a bandwidth. Then let

$$\widehat{\mu}_{1r}(x) = \widehat{\vartheta}_0. \tag{9}$$

This estimator is linear in the dependent variable and has an explicit form. When there are some discrete components to X, it maybe advantageous to modify the kernel window to reflect this along the lines discussed in Li and Racine (2002) for example.

In computing the estimator $\widehat{\mu}_{2r}(x)$ we require an estimator of $G(v \mid x)$, which is given by the local linear smooth of Y_i on X_i, V_i^n. Specifically, for any $v \in \mathcal{V}_n$ let $(\widehat{\vartheta}_0, \widehat{\vartheta}_v, \widehat{\vartheta}_x)$ minimize the following localized least squares criterion

$$\sum_{i=1}^{n} k_b\left(v - V_i^n\right) K_b\left(x - X_i\right)\left[Y_i - \vartheta_0 - \vartheta_v(V_i^n - v) - \vartheta_x^\top(X_i - x)\right]^2,$$

[11] 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.

and let $\widehat{G}(v \mid x) = \widehat{\vartheta}_0$. Then define

$$\widehat{\mu}_{2r}(x) = r(\kappa(x), x) + \int_{\rho_0(x)}^{\rho_1(x)} r'(v, x)[\widehat{G}(v \mid x) - 1(v < \kappa(x))]dv, \tag{10}$$

where the univariate integral is interpreted in the Stieltjes sense. Specifically, if V has a discrete distribution as described above,

$$\begin{aligned}\int_{\rho_0(x)}^{\rho_1(x)} r'(v, x)[\widehat{G}(v \quad \mid \quad x) - 1(v < \kappa(x))]dv \\ = \quad \frac{1}{J}\sum_{j=1}^{J} r'(v_{nj}, x)[\widehat{G}(v_{nj} \mid x) - 1(v_{nj} < \kappa(x))]1(\rho_0(x) \le v_{nj} \le \rho_1(x)).\end{aligned}$$

When V is continuously distributed we would compute (10) by an approximation scheme that replaces the integral by a finite sum.

The estimator (10) is in the class of marginal integration/partial mean estimators sometimes used for estimating additive nonparametric regression models, see Linton and Nielsen (1995), Newey (1994), and Tjø stheim and Auestad (1994), except that the integrand is not just a regression function and the integrating measure λ, where (asymptotically) $d\lambda(v) = r'(v, x)1(\rho_0(x) \le v \le \rho_1(x))dv$, 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, see the above references.

We make the following assumptions.

Assumption B.1. k is a symmetric probability density with bounded support, and is Lipschitz continuous on its support, i.e., $|k(t) - k(s)| \le c|t - s|$ for some finite constant c. Define $\mu_2(k) = \int t^2 k(t)dt$.

Assumption B.2. The random variables (V^n, X) are asymptotically continuously distributed, i.e., there exists a Lebesgue density $h_{V,X}(v, x)$ along with conditionals $h(v|x)$ and marginal $h_X(x)$ such that for some finite constant c_h

$$\sup_{v,x} |H(v|x, n) - H(v|x)| \le \frac{c_h}{J}, \tag{11}$$

where $h(v|x)$ is the density of $H(v|x)$. Furthermore, $\inf_{\rho_0(x) \le v \le \rho_1(x)} h_{V,X}(v, x) > 0$. For all n larger than some n_0, $\mathrm{var}(Y_i \mid V_i^n = v, X_i = x) < \infty$, and the limiting conditional variance $\sigma^2(v, x) = \mathrm{var}(Y_i \mid V_i = v, X_i = x) = G(v \mid x)[1 - G(v \mid x)]$. Furthermore, $G(v \mid x)$ and $h_{V,X}(v, x)$ are twice continuously differentiable for all v with $\rho_0(x) \le v \le \rho_1(x)$. The set $[\rho_0(x), \rho_1(x)] \times \{x\}$ is strictly contained in the support of (V, X) for large enough n.

These are standard regularity conditions for nonparametric estimation. Let ∇, ∇^2 denote the first and second derivative operators. Define

$$\beta_1(x) = \frac{\mu_2(k)}{2}\text{tr}(\nabla^2 \mu_r(x)) \; ; \; \beta_2(x) = \frac{\mu_2(k)}{2}\int \text{tr}(\nabla^2 G(v \mid x))d\lambda(v),$$

$$\omega_1(x) = \|K\|^2 \frac{\text{var}[s_r(Z) \mid X = x]}{h_X(x)} \; ; \; \omega_2(x) = \|K\|^2 \int_{\rho_0(x)}^{\rho_1(x)} \sigma^2(v,x) \left(\frac{r'(v,x)}{h_{V,X}(v,x)}\right)^2 h_{V,X}(v,x)dv.$$

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* $j = 1, 2$,

$$\frac{\widehat{\mu}_{jr}(x) - \mu_r(x) - b^2\beta_j(x)}{\sqrt{\omega_j(x)/nb^d}} \Longrightarrow N(0,1).$$

Consistent standard errors can be obtained by plugging in consistent estimators of the unknown quantities in $\omega_j(x)$.

3.1.1 Efficiency Comparison

When the limiting design distribution $h(V \mid X)$ is known, either $\widehat{\mu}_{1r}$ or $\widehat{\mu}_{2r}$ may be applied. These two estimators are not in general rankable in terms of mean squared error, but can be compared in some special cases.

Suppose that $r(w,x) = w$. In this case

$$\omega_1(x) \propto \text{var}\left(\frac{[Y - 1(V < \kappa(x))]}{h(V \mid X)} \mid X = x\right) \text{ and } \omega_2(x) \propto \int_{\rho_0(x)}^{\rho_1(x)} \left(\frac{G(v \mid x)[1 - G(v \mid x)]}{h(v \mid x)}\right) dv.$$

If furthermore, $V|X$ is uniform on $[0,1]$,

$$\omega_1(x) \propto \int_0^1 G(v \mid x)dv[1 - \int_0^1 G(v \mid x)dv] + \kappa(x)(1-\kappa(x)) + 2\left[\int_0^{\kappa(x)} G(v \mid x)dv - \kappa(x)\int_0^1 G(v \mid x)dv\right]$$

$$\omega_2(x) \propto \int_0^1 G(v \mid x)[1 - G(v \mid x)]dv.$$

Generally, $\omega_1(x)$ depends on $\kappa(x)$ except in the special case that W is uniform on $[0,1]$. If $W|X$ is uniform on $[0,1]$, the asymptotic variance of $\widehat{\mu}_{2r}(x)$ is proportional to $1/6$, while the asymptotic variance of $\widehat{\mu}_{1r}(x)$ is proportional to $1/4$, with proportionality factor having to do with the bandwidth, kernel, and covariate density. In this case, $\widehat{\mu}_{2r}(x)$ is more efficient in variance terms.[12]

[12]The estimators $\widehat{\mu}_{1r}(x)$ for different κ are correlated, but not perfectly so, so that there is scope for improving

Regarding the bias of the two estimators in the case that $r(w,x) = w$:

$$\begin{aligned} \mathrm{tr}(\nabla_x^2 \mu_r(x)) &= \sum_{j=1}^{d} \int v \frac{\partial^2 g(v \mid x)}{\partial x_j^2} dv = \sum_{j=1}^{d} \int \mathrm{tr}(\nabla_x^2 G(v \mid x)) dv \\ \int \mathrm{tr}(\nabla^2 G(v \quad &\mid \quad x)) d\lambda(v) = \int \left[\sum_{j=1}^{d} \frac{\partial^2 G(v \mid x)}{\partial x_j^2} + \frac{\partial^2 G(v \mid x)}{\partial v^2} \right] dv, \end{aligned}$$

where $g(v \mid x)$ is the conditional density of $W|X$. Under certain conditions these two biases are the same applying integration by parts. In order for $\int [\partial^2 G(v \mid x)/\partial v^2] dv = -\int [\partial g(v \mid x)/\partial v] dv = 0$ it is sufficient that the conditional density and its derivative be zero on the boundary.

We have shown than in an important special case $\widehat{\mu}_{2r}(x)$ has smaller mean squared error than $\widehat{\mu}_{1r}(x)$. However, either estimator could be more efficient in other situations. There are other comparisons between the estimators that are also relevant. For example, the estimator $\widehat{\mu}_{1r}(x)$ requires prior knowledge of $h(v \mid x)$, and entails more smoothness than $\widehat{\mu}_{2r}(x)$, as can be seen from the bias expressions given above. 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.[13] 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, whereas a comparable theory relevant for $\widehat{\mu}_{2r}(x)$ is not so well developed. Similar comments apply to standard error construction based on standard asymptotic or bootstrap principles.

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 $U_i = m(X_i, \theta_0) - \Lambda^{-1}(V_i)$

efficiency of $\widehat{\mu}_{1r}(x)$. The covariance function of $\widehat{\mu}_{1r}(x;\kappa_1), \widehat{\mu}_{1r}(x;\kappa_2)$ is proportional to

$$C(\kappa_1, \kappa_2) = \frac{1}{4} + \kappa_1 \wedge \kappa_2 - \kappa_1 \kappa_2 - \left[\frac{\kappa_1}{2} - \int_0^{\kappa_1} G(v) dv \right] - \left[\frac{\kappa_2}{2} - \int_0^{\kappa_2} G(v) dv \right],$$

which is equal to

$$\frac{1}{4} + \kappa_1 \wedge \kappa_2 - \kappa_1 \kappa_2 - \frac{\kappa_1}{2} + \frac{\kappa_1^2}{2} - \frac{\kappa_2}{2} + \frac{\kappa_2^2}{2}$$

in the uniform case. The correlation function is maximized at $\kappa_1 = \kappa_2$ whereupon it is one, but is minimized at the point where $\kappa_1 = \kappa_2 \pm 1/2$ and the minimizing value is 0.5. Furthermore, we can establish a functional central limit theorem in κ. Now consider the class of estimators $\int \widehat{\mu}_{1r}(x;\kappa)\omega(\kappa)d\kappa$ for some weighting function $\omega(.)$. One can show that with the optimal combination we achieve the same variance factor, 1/6 in the uniform case, as the estimator $\widehat{\mu}_{2r}(x)$.

[13] The evidence on the finite sample performance of marginal integration estimators is mixed, see Sperlich, Linton, and Härdle (1999).

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 (11).

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

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

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 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 $nb^6 \to \infty$.

These regularity conditions are fairly standard.

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

$$\Gamma_F = \left(\frac{\partial}{\partial\theta}E\left[f_1(Z_i,\theta)\right]\right)\Big|_{\theta=\theta_0} = \left(\frac{\partial}{\partial\theta}E\left[f_0(Z_i,\theta)\right]\right)\Big|_{\theta=\theta_0} + r'[\Lambda(m(x,\theta_0)),x]\Lambda'(m(x,\theta_0))\frac{\partial m(x,\theta_0)}{\partial\theta}$$

$$\Psi_F = E\left[f_0(Z_i,\theta_0)\frac{\psi'(U_i)}{\psi(U_i)}\widetilde{\gamma}_i\right] + E\left[\frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\widetilde{\zeta}_{ij}\widetilde{\gamma}_j\right]$$

$$\widetilde{\gamma}_i = \frac{\partial m}{\partial\theta^\top}(X_i,\theta_0) - E\left[\frac{\partial m}{\partial\theta^\top}(X_i,\theta_0)\right]$$

and $\zeta_{ij} = [h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda'']\,(m(X_j,\theta_0)-U_i)$, 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 $n\to\infty$,*

$$\frac{\sqrt{n}[\widehat{\mu}_{3r}(x) - \mu_r(x)]}{\sigma_\eta(x)} \Longrightarrow N(0,1), \tag{12}$$

where $0 < \sigma_\eta^2(x) = \mathrm{var}(\eta_j) < \infty$ 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.

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

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

$$\gamma_i^* = \frac{\partial m}{\partial\theta^\top}(X_i,\theta_0) - \overline{m}_\theta(U_i),$$

where $\overline{m}_\theta(U_i) = E[\partial m(X_i,\theta_0)/\partial\theta \mid U_i]$ and $\overline{f}_0(U_i) = E[f_0(Z_i,\theta_0)|U_i]$.

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

$$\frac{\sqrt{n}[\widehat{\mu}_{4r}(x) - \mu_r(x)]}{\sigma^*_\eta(x)} \Longrightarrow N(0,1),$$

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

$$\eta^*_{2j} = (\Gamma_F - \Psi^*_F)\,\varsigma(Z_j;\theta_0)$$

$$\eta^*_{3j} = -[\overline{f}_0(U_j) - E\overline{f}_0(U_j)].$$

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.

Standard errors can be constructed by substituting population quantities by estimated ones along the lines discussed in Newey and McFadden (1994). Alternatively, one can use the bootstrap as we do in our application below.

Regarding efficiency, it is not possible to provide a 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)

3.2.1 Quantile estimators

The distribution theory for our quantile estimators is immediate. The estimator $\widehat{w}_q(x) = \widehat{G}^{-1}(1-q \mid x)$ has the standard distribution theory for conditional quantile estimation. See, e.g., Chaudhuri (1991). The distribution theory for $\widetilde{w}_q(x) = m(x,\widehat{\theta}) - \widehat{F_\varepsilon}^{-1}(1-q)$ is the same as the distribution theory for $\widetilde{w}_q(x) = m(x,\theta_0) - \widetilde{F_\varepsilon}^{-1}(1-q)$, where

$$\widetilde{F_\varepsilon}(u) = \widehat{E}(Y \mid U = u),$$

which is again a standard one-dimensional conditional quantile estimator. This is because $\widehat{\theta}$ converges at rate root-n, so the estimation error in $\widehat{\theta}$ is asymptotically irrelevant given the slower convergence rate of quantiles.

4 Numerical Results

4.1 Monte Carlo

We report the results of a small simulation experiment based on a design of Crooker and Herriges (2000). 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 $\sigma \in \{5, 10, 50\}$ and sample size $n \in \{100, 300, 500\}$. For our first set of experiments the bid values are chosen equally randomly from $\{25, 50, 75, 125, 175\}$, and $\kappa = 100$. This design was chosen because it permits direct comparison with the parametric and SNP estimators of WTP considered by Crooker and Herriges (2000).

The moments we estimate are $\text{std}(W \mid X = x) = \sqrt{E[W^2 \mid X = x] - E^2[W \mid X = x]}$ and $E[W \mid X = x]$. We compute estimators $\widehat{\mu}_\lambda(.)$ for $\lambda = 1, 3, 4, 5$. We do not compute $\widehat{\mu}_2(.)$ here, because it is very time consuming, and the small sample performance of this class of estimators (integrals of nonparametric conditional expectations) has been extensively documented in Sperlich, Linton, and Härdle (1999) and elsewhere. In the computation of $\widehat{\mu}_1(.)$ and $\widehat{\mu}_4(.)$ we used a Gaussian kernel and Silverman's rule of thumb bandwidth. This kernel and bandwidth is not likely to be optimal for this problem, but they are convenient and hence fairly widely used choices in practice.

In Table 1 and 2 we report four different performance measures: pointwise root mean squared error (PRMSE), pointwise mean absolute error (PMAE), integrated root mean squared error (IRMSE), and integrated mean absolute error (IMAE). Crooker and Herriges (2000) 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 generally improves with sample size according to all measures: the pointwise measures improve at approximately our theoretical asymptotic rate, while the integrated measures improve much more slowly. Note that, unlike our limiting distribution theory, in this set of experiments the number of mass points of the bid distribution does not increase with the sample size.

(A2) The nonparametric estimators $\widehat{\mu}_1(.)$ and $\widehat{\mu}_4(.)$ perform very well, in some cases better than the semiparametric estimators $\widehat{\mu}_3(.)$ and $\widehat{\mu}_5(.)$. From Table 1, the rankings of the estimators from best to worst by pointwise performance criteria are $\widehat{\mu}_1$, $\widehat{\mu}_5$, $\widehat{\mu}_4$, $\widehat{\mu}_3$ in small samples and $\widehat{\mu}_1$, $\widehat{\mu}_4$, $\widehat{\mu}_5$, $\widehat{\mu}_3$ in large samples. By integrated performance criteria, the rankings from best to worst are $\widehat{\mu}_5$, $\widehat{\mu}_4$,

$\widehat{\mu}_1$, $\widehat{\mu}_3$ in small samples and $\widehat{\mu}_4$, $\widehat{\mu}_5$, $\widehat{\mu}_1$, $\widehat{\mu}_3$ in large samples.

(A3) The only consistent ranking across designs is that $\widehat{\mu}_3$ always performs the worst.

(A4) Estimators $\widehat{\mu}_1(.)$, $\widehat{\mu}_4(.)$, and $\widehat{\mu}_5(.)$ seem to perform better than the Crooker and Herriges SNP estimator, especially in the large σ case.

(A5) The estimates of 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. We repeated the above experiments with bid distribution uniform on $[25, 175]$ and report the results in Tables 3 and 4. Our conclusions are:

(B1) The performance in the continuous design is somewhat better than in the discrete design, e.g., 70% of the numbers are larger in Table 1 than in Table 3. For some designs the pointwise results in Table 1 are better, but the integrated results are always better in Table 3. Note that for the pointwise results the chosen point of evaluation $x = 0$ corresponds to $E[W \mid X = 0] = 100$ and in Table 1 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 best in Table 4.

(B3) The ranking of the estimators is the same in Table 3 as Table 1. Once again $\widehat{\mu}_3$ always performs the worst, but the rankings of the other estimators vary depending on the criterion and sample size.

4.2 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 is 14. The dataset consists of bid responses and some personal characteristics of the respondents. The covariates 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 $n = 530$.

Because there are seven covariates and only a limited number of bid values, we first consider semiparametric specifications for W, in particular:

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

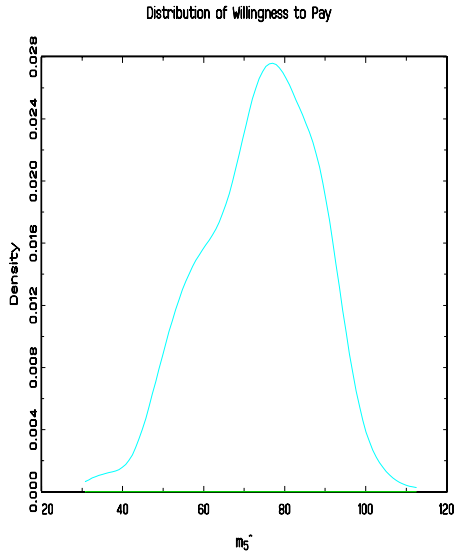

Figure 1:

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 $\widehat{\mu}_j(x)$, $j = 3, 4, 5$. To check for possible framing effects, we estimate this conditional mean WTP separately using first bid data and second bid data.

Figure 1 shows a kernel based estimate of the density of the data points $\widehat{\mu}_5(X_i)$, $i = 1, ..., n$ using the linear specification with first bid data. This may be interpreted as the distribution of estimated WTP across the sample. Similar results are obtained using $\widehat{\mu}_3$ and $\widehat{\mu}_4$ and (in shape but not location) using second bid data. There is obviously quite a bit of dispersion around the mean, but the distribution looks quite symmetric.

In Table 5 we report the sample average of the estimates of $E(W \mid X = X_i)$, denoted $\overline{\widehat{\mu}_j}$, as well as $E(W \mid X = \overline{X})$, denoted $\widehat{\mu}_j(\overline{X})$, along with bootstrap standard errors. The bootstrap data were drawn with replacement from $\{Y_i, V_i, X_i\}_{i=1}^n$. The computation of $\widehat{\mu}_j$ is done exactly as described in the simulation section.

	Linear		Log Linear	
	bid1	bid2	bid1	bid2
$\widehat{\overline{\mu_3}}$	68.3417 (95.6285)	274.0677 (172.0449)	62.0320 (4.4683)	306.0211 (411.4603)
$\widehat{\mu}_3(\overline{X})$	68.3417 (95.6285)	274.0677 (172.0449)	61.5918 (4.2751)	302.4752 (328.7766)
$\widehat{\overline{\mu_4}}$	61.3042 (90.5139)	213.0857 (120.3624)	64.6992 (5.0823)	369.2809 (394.6291)
$\widehat{\mu}_4(\overline{X})$	61.3042 (90.5139)	213.0857 (120.3624)	63.7869 (4.4995)	472.5140 (328.2098)
$\widehat{\overline{\mu_5}}$	73.7925 (8.4186)	143.4519 (13.6322)	99.1164 (4.1348)	141.5369 (9.0742)
$\widehat{\mu}_5(\overline{X})$	73.7925 (8.4186)	143.4519 (13.6322)	98.7726 (6.6526)	134.0196 (21.4996)

Table 5: Estimates of WTP

Table 6 provides parameter estimates along with their bootstrap standard errors, and asterisks indicating significant departure from zero at the 5% level.

	Linear		Log Linear	
	bid1	bid2	bid1	bid2
YEARCA	0.6869 (0.4724)	1.6964 (0.8023)	0.0021 (0.0022)	0.0131 (0.0062*)
SEX	10.3763 (12.3460)	33.3371 (22.4836)	−0.0460 (0.0632)	0.2579 (0.1740)
ln(AGE)	−42.0977 (19.2617*)	−62.0518 (33.1249)	−0.2040 (0.1088)	−0.4801 (0.2563)
EDUC	−0.9696 (2.9860)	3.9614 (5.2206)	0.0119 (0.0154)	0.0307 (0.0404)
WHITE	5.0222 (14.3550)	27.9634 (28.0857)	0.1338 (0.0797)	0.2164 (0.2173)
ENVORG	3.4128 (14.6966)	12.2217 (30.1263)	−0.1085 (0.0792)	0.0946 (0.2331)
ln(INCOME)	7.0079 (9.3697)	49.0606 (19.0478*)	0.0972 (0.0500*)	0.3796 (0.1474*)

Table 6

The most striking feature of Tables 5 and 6 is that the second bid data yield far larger coefficients, with corresponding larger estimates of WTP. 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.

Looking across estimators and specifications, few of the regressors are statistically significant. Income is generally most significant, having a positive effect. Table 5 shows a moderate range of mean WTP estimates from the first bid, while the second bid WTP estimates are far more dispersed and have much larger standard errors. 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 62 to 99 for first bid data and 141 to 369 using only second bids.

Finally, we conducted a purely nonparametric analysis with each of the four continuous covariates, one at a time. In Table 7 we report the estimated value of $\mu_w(\overline{X}_j)$ along with bootstrap standard errors for each separate covariate. The implementation is as described in the simulation section. The estimated values of $\mu_w(\overline{X}_j)$ are quite precise and are in the ballpark reported earlier.

	$\widehat{\mu}_1(\overline{X}_j)$	
X_j	bid1	bid2
YEARCA	75.5147 (10.75004)	167.9237 (17.7617)
AGE	91.5519 (9.6176)	171.3500 (22.1864)
EDUC	94.6993 (11.4778)	142.3537 (34.2841)
ln(INCOME)	75.5148 (10.7500)	167.9237 (17.7617)

Table 7

In Figures 2 and 3 we provide the marginal smooths themselves along with a pointwise 95% confidence interval. In contrast to the semiparametric model which assumes a linear or loglinear relationship, these figures from the nonparametric estimator show some nonlinear effects.

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

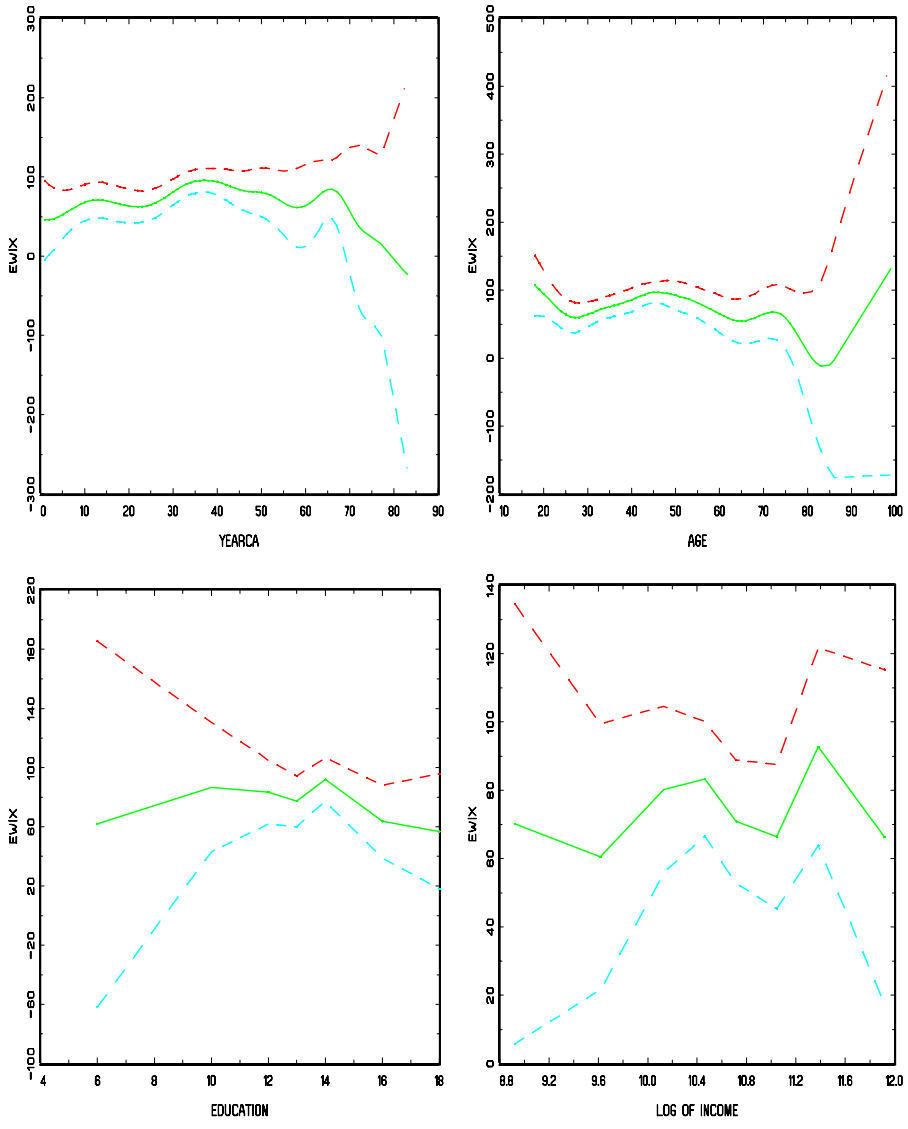

Figure 2: Marginal Local Linear regression Smooths using Bid1 data along with pointwise 95% confidence interval

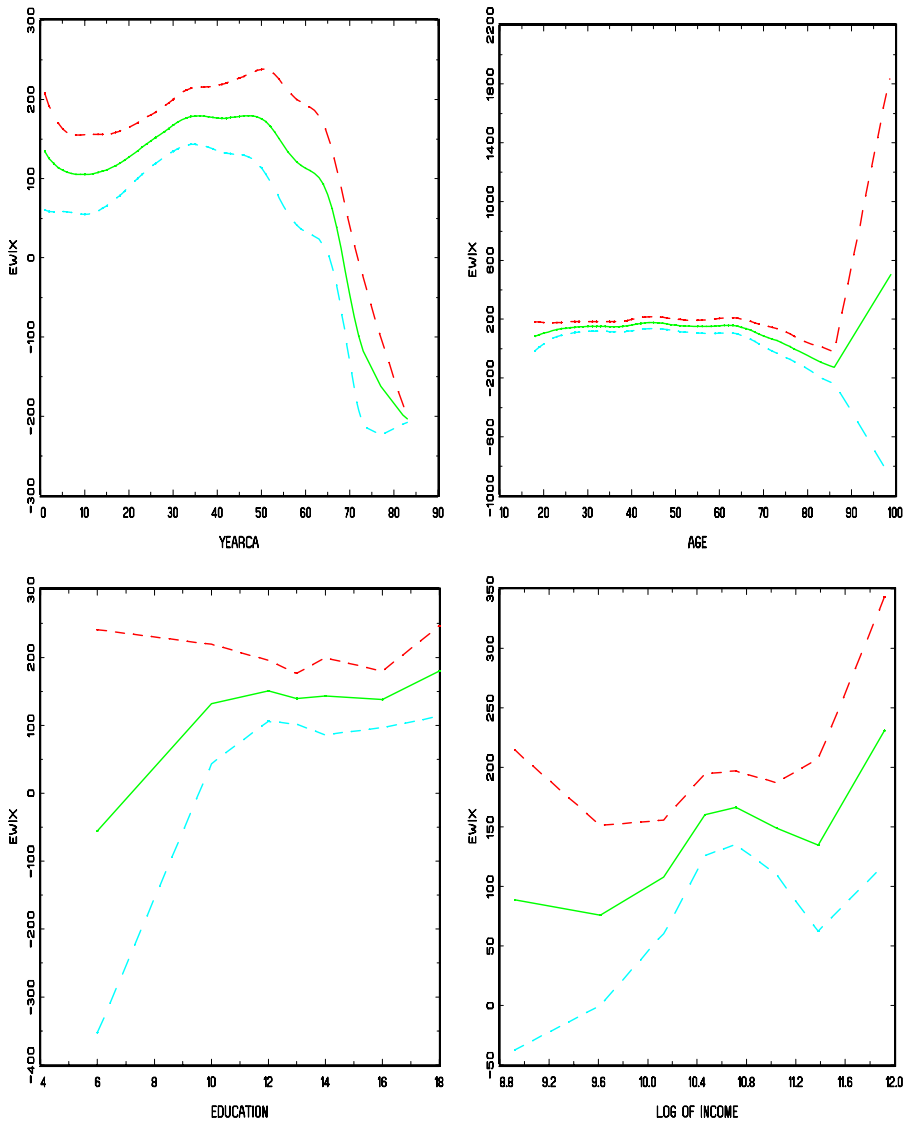

Figure 3: Marginal Local Linear regression Smooths using Bid2 data along with pointwise 95% confidence interval

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

6.1 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 5. Assume $\text{supp}(X)$ is some open or closed interval on the real line, $\text{supp}(V) = \{-\delta, 0\}$ 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 5. 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 Theorem 5 for bid design differ markedly from results on optimal bid design in

parametric or semiparametric models. Summarizing Kanninen (1993), Crooker and Herriges (2000) say, in refering 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. Theorem 5 shows that such models are generally not identified.[14]

6.2 Distribution Theory for Nonparametric Estimators

Proof of Theorem 2. The properties of $\widehat{\mu}_{1r}(x)$ are more or less standard, because V_i^n is part of the variable being smoothed. The only difference is the triangular array nature of the sampling scheme, but given the conditions we made on the way this distribution changes with n, the limiting distribution $H(v|x)$ can replace $H(v|x,n)$ with error of smaller order than the leading term.

We now turn to $\widehat{\mu}_{2r}(x)$. First, we introduce some notation to define the local linear estimator $\widehat{G}(v \mid x)$. Define $\widetilde{X}_i = (V_i^n, X_i)$ 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. Then

$$\widehat{G}(\widetilde{x}) - G(\widetilde{x}) = e_1^\top M_n^{-1}(\widetilde{x}) \left[\Psi_{n1}(\widetilde{x}) + \Psi_{n2}(\widetilde{x})\right], \tag{13}$$

where

$$\begin{aligned}
M_n(\widetilde{x}) &= \frac{1}{n}\sum_{i=1}^n \widetilde{K}_b(\widetilde{x} - \widetilde{X}_i) \begin{bmatrix} 1 \\ (\widetilde{x} - \widetilde{X}_i) \end{bmatrix} \begin{bmatrix} 1 \\ (\widetilde{x} - \widetilde{X}_i) \end{bmatrix}^\top, \\
\Psi_{n1}(\widetilde{x}) &= \frac{1}{n}\sum_{i=1}^n \widetilde{K}_b(\widetilde{x} - \widetilde{X}_i) \begin{bmatrix} 1 \\ (\widetilde{x} - \widetilde{X}_i) \end{bmatrix} \varepsilon_i \\
\Psi_{n2}(\widetilde{x}) &= \frac{1}{n}\sum_{i=1}^n \widetilde{K}_b(\widetilde{x} - \widetilde{X}_i) \begin{bmatrix} 1 \\ (\widetilde{x} - \widetilde{X}_i) \end{bmatrix} r_i(\widetilde{x}),
\end{aligned}$$

where $r_i(\widetilde{x}) = G(\widetilde{X}_i) - G(\widetilde{x}) - (\widetilde{x} - \widetilde{X}_i)^\top \nabla G(\widetilde{x})$. The argument is very similar to Fan, Mammen, and Härdle (1998, Theorem 1), and 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 (13). Letting

[14] 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.

$B = \text{diag}(1, b^{-(d+1)}, \ldots, b^{-(d+1)})$, we have

$$\begin{aligned} & BM_n(\widetilde{x})B \\ = \; & \frac{1}{n}\sum_{i=1}^{n} \widetilde{K}_b(\widetilde{x} - \widetilde{X}_i) \begin{bmatrix} 1 \\ (\widetilde{x} - \widetilde{X}_i)/b \end{bmatrix} \begin{bmatrix} 1 \\ (\widetilde{x} - \widetilde{X}_i)/b \end{bmatrix}^\top \\ = \; & E\widetilde{K}_b(\widetilde{x} - \widetilde{X}_i) \begin{bmatrix} 1 \\ (\widetilde{x} - \widetilde{X}_i)/b \end{bmatrix} \begin{bmatrix} 1 \\ (\widetilde{x} - \widetilde{X}_i)/b \end{bmatrix}^\top + O_p(a_n), \end{aligned} \tag{14}$$

where $a_n = \sqrt{\log n / nb^{d+1}}$ uniformly over v in the support of $H(V|x,n)$. The justification for this comes from Masry (1996a). Although he assumed continuous density, it is clear from the proofs that the argument goes through in our case provided the supremum over the compact set is replaced by the maximum over the sample realizations. We calculate the upper diagonal expectation in (14) by first conditioning on X,

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

Then using integration by parts

$$\begin{aligned} & \left| \int k_b(v - v')K_b(x - x')\left[dH(v', x'|n) - dH(v', x')\right] \right| \\ = \; & \left| \int k_b(v - v')K_b(x - x')\left[dH(v'|x', n) - dH(v'|x')\right] dH(x') \right| \\ = \; & \left| \frac{1}{b^2} \int k'\left(\frac{v - v'}{b}\right)\left[H(v'|x', n) - H(v'|x')\right] K_b(x - x')\, dH(x')dv' \right| \\ \leq \; & \sup_{v', x'} |H(v'|x', n) - H(v'|x')| \, \frac{1}{b^2} \int \left| k'\left(\frac{v - v'}{b}\right) \right| |K_b(x - x')|\, dH(x')dv' \\ \leq \; & \frac{1}{b} \sup_{v', x'} |H(v'|x', n) - H(v'|x')| = O(J^{-1}b^{-1}), \end{aligned}$$

by the integrability and smoothness on k. Similar arguments apply to the other terms in (14). Therefore,

$$BM_n(\widetilde{x})B = \begin{bmatrix} h(\widetilde{x}) & 0 \\ 0 & h(\widetilde{x})\mu_2(\widetilde{K}) \end{bmatrix} + O_p(c_n),$$

where $c_n = a_n + b + J^{-1}b^{-1}$, and the error is uniform over v in the support of $H(V|x,n)$. Then

$$e_1^\top M_n^{-1}(\widetilde{x})\Psi_{n1}(\widetilde{x}) = \frac{1}{h(\widetilde{x})}\frac{1}{n}\sum_{i=1}^{n}\widetilde{K}_b(\widetilde{x}-\widetilde{X}_i)\varepsilon_i + \mathrm{rem}(\widetilde{x}),$$

where $\mathrm{rem}(\widetilde{x})$ is a remainder term that is $o_p(n^{-1/2}b^{-(d+1)/2})$. This argument is repeated for the bias terms, and defining

$$\beta(\widetilde{x}) = \frac{\mu_2(k)}{2}\mathrm{tr}(\nabla^2 G(\widetilde{x})),$$

we obtain

$$\widehat{G}(\widetilde{x}) - G(\widetilde{x}) = \frac{1}{h(\widetilde{x})}\frac{1}{n}\sum_{i=1}^{n}\widetilde{K}_b(\widetilde{x}-\widetilde{X}_i)\varepsilon_i + b^2\beta(v,x) + \mathrm{rem}(\widetilde{x}),$$

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

$$\widehat{\mu}_{2r}(x) - \mu_r(x) = \int_{\rho_0(x)}^{\rho_1(x)} r'(v,x)[\widehat{G}(v \mid x) - G(v \mid x)]dv + O_p(J^{-1}b^{-1}).$$

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

$$\widehat{\mu}_{2r}(x) - \mu_r(x) = \frac{1}{n}\sum_{i=1}^{n}K_b(x-X_i)\frac{r'(V_i^n,x)}{h_{V,X}(V_i^n,x)}\varepsilon_i + b^2\overline{\beta}(x) + o_p(n^{-1/2}b^{-d/2}),$$

where $\overline{\beta}(x) = \int \beta(v,x)d\lambda(v)$. It follows that the asymptotic variance of $\widehat{\mu}_{2r}(x)$ is

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

by a change of variables and dominated convergence. Furthermore, the central limit theorem holds by the arguments used in Gozalo and Linton (1999, Lemma CLT). ■

6.3 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 below. 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 we have

$$\psi'(u) = E\frac{\partial \rho_j(u,\theta_0)}{\partial u} = -E\left(\left[h'(\Lambda|X_j)(\Lambda')^2 + h(\Lambda|X_j)\Lambda''\right](m(X_j,\theta_0) - u)\right). \tag{15}$$

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

$$\begin{aligned}
\widehat{\mu}_{3r}(x) &= \frac{1}{n}\sum_{i=1}^n f_1(Z_i,\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^n f_2(Z_i,\theta_0)[\widehat{\psi}(\widehat{U}_i) - \psi(U_i)] && (16)\\
&\quad -\frac{1}{n}\sum_{i=1}^n [f_2(Z_i,\widehat{\theta}) - f_2(Z_i,\theta_0)][\widehat{\psi}(\widehat{U}_i) - \psi(U_i)] && (17)\\
&\quad +\frac{1}{n}\sum_{i=1}^n \frac{r'[\Lambda(m(x,\widehat{\theta}) - \widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta}) - \widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\psi^2(U_i)\widehat{\psi}(\widehat{U}_i)}[\widehat{\psi}(\widehat{U}_i) - \psi(U_i)]^2, && (18)
\end{aligned}$$

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 (16), while (17) and (18) contain remainder terms.

Leading Terms. Lemma 1 implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n [f_1(Z_i,\widehat{\theta}) - Ef_1(Z_i,\theta_0)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\Gamma_F \varsigma(Z_i,\theta_0) + [f_1(Z_i,\theta_0) - Ef_1(Z_i,\theta_0)]\} + o_p(1), \tag{19}$$

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. Then, by Lemma 2 and the fact that $E|f_2(Z_i,\theta_0)| < \infty$, we have

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

where $L(Z_i, Z_j) = \xi_j(U_i) + \Gamma(Z_i)\varsigma(Z_j;\theta_0)$, 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 $E_j[\xi_j(U_i)] \neq 0$. 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

$$\sqrt{n}(\mathcal{Q}_n - \widehat{\mathcal{Q}}_n) = o_p(1).$$

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

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

Therefore,

$$
\begin{aligned}
\widehat{\mathcal{Q}}_n \;=\; & E\left[f_0(Z_i,\theta_0)\frac{\psi'(U_i)}{\psi(U_i)}\widetilde{\gamma}_i + \frac{f_0(Z_i,\theta_0)}{\psi(U_i)}\widetilde{\zeta}_{ij}\widetilde{\gamma}_j\right]\frac{1}{n}\sum_{j=1}^n \varsigma(Z_j;\theta_0) \\
& + \frac{1}{n}\sum_{j=1}^n E_j\left[f_0(Z_i,\theta_0)\frac{[\rho_j(U_i,\theta_0) - \psi(U_i)]}{\psi(U_i)}\right].
\end{aligned} \tag{20}
$$

We have shown that

$$
\frac{1}{n}\sum_{i=1}^n f_2(Z_i,\theta_0)[\widehat{\psi}(\widehat{U}_i) - \psi(U_i)] = \widehat{\mathcal{Q}}_n + o_p(n^{-1/2}),
$$

where $\widehat{\mathcal{Q}}_n$ is given in (20). This concludes the analysis of the leading terms.

Remainders. By the Cauchy-Schwarz inequality

$$
\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^n [f_2(Z_i,\widehat{\theta}) - f_2(Z_i,\theta_0)][\widehat{\psi}(\widehat{U}_i) - \psi(U_i)]\right| \\
\leq \;& \left(\frac{1}{n}\sum_{i=1}^n [f_2(Z_i,\widehat{\theta}) - f_2(Z_i,\theta_0)]^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n [\widehat{\psi}(\widehat{U}_i) - \psi(U_i)]^2\right)^{1/2} \\
= \;& O_p(n^{-1})
\end{aligned}
$$

from another application of Lemmas 1 and 2.

Therefore,

$$
\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^n \frac{r'[\Lambda(m(x,\widehat{\theta}) - \widehat{U}_i), x]\Lambda'(m(x,\widehat{\theta}) - \widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\psi^2(U_i)\widehat{\psi}(\widehat{U}_i)}[\widehat{\psi}(\widehat{U}_i) - \psi(U_i)]^2\right| \\
\leq \;& \frac{\sup_{u\in\mathcal{U}}[\widehat{\psi}(u) - \psi(u)]^2 + o_p(n^{-1/2})}{\inf_{u\in\mathcal{U}}\psi^3(u) + o_p(1)}\frac{1}{n}\sum_{i=1}^n |r'[\Lambda(m(x,\widehat{\theta}) - \widehat{U}_i), x]\Lambda'(m(x,\widehat{\theta}) - \widehat{U}_i)| \cdot (|Y_i| + 1) \\
= \;& o_p(n^{-1/2}).
\end{aligned}
$$

This result used the fact that $\min_{1\leq i\leq n}\widehat{\psi}(\widehat{U}_i) \geq \inf_{u\in\mathcal{U}}\psi(u) + o_p(1)$, which is proved in Lemma 2. Also $\inf_{u\in\mathcal{U}}\psi(u) > 0$.

In conclusion,

$$\sqrt{n}[\widehat{\mu}_{3r}(x) - \mu_r(x;\theta_0)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i + o_p(1),$$

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

$$\begin{aligned}
\widehat{\mu}^*_{4r}(x;\widehat{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} f_1(Z_i,\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} f_2(Z_i,\theta_0)[\widetilde{\psi}(\widehat{U}_i) - \psi(U_i)] && (21)\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}[f_2(Z_i,\widehat{\theta}) - f_2(Z_i,\theta_0)] \times [\widetilde{\psi}(\widehat{U}_i) - \psi(U_i)] && (22)\\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\frac{r'[\Lambda(m(x,\widehat{\theta}) - \widehat{U}_i), x]\Lambda'(m(x,\widehat{\theta}) - \widehat{U}_i)[Y_i - 1(\widehat{U}_i > 0)]}{\psi^2(U_i)\widetilde{\psi}(\widehat{U}_i)}[\widetilde{\psi}(\widehat{U}_i) - \psi(U_i)]^2. && (23)
\end{aligned}$$

The leading terms in this expansion are derived from (21), while (22) and (23) contain remainder terms.

Leading Terms. We make use of Lemma 3 given below. 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

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n} f_2(Z_i,\theta_0)\left[\widetilde{\psi}(\widehat{U}_i) - \psi(U_i) - \frac{1}{n}\sum_{j=1}^{n} L^*(Z_i,Z_j)\right]\right| &\leq \frac{1}{nb^3}\left(\frac{1}{n}\sum_{i=1}^{n}|f_2(Z_i,\theta_0)|d(X_i)\right)\\
&= O_p(n^{-1}b^{-3}). && (24)
\end{aligned}$$

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

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

Under our bandwidth conditions, the right hand side of (24) 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

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

uniformly in i. Define

$$\overline{f}_2(U_i) = E[f_2(Z_i, \theta_0) | U_i].$$

Then by iterated expectation

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

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

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

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,

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

by substituting in for f_2 and decomposing $f_0(Z_i, \theta_0) = \overline{f}_0(U_i) + f_0(Z_i, \theta_0) - \overline{f}_0(U_i)$. Using the same U-statististic 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

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

Remainder Terms. First,

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

Second

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\frac{r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)[Y_i-1(\widehat{U}_i>0)]}{\psi^2(U_i)\widetilde{\psi}(\widehat{U}_i)}[\widetilde{\psi}(\widehat{U}_i)-\psi(U_i)]^2\right| \\
\leq\ & \frac{\sup_{u\in\mathcal{U}}[\widetilde{\psi}(u)-\psi(u)]^2+o_p(n^{-1/2})}{\inf_{u\in\mathcal{U}}\psi^3(u)+o_p(1)}\frac{1}{n}\sum_{i=1}^{n}|r'[\Lambda(m(x,\widehat{\theta})-\widehat{U}_i),x]\Lambda'(m(x,\widehat{\theta})-\widehat{U}_i)|\cdot(|Y_i|+1) \\
=\ & o_p(n^{-1/2}).
\end{aligned}$$

This result used the fact that $\min_{1\leq i\leq n}\widetilde{\psi}(\widehat{U}_i)\geq\inf_{u\in\mathcal{U}}\psi(u)+o_p(1)$, which is proved in Lemma 3. ■

6.4 Subsidiary Results

Define

$$F_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}f(Z_i,\theta)$$

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* $\delta_n\to 0$,

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

Then

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

where

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

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

$$\Pr[||\sqrt{n}(\widehat{\theta} - \theta_0)|| > \delta_n] \to 0$$

as $n \to \infty$. We can therefore suppose that $||\sqrt{n}(\widehat{\theta} - \theta_0)|| \leq \delta_n$ with probability tending to one. 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* $n \to \infty$

$$\max_{1 \leq i \leq n} \left| \widehat{\psi}(\widehat{U}_i) - \psi(U_i) - \frac{1}{n} \sum_{j=1}^{n} L(Z_i, Z_j) \right| = o_p(n^{-1/2}) \tag{25}$$

$$\max_{1 \leq i \leq n} \left| \frac{1}{n} \sum_{j=1}^{n} L(Z_i, Z_j) \right| = O_p(n^{-1/2}) \tag{26}$$

$$\min_{1 \leq i \leq n} \widehat{\psi}(\widehat{U}_i) \geq \inf_{u \in \mathcal{U}} \psi(u) + o_p(1) \tag{27}$$

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

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

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

Proof. Regarding (26), 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.

Result (27) 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 (25) and (26).

Before showing (25) we show that

$$\max_{1\le i\le n}\widehat{U}_i \le \max_{1\le i\le n}U_i + o_p(1) \tag{28}$$

$$\min_{1\le i\le n}\widehat{U}_i \ge \min_{1\le i\le n}U_i + o_p(1), \tag{29}$$

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 $\mathcal{A}$

$$\begin{aligned}\Pr[\mathcal{A}] &\le \Pr\left[\mathcal{A} \text{ and } \{\widehat{U}_1,\ldots,\widehat{U}_n\}\subset\mathcal{U}\right] + \Pr\left[\widehat{U}_j \notin \mathcal{U} \text{ for some } j\right] \\ &\le \Pr\left[\mathcal{A} \text{ and } \{\widehat{U}_1,\ldots,\widehat{U}_n\}\subset\mathcal{U}\right] + o(1) = o(1).\end{aligned} \tag{30}$$

Proof of (28). 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 $\Pr[||\widehat{\theta}-\theta_0|| \ge \delta_n] \to 0$. Therefore, with probability tending to one

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

Furthermore, applying the Bonferroni and Markov inequalities

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

for any $\epsilon > 0$ when $r > 2$. This yields (28); (29) follows similarly.

We next prove (25). 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

$$\widehat{\psi}(\widehat{U}_i) - \widehat{\psi}(U_i) = \frac{1}{n}\sum_{j=1}^{n}\frac{\partial\rho_j(U_i(\theta_0),\theta_0)}{\partial\theta^\top}(\widehat{\theta}-\theta_0) + R_{ni}, \tag{31}$$

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 (30) we have in (31) that with probability tending to one

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

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

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

where $\partial \rho_j(u,\theta_0)/\partial u$ was defined in (15).

■

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

$$\max_{1\le i\le n}\left|\widetilde{\psi}(\widehat{U}_i) - \psi(U_i) - \frac{1}{n}\sum_{j=1}^n L^*(Z_i,Z_j)\right| \le \frac{k}{nb^3}d(X_i) \tag{32}$$

$$\max_{1\le i\le n}\left|\frac{1}{n}\sum_{j=1}^n L^*(Z_i,Z_j)\right| = O_p\left\{\left(\frac{\log n}{nb}\right)^{1/2}\right\} + O_p(b^2) \tag{33}$$

$$\min_{1\le i\le n}\widetilde{\psi}(\widehat{U}_i) \ge \inf_{u\in\mathcal{U}}\psi(u) + o_p(1) \tag{34}$$

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

$$\widetilde{\psi}(\widehat{U}_i) - \psi(U_i) = T_{ni} + R_{ni}, \tag{35}$$

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):

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

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

$$\max_{1\leq i\leq n}\left|\frac{1}{nb^2}\sum_{j=1}^{n}k'\left(\frac{U_i-U_j}{b}\right)d(X_j)-E_i\left[\frac{1}{b^2}k'\left(\frac{U_i-U_j}{b}\right)\overline{d}(U_j)\right]\right|=O_p\left\{\left(\frac{\log n}{nb^3}\right)^{1/2}\right\}, \tag{36}$$

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

$$\begin{aligned}
&\left|E\left[\frac{1}{b^2}k'\left(\frac{U_i-U_j}{b}\right)\overline{d}(U_j)\mid U_i\right]-[\overline{d}(U_i)\psi(U_i)]'\right| \qquad (37)\\
=&\left|\int\frac{1}{b^2}k'\left(\frac{U_i-u}{b}\right)\overline{d}(u)\psi(u)du-[\overline{d}(U_i)\psi(U_i)]'\right|\\
=&\left|\int\frac{1}{b}k\left(\frac{U_i-u}{b}\right)[\overline{d}(u)\psi(u)]'du-[\overline{d}(U_i)\psi(U_i)]'\right|\\
=&\left|\int k(t)\left([\overline{d}(U_i+tb)\psi(U_i+tb)]'-[\overline{d}(U_i)\psi(U_i)]'\right)dt\right|\\
=&\ O_p(b^2)
\end{aligned}$$

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 (36) and (37) 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,

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

uniformly in i.

In conclusion,

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

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

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

by the Masry results.

The proof of (34) follows as for (27).

■

6.5 Stochastic Equicontinuity Results

We now show that condition (iii) of Lemma 1 is satisfied in our case. Let $\Theta_n(c) = \{\theta\colon \sqrt{n}|\theta-\theta^0| \leq c\}$. 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 $n \geq n_0$, $\Pr[\widehat{\theta} \in \Theta_n(c_\epsilon)] \geq 1 - \epsilon$. 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

$$\rho(\theta, \theta') = E\left([f(Z_i, \theta) - f(Z_i, \theta')]^2 \right),$$

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 $\gamma \in \Gamma(c) \subset \mathsf{R}^p$. We establish the following result:

$$\sup_{\gamma \in \Gamma} |\nu_n(\gamma) - \nu_n(0)| = o_p(1) \tag{38}$$

To prove (38) 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 (38) 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 $m(Z_i, \theta^0)$:

$$m(Z_i, \theta^0 + \gamma n^{-1/2}) = m(Z_i, \theta^0) + \frac{1}{\sqrt{n}} \sum_{k=1}^{p} \frac{\partial m}{\partial \theta_k}(Z_i, \theta^0)\gamma_k + \frac{1}{n} \sum_{k=1}^{p} \sum_{r=1}^{p} \frac{\partial^2 m}{\partial \theta_k \partial \theta_r}(Z_i; \overline{\theta})\gamma_k \gamma_r \tag{39}$$

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

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

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). ■

References

[1] An, M.Y. (2000). "A Semiparametric Distribution for Willingness to Pay and Statistical Inference with Dichotomous Choice CV Data,"

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

[6] 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.

[7] Coppejans, M. (2003): "Effective Nonparametric Estimation in the Case of Severly Discretized Data," Duke University.

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

[9] Crooker, J.R., and J.A. Herriges (2000): "Parametric and Semi-Nonparametric Estimation of Willingness-to-Pay in the Dichotomous Choice Contingent Valuation Framework," Manuscript, Texas Tech University.

[10] Das, M. (2002), "Minimum Distance Estimators for Nonparametric Models With Grouped Dependent Variables," unpublished manuscript.

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

[12] Gozalo, P., and O.B. Linton (1999): "Local Nonlinear Least Squares: Using Parametric Information in Nonparametric Regression." *Journal of Econometrics,* 99, 63-106.

[13] Hardle, 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.

[14] Hengartner, N., and O. Linton (1996): "Nonparametric regression estimation at design poles and zeros," *The Canadian Journal of Statistics* 24, 583-591.

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

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

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

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

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

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

[21] Lu, Z-Q. (2002): "Nonparametric regression with Singular Design," Kowloon.

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

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

[24] Masry, E. (1996a), "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *J. Time Ser. Anal.* 17, 571-599.

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

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

[27] Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

[28] 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.

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

[30] Racine, J., and Q. Li (2002): "Nonparametric estimation of regression functions with categorical and continuous data," Manuscript, College Station.

[31] 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.

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

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

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

[35] 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.

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

[37] Tjostheim, D. and B. H. Auestad (1994): "Nonparametric Identification of Nonlinear Time Series: Projections," *Journal of the American Statistical Association,* 89, 1398-1409.

[38] 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
PRMSE	$\widehat{\mu}_1$	3.97	2.73	2.27	4.01	2.73	2.30	3.97	2.67	2.23
	$\widehat{\mu}_3$	14.10	12.79	12.56	14.04	12.85	12.60	13.92	12.79	12.52
	$\widehat{\mu}_4$	7.00	3.69	2.66	6.98	3.67	2.65	6.91	3.71	2.65
	$\widehat{\mu}_5$	4.90	3.48	3.14	4.98	3.48	3.13	4.97	3.50	3.16
PMAE	$\widehat{\mu}_1$	3.15	2.21	1.82	3.18	2.19	1.84	3.18	2.14	1.79
	$\widehat{\mu}_3$	12.49	12.13	12.15	12.37	12.17	12.20	12.27	12.12	12.12
	$\widehat{\mu}_4$	5.72	3.00	2.15	5.72	2.99	2.14	5.62	3.02	2.14
	$\widehat{\mu}_5$	3.94	2.88	2.66	4.01	2.87	2.65	3.99	2.89	2.68
IRMSE	$\widehat{\mu}_1$	14.56	12.63	12.21	14.55	12.64	12.20	14.59	12.59	12.17
	$\widehat{\mu}_3$	17.59	16.04	15.74	17.55	16.07	15.79	17.57	16.10	15.79
	$\widehat{\mu}_4$	12.74	10.41	9.89	12.73	10.38	9.90	12.85	10.51	10.00
	$\widehat{\mu}_5$	11.74	10.33	10.02	11.76	10.31	10.03	11.92	10.43	10.14
IMAE	$\widehat{\mu}_1$	10.96	10.16	10.05	10.95	10.15	10.04	10.98	10.08	9.97
	$\widehat{\mu}_3$	14.42	13.36	13.13	14.35	13.38	13.18	14.34	13.40	13.16
	$\widehat{\mu}_4$	10.08	8.50	8.22	10.06	8.48	8.23	10.17	8.59	8.32
	$\widehat{\mu}_5$	9.23	8.44	8.31	9.25	8.42	8.32	9.38	8.53	8.42

Table 1: Conditional Mean in 5-bid design; 10,000 replications; Bandwidth by Silverman's Thumb; Pointwise Root Mean Squared and Absolute Errors (PRMSE and PMAE) calculated at $x = 0$.

		$\sigma = 5$			$\sigma = 10$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
PRMSE	$\widehat{\mu}_1$	7.90	3.07	1.24	4.12	2.98	4.50	37.49	41.47	43.19
	$\widehat{\mu}_3$	11.18	6.43	5.40	12.09	9.99	9.86	45.97	48.60	49.37
	$\widehat{\mu}_4$	9.91	7.47	6.85	11.10	9.70	9.30	45.90	46.78	46.75
	$\widehat{\mu}_5$	28.14	26.41	26.03	24.18	22.07	21.33	28.74	24.32	22.50
PMAE	$\widehat{\mu}_1$	7.10	2.61	0.94	3.26	2.61	4.36	37.29	41.42	43.16
	$\widehat{\mu}_3$	8.12	5.65	5.17	10.97	9.76	9.75	44.62	48.29	49.25
	$\widehat{\mu}_4$	7.59	6.34	5.98	10.33	9.26	8.84	44.78	46.19	46.24
	$\widehat{\mu}_5$	24.26	24.38	24.65	21.12	20.17	19.92	23.56	21.38	20.59
IRMSE	$\widehat{\mu}_1$	17.37	17.49	17.42	13.11	13.24	13.24	30.48	30.14	30.30
	$\widehat{\mu}_3$	11.18	6.43	5.40	12.09	9.99	9.86	45.97	48.60	49.37
	$\widehat{\mu}_4$	9.91	7.47	6.85	11.10	9.70	9.30	45.90	46.78	46.75
	$\widehat{\mu}_5$	21.93	20.17	19.75	18.66	16.69	16.08	33.37	32.01	31.66
IMAE	$\widehat{\mu}_1$	15.74	15.71	15.46	11.39	11.47	11.57	29.40	29.13	29.25
	$\widehat{\mu}_3$	8.12	5.65	5.17	10.97	9.76	9.75	44.62	48.29	49.25
	$\widehat{\mu}_4$	7.59	6.34	5.98	10.33	9.26	8.84	44.78	46.19	46.24
	$\widehat{\mu}_5$	18.05	17.50	17.39	15.83	14.71	14.38	29.72	29.33	29.25

Table 2: Conditional Standard Deviation in 5-bid design; 10,000 replications;
Bandwidth by Silverman's Thumb

		$\sigma = 5$			$\sigma = 10$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
PRMSE	$\widehat{\mu}_1$	5.59	3.32	2.60	6.19	3.80	3.05	11.30	7.34	5.92
	$\widehat{\mu}_3$	10.65	7.05	6.28	10.42	7.11	6.24	12.42	9.63	8.90
	$\widehat{\mu}_4$	6.27	3.33	2.51	6.42	3.45	2.67	9.12	5.11	3.96
	$\widehat{\mu}_5$	5.78	3.35	2.61	5.88	3.39	2.60	7.56	4.38	3.42
PMAE	$\widehat{\mu}_1$	4.40	2.63	2.06	4.91	3.01	2.43	9.00	5.83	4.72
	$\widehat{\mu}_3$	8.56	5.73	5.22	8.35	5.80	5.19	10.16	8.30	7.93
	$\widehat{\mu}_4$	5.02	2.68	2.01	5.13	2.75	2.14	7.26	4.09	3.16
	$\widehat{\mu}_5$	4.59	2.67	2.08	4.70	2.71	2.08	6.04	3.49	2.72
IRMSE	$\widehat{\mu}_1$	12.30	7.54	6.02	12.37	7.57	6.05	15.76	11.12	9.82
	$\widehat{\mu}_3$	12.65	8.05	6.98	12.46	8.10	6.94	15.90	12.38	11.55
	$\widehat{\mu}_4$	9.22	5.12	3.93	9.35	5.19	4.03	13.37	9.24	8.32
	$\widehat{\mu}_5$	8.91	5.13	4.00	9.00	5.15	3.99	12.31	8.85	8.07
IMAE	$\widehat{\mu}_1$	8.81	5.41	4.33	8.96	5.47	4.40	12.05	8.51	7.54
	$\widehat{\mu}_3$	9.99	6.46	5.70	9.80	6.52	5.66	12.35	9.80	9.24
	$\widehat{\mu}_4$	7.14	3.95	3.02	7.24	4.01	3.12	10.40	7.23	6.57
	$\widehat{\mu}_5$	6.87	3.97	3.08	6.95	3.98	3.09	9.53	6.90	6.36

Table 3: Conditional Mean in Continuous design; 10,000 replications; Bandwidth by Silverman's Thumb

		$\sigma = 5$			$\sigma = 10$			$\sigma = 50$		
		n=100	n=300	n=500	n=100	n=300	n=500	n=100	n=300	n=500
PRMSE	$\widehat{\mu}_1$	9.59	7.43	6.49	7.49	5.43	4.55	10.74	7.91	7.17
	$\widehat{\mu}_3$	40.77	41.70	41.96	37.21	37.90	38.04	18.75	18.18	18.08
	$\widehat{\mu}_4$	18.37	12.46	10.44	16.99	12.00	10.35	15.96	8.70	6.92
	$\widehat{\mu}_5$	20.38	15.14	13.32	19.07	14.52	12.72	24.20	15.84	12.30
PMAE	$\widehat{\mu}_1$	8.86	7.10	6.25	6.41	4.79	4.05	8.68	6.65	6.16
	$\widehat{\mu}_3$	38.72	41.16	41.66	35.33	37.34	37.73	17.15	17.58	17.73
	$\widehat{\mu}_4$	14.01	9.95	8.54	14.39	10.62	9.25	11.89	6.72	5.49
	$\widehat{\mu}_5$	15.15	11.68	10.55	15.88	12.62	11.21	18.37	11.54	9.12
IRMSE	$\widehat{\mu}_1$	12.13	10.00	9.08	10.98	8.98	8.08	18.80	14.35	12.92
	$\widehat{\mu}_3$	40.77	41.70	41.96	37.21	37.90	38.04	18.75	18.18	18.08
	$\widehat{\mu}_4$	18.37	12.46	10.44	16.99	12.00	10.35	15.96	8.70	6.92
	$\widehat{\mu}_5$	18.51	14.00	12.20	17.54	13.36	11.82	24.18	18.16	16.00
IMAE	$\widehat{\mu}_1$	10.05	8.43	7.69	9.35	7.59	6.77	14.06	11.05	10.33
	$\widehat{\mu}_3$	38.72	41.16	41.66	35.33	37.34	37.73	17.15	17.58	17.73
	$\widehat{\mu}_4$	14.01	9.95	8.54	14.39	10.62	9.25	11.89	6.72	5.49
	$\widehat{\mu}_5$	13.34	10.53	9.43	14.20	11.28	10.08	18.32	13.42	11.95

Table 4: Conditional Standard Deviation in Continuous design; 10,000 replications;
Bandwidth by Silverman's Thumb