ESTIMATING FEATURES OF A DISTRIBUTION FROM BINOMIAL DATA

Arthur Lewbel
Oliver Linton
Daniel McFadden

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP07/01

ESTIMATING FEATURES OF A DISTRIBUTION FROM BINOMIAL DATA[¤]

Arthur Lewbel[y]
Boston College

Oliver Linton[z]
London School of Economics

Daniel McFadden
University of California, Berkeley[x]

December 4, 2001

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. One application is destructive duration analysis, where the process is censored at an observation test time. Another is referendum contingent valuation in resource economics, where one is interested in features of the distribution of values placed by consumers on a public good such as endangered species. Sample consumers are asked whether they would vote for 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.

[¤] This research was supported in part by the National Science Foundation through grants SES-9905010 and SBR-9730282, by the Economic and Social Research Council grant R000239252, and by the E. Morris Cox Endowment.

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

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

[x] 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. Three applications illustrate the problem:

Bioassay - Find the distribution of survival times 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, a test time, and an indicator for the test result.

Destructive Testing - Find the distribution of speeds at which air bags fail to protect passengers in automobile crashes. At speeds selected by experimental design, drive cars into a barrier and determine whether a dummy occupant is injured. An observation consists of covariates, a test speed, and an indicator for injury.

Survey research with Shadow Effects - Find the distribution of a household economic variable such as wealth. Subjects are asked if their economic variable exceeds a test value chosen by design. An observation consists of covariates, a test value, an indicator for the response. Follow up queries are 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 willingness to pay (WTP) for economic goods.[1]

Given a set of covariates, when the experimental design is randomized with a strictly positive test value density and mild regularity conditions, we propose consistent estimators for conditional (on covariates) moments of the unknown distribution. We also provide root n consistent estimators for the case where the unknown distribution depends on covariates through a single index location shift. In addition, we provide estimators of conditional quantiles of the unknown distribution.

2 Model Specification

The goal is estimation of conditional moments or quantiles of a latent, unobserved random scalar W, conditioned on a vector of observed covariates X. The conditional cumulative distribution function

[1] McFadden (1994) provides references and experminatal 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.

of W, denoted $G(w \mid x)$, is unknown but assumed to be smooth.

A test value v is set by a randomized experimental design or natural experiment. The value v is a realization of a random variable V; drawn from either a known or unknown conditional density $h(v \mid x)$ (we consider both cases). It is assumed that W is conditionally independent of V, conditioning on X (consistent with experimental design).

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 random sample of realizations of covariates X, test values V, and outcomes Y. The framework is similar to random censored regressions (with censoring point v), except that for random censoring we would observe w for observations having $w > v$, whereas in the present context we only observe $y = I(w > v)$.

Given a function $r(w; x)$, the goal is estimation of the conditional moment $\mu_r(x) = E[r(W; X) \mid X = x]$ for any chosen x in the support of X. Of particular interest are the moments based on $r(W; X) = W^k$ for integers k: In addition to moments we may also be interested in quantiles. Let $w_q(x)$ denote the q'th quantile of W given x:

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 conditional moments $\mu_r(x)$ and quantiles $w_q(x)$ defined in terms of those parameters.

Assuming that the conditional distribution of W given X is not finitely parameterized, we propose semiparametric and nonparametric estimators for these moments and quantiles. The semiparametric estimators assume that the conditional mean of W is finitely parameterized. The nonparametric estimators only require smoothness assumptions, but suffer from the usual curse of dimensionality. We provide limit normal distributions for these estimators. The semiparametric estimators all converge at the rate that would be obtained if draws w were observed.

In the example of willingness-to-pay models, W_i would be an individual i's unknown willingness to pay for a resource, V_i would be a bid that was posed to the individual, and X_i would be observable characteristics of the individual (such as age, income level, geographic location, and political party affiliation). Objects of interest might include the average willingness-to-pay for the resource among individuals in certain locations and income levels, or for voting models, the median willingness-to-pay among subsets of likely voters.

The next section provides results that will form the basis for the proposed estimators. Later sections provide limiting distributions.

3 Identification

Make the following assumptions.

Assumption A.1. The covariate vector X has compact support $\mathcal{X} \subseteq R^d$. The latent scalar W has an unknown, twice continuously differentiable conditional c.d.f. $G(w \mid x)$, with a compact support $[\alpha_0(X), \alpha_1(X)]$. The test variable V is continuously distributed with a known or unknown positive probability density function $h(v \mid x)$ having compact support $[\delta_0(X), \delta_1(X)]$ such that $\delta_0(X) \leq \alpha_0(X)$ and $\delta_1(X) \geq \alpha_1(X)$. The variables W and V are conditionally independent, given X. Let $Z = (X, V, Y)$.

Define $m(v, x)$ by

$$m(v,x) = E[Y|V = v, X = x]$$

and let m^{-1} be the inverse of the function m with respect to its first element (which exists on the support of W given assumption A.1), so if $t = m(v, x)$ then $v = m^{-1}(t \mid x)$ for $v \in [\alpha_0(X), \alpha_1(X)]$.

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. Let κ be a known constant that is in the support of W. The moment $\mu_r(x)$ exists, where $\mu_r(x)$ is defined by

$$\mu_r(x) = E[r(W, X) \mid X = x].$$

Define $r'(w, x) = \partial r(w, x)/\partial w$ and $s_r(z)$ by

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

For any regular function r, Theorem 1 below provides an expression for the conditional mean $\mu_r(x)$. Also provided is the q'th conditional quantile of W given x, denoted $w_q(x)$:

Theorem 1. Let Assumptions A.1 and A.2 hold. Then

$$\begin{aligned} \mu_r(x) &= E[s_r(Z) \mid X = x]. \\ w_q(x) &= m^{-1}(1 - q \mid x) \end{aligned}$$

Proof of Theorem 1. First observe that, given the conditional independence of W and V,

$$m(v,x) = E[Y|V = v, X = x] = 1 - G(v|x).$$

Next, by definition, $\mu_r(x) = \int_{\alpha_0(x)}^{\alpha_1(x)} r(v,x)[\partial G(v|x)/\partial v]dv$. Integration by parts yields

$$\mu_r(x) = r(v,x)[G(v|x) - 1(v \geq \theta)]|_{v=\alpha_0(x)}^{\alpha_1(x)} - \int_{\alpha_0(x)}^{\alpha_1(x)} r'(v,x)[G(v|x) - 1(v \geq \theta)]dv.$$

Therefore, collecting terms we find that

$$\begin{aligned} \mu_r(x) &= r(\theta,x) + \int_{\alpha_0(x)}^{\alpha_1(x)} r'(v,x)[E(Y|V = v, X = x) - 1(v < \theta)]dv \\ &= \int_{\alpha_0(x)}^{\alpha_1(x)} E[s_r(Z)|V = v, X = x)h(v \mid x)dv \\ &= E[s_r(Z) \mid X = x], \end{aligned}$$

where the last equality uses the assumptions regarding the supports of W and V, and the law of iterated expectation. The conditional quantile expression follows from $G(v|x) = 1 - m(v,x)$. ■

Theorem 1 provides the basis for the nonparametric moment estimators described in the next section, and for some semiparametric and quantile estimators. Essentially, based on Theorem 1, $\mu_r(x)$ may be estimated as the fitted values of either a parametric or nonparametric regression of $s_r(z)$ on x.

Corollary 1 below will be used to obtain faster converging moment and quantile estimators, based on stronger assumptions.

Assumption A.3. The latent W satisfies $W = g(X,\theta_0) - \varepsilon$, where g is a known function, $\theta_0 \in \Theta$ is a vector of parameters, and ε is a disturbance that is distributed independently of V, X, with unknown, twice continuously differentiable c.d.f. $G^*(\varepsilon)$ and compact support $[a_0, a_1]$ that contains zero. Define $U = g(X,\theta_0) - V$. Let $\psi(U)$ denote the unconditional probability density function of U. The support of U contains the interval $[a_0, a_1]$.

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

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

Corollary 1. Let Assumptions A.1, A.2, and A.3 hold. Then

$$\begin{aligned} Ã(u) &= E[h(g(X;\mu_0)\ ¡\ U)\ j\ U = u] \\ G^{¤}(u) &= E(Y\ j\ U = u) \\ ¹_r(x) &= E[s_r^{¤}(x;U;Y)] \\ w_q(x) &= g(x;\mu_0)\ ¡\ G^{¤\,¡\,1}(1\ ¡\ q) \end{aligned}$$

Proof of Corollary 1. Having $Ã(u) = E[h(g(X;\mu_0)\ ¡\ u)]$ follows from the definitions of U, $Ã$, and h: Also from definitions, $Y = I(" < U)$ which implies that $G^{¤}(u) = E(Y\ j\ U = u)$. Next, following the same steps as in Theorem 1 we have

$$\begin{aligned} ¹_r(x) &= \overset{a_1}{\underset{a_0}{s}} r[g(x;\mu_0)\ ¡\ u;x][@G^{¤}(u)=@u]du \\ &= r[g(x;\mu_0);x] + \overset{a_1}{\underset{a_0}{s}} r^0[g(x;\mu_0)\ ¡\ u;x][G^{¤}(u)\ ¡\ 1(u > 0)]du \\ &= \overset{a_1}{\underset{a_0}{s}} E[s_r^{¤}(x;U;Y)\ j\ U = u]Ã(u)du = E[s_r^{¤}(x;U;Y)] \end{aligned}$$

Finally, the quantile expression follows from $G(W\ j\ X = x) = 1\ ¡\ G^{¤}[g(X;\mu_0)\ ¡\ W]$. ■

The advantage of Corollary 1 over Theorem 1 for estimation is that in Corollary 1, $¹_r(x)$ and $Ã(u)$ are expressed as unconditional expectations and so can be estimated using ordinary sample averages (given an estimate of μ). Similarly, using Corollary 1 estimation of the quantiles $w_q(x)$ given μ only requires estimation of the one dimensional regression $G^{¤}(u) = E(Y\ j\ U = u)$, instead of the high dimensional $m(v;x)$.

4 Estimators

Assume that a random sample $Z_i = (X_i;V_i;Y_i)$ for $i = 1;:::;n$ is observed, where V_i is a realization of V, Y_i is a realization of Y, and X_i is a realization of X. Theorem 1 and Corollary 1 suggest a number of possible estimators for $¹_r(x)$: To describe these estimators, let $\overset{b}{E}$ denote an estimated expectation. An unconditional estimated expectation just denotes the sample average, while a conditional estimated expectation denotes a nonparametric regression.

4.1 Nonparametric Estimators

Let Assumptions A.1 and A.2 hold.

If the experimental design, and hence the density function h, is known, then $s_r(z_i)$ can be constructed for each observation i, and $\mu_r(x)$ may then be consistently estimating by nonparametrically regressing $s_r(z)$ on x. This first estimator is

$$\hat\mu_{1r}(x) = \hat E[s_r(z) \mid X = x]$$

Note that $\hat\mu_{1r}(x)$ depends on the design density h. One could replace $h(v \mid x)$ with an estimate $\hat h(v \mid x)$ (using, e.g., kernel density estimation) in the definition of $s_r(z)$. Call the result $\hat s_r(z)$. An estimator of $\mu_r(x)$ that can be used when h is unknown is then $\hat\mu^{*}_{1r}(x) = \hat E[\hat s_r(z) \mid X = x]$.

An estimator that does not entail knowing or estimating the density h is the following. Recall that $m(v,x) = E[Y \mid V = v, X = x]$. Let $\hat m(v,x)$ be a consistent estimator of m, that is, a nonparametric regression of y on x, v, so

$$\hat m(v,x) = \hat E[Y \mid V = v, X = x]$$

Let a_0 and a_1 be known or estimated constants such that $a_0 \le \alpha_0(x)$ and $a_1 \ge \alpha_1(x)$. Then, based on the proof of Theorem 1, a consistent estimator of $\mu_r(x)$ is given by

$$\hat\mu_{2r}(x) = r(\cdot, x) + \int_{a_0}^{a_1} r'(v,x)[\hat m(v,x) - 1(v < \cdot)]dv,$$

where the integral may be evaluated numerically. We give some more details later about the construction of the estimators.

4.2 Conditional Mean Estimation

This section considers estimation of the conditional mean of the unobserved W, when this conditional mean is finitely parameterized. This structure permits estimation at rate root n, instead of the slow convergence of the estimators described in the previous section.

For this section, continue to let Assumptions A.1 and A.2 hold, but assume also that $E(W \mid X = x) = \mu(x, \beta_0)$ for all $x \in \mathcal{X}$, where $\mu(x, \beta_0)$ is a known function and $\beta_0 \in \mathcal{B}$ is an unknown vector of parameters. The function $\mu(x, \beta)$ equals $\mu_r(x)$ for $r(w, x) = w$. Whatever other moments may be considered, it is likely that the conditional mean function μ would be of interest as well.

Assume the identification condition $\Pr[E(W \mid X = x) \neq \mu(X, \beta)] > 0$ for all $\beta \neq \beta_0$, $\beta \in \mathcal{B}$. It follows from Theorem 1 that

$$\mu(x, \beta_0) = E\left[\frac{Y - 1(V < 0)}{h(V \mid X)} \mid X = x\right]$$

This suggests the conditional mean estimator $\mu(x, \hat\beta)$, where $\hat\beta$ is defined by

$$\hat\beta = \arg\min_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - 1(V_i < 0)}{h(V_i \mid X_i)} - \mu(X_i, \beta)\right)^2 \hat\tau(Z_i),$$

where $\widehat{\omega}(z)$ is a known or estimated positive weight function chosen for efficiency. The estimator $\widehat{\theta}$ is an ordinary nonlinear weighted least squares, and so $\widehat{\theta}$, and therefore also $\mu(x;\widehat{\theta})$ is root n consistent and asymptotically normal with a standard limiting distribution, under standard regularity conditions. No nonparametric plug in functions are required. This estimator might not be efficient, since it violates the principle of ancillarity due to its dependence on the design density h.

If h is not known, one could replace $h(V \mid X)$ with an estimate $\widehat{h}(V \mid X)$ in the definition of $\widehat{\theta}$. The resulting estimator would then take the form of an ordinary two step estimator with a nonparametric first step (the estimation of h) which, with regularity, will be root n consistent and asymptotically normal. This estimator 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.

4.3 Semiparametric Estimators

This section discusses rate root n estimation of arbitrary conditional moments based on Corollary 1. For these estimators we let Assumption A.3 hold, in addition to Assumptions A.1 and A.2. It will be convenient to first consider the case where θ_0 in Assumption 3 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 g to equal zero.

4.3.1 Estimation With Known θ

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

$$\widehat{\psi}(u) = \frac{1}{n}\sum_{i=1}^{n} h[g(X_i;\theta_0) - u].$$

Then, based on Corollary 1, we have the estimator

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

This estimator is computationally extremely simple, since it entails only sample averages. Special cases of the estimator $\widehat{\mu}_{3r}(x)$ were proposed by McFadden (1994) and by Lewbel (1997).

Let $\widetilde{\psi}(u)$ be an estimator of $\psi(u)$ that does not depend on 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

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

which may be used when h is unknown.

Another approach to estimating $\mu_r(x)$ uses ordered observations in place of a preliminary estimate of ψ. Augment the observed (U,Y) pairs with the artificial observations $(\alpha_0,0)$ and $(\alpha_1,1)$. Recode each observation (U,Y) having $U<\alpha_0$ as $(\alpha_0,0)$ and each observation having $U>\alpha_1$ as $(\alpha_1,1)$. Then, index the observations so that the U's, including the artificial ones, are in non-decreasing order, and denote them by $U_{n0}\le\cdots\le U_{n,n+1}$. The probability of ties in the interior of the support is zero. Let Y_{ni} denote the observed Y associated with U_{ni}. The proposed ordered data estimator is then

$$\hat{\mu}_{5r}(x) = r[g(x,\theta_0),x] + \sum_{i=1} r'[g(x,\theta_0) - U_{ni},x][Y_{ni} - 1(U_{ni}>0)]\frac{U_{n,i+1} - U_{n,i-1}}{2}$$

This estimator can be interpreted as being the same as $\hat{\mu}_{4r}(x)$, with the density $\psi(u)$ estimated by differencing the empirical distribution of U, although this function, $2n/(U_{n,i+1} - U_{n,i-1})$, is not a consistent estimator of $\psi(U_i)$. The estimator $\hat{\mu}_{5r}(x)$ has the advantages of just equalling a sample average and of not requiring knowledge of h, however, it can be shown to be less efficient than $\hat{\mu}_{4r}(x)$ with an optimally estimated $\hat{\psi}(U_i)$.

4.3.2 Estimation with Unknown θ

Assumptions A.1, A.2, and A.3 imply that $E(W \mid X=x) = \alpha + g(x,\theta)$ for some arbitrary location constant α (since no location constraint is imposed upon ε). Therefore, based on the estimator of $\mu(x,\beta)$ described in section 4.2, we may obtain, a root n consistent, asymptotically normal estimate $\tilde{\theta}$ of θ by the simple least squares criterion

$$\tilde{\theta} = \arg\min_{\theta}\left[\min_{\alpha}\frac{1}{n}\sum_{i=1}\left(\frac{Y_i - 1(V_i<0)}{h(V_i \mid X_i)} - \alpha - g(X_i,\theta)\right)^2\right]$$

An estimator $\hat{h}$ may be used in place of h if h is unknown.

Assumptions A.1, A.2, and A.3 make the latent error ε independent of X, and therefore the binary choice estimator of Klein and Spady (1993) will provide a semiparametrically efficient estimator of θ (note that the Klein and Spady estimator does not identify a location constant α, but that is not required, since no location constraint is imposed upon ε). We will provide a numerically simpler estimator that is asymptotically equivalent to Klein and Spady.

Let $\tilde{\theta}$ denote the chosen 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 $\hat{\mu}_{\lambda r}(x,\theta)$ for $\lambda = 3,4,$ or 5. Note that in addition to directly appearing in the equations for $\hat{\mu}_{\lambda r}$, θ also appears in the definition of $U_i = g(X_i,\theta) - V_i$. We later derive the root N consistent, asymptotically normal limiting distribution for each estimator $\hat{\mu}_{\lambda r}(x,\tilde{\theta})$. The estimators are not differentiable in U_i, which complicates the derivation of their limiting distribution (e.g., Theorem 6.1 of Newey and McFadden (1994) is not directly applicable due to this nondifferentiability).

5 Quantile Estimators

In addition to moments, one may also desire estimates of conditional quantiles of W. Let Assumptions A.1 and A.2 hold and define $\widehat{G}(v \mid x) = 1 - \widehat{m}(v;x)$. Then, based on Theorem 1, an estimate of the q'th quantile of W given $X = x$ is just

$$\widehat{w}_q(x) = \widehat{G}^{-1}(q \mid x)$$

The rate of convergence of this estimate will be slow, because of the high dimension of $\widehat{m}$.

If Assumption A.3 holds in addition to A.1 and A.2, then faster convergence is possible. Given Corollary 1 we have $U = g(X;\theta_0) - V$, $G^*(u) = E(Y \mid U = u)$, and $w_q(x) = g(x;\theta_0) - G^{*-1}(1 - q)$. Therefore, let $\widehat{U}_i = g(X_i;\widehat{\theta}) - V_i$ and estimate the conditional quantile $w_q(x)$ by

$$\begin{aligned} \widehat{G}^*(u) &= \widehat{E}(Y \mid \widehat{U} = u) \\ \widetilde{w}_q(x) &= g(x;\widehat{\theta}) - \widehat{G}^{*-1}(1 - q) \end{aligned}$$

where the function $\widehat{G}^*$ is obtained by nonparametrically regressing Y on $\widehat{U}$, and is then numerically inverted to obtain $\widehat{G}^{*-1}$. This estimator $\widetilde{w}_q(x)$ will converge at a faster rate than the nonparametric quantile estimator $\widehat{w}_q(x)$: With sufficient regularity, $\widetilde{w}_q(x)$ is asymptotically normal and converges at the same rate as a one dimensional nonparametric regression estimator, i.e., the same as the best rate that could be obtained if realizations of the latent w were observed.

6 Estimation Details and Distribution Theory

In this section we provide a bit more detail about the computation of the estimators $\widehat{\mu}_{jr}(x)$ and their distribution theory.

6.1 Nonparametric Estimators

We first consider a fairly general class of nonparametric estimators and then specialize to kernels. Specifically, we consider a class of linear estimators of m; that is, let

$$\widehat{m}(v;x) = \sum_{i=1} w_{ni}(v;x)Y_i; \qquad (1)$$

where $w_{ni}(v;x)$ are some smoothing weights that depend only on $\{(X_1;V_1);\ldots;(X_n;V_n)\}$ and satisfy certain conditions as in Stone (1982). This includes a large class of commonly used estimation schemes such as kernels, local polynomial, nearest neighbor, series, smoothing splines, etc., see Härdle and Linton (1994) for further discussion. It does exclude local median or local quantile estimators whether they be based on kernels or nearest neighbors; it also excludes the popular neural networks class of estimators. Finally, this framework so far excludes methods that have selected the smoothing parameter based on the data. However, it should be possible to extend the treatment to a class of asymptotically linear smoothers, which includes almost any smoothing method that can be asymptotically normal.

We shall suppose also that the smoothing weights in (1) satisfy exactly $\sum_{i=1}^n w_{ni}(v;x) = 1$; this is the case for many linear estimators. In this case, we can write

$$\hat{m}(v;x) - m(v;x) = \sum_{i=1} w_{ni}(v;x)\varepsilon_i + \sum_{i=1} w_{ni}(v;x)\{m(V_i;X_i) - m(v;x)\}; \tag{2}$$

where the error term $\varepsilon_i = Y_i - m(V_i;X_i)$ is independent across i and satisfies $E(\varepsilon_i|V_i;X_i) = 0$; define also $\sigma_i^2 = \mathrm{var}(\varepsilon_i|V_i;X_i)$: The first term on the right hand side of (2) is conditional mean zero and determines the limiting variance, while the second term determines the bias. Under additional conditions, we can approximate the bias term $\sum_{i=1}^n w_{ni}(v;x)\{m(V_i;X_i) - m(v;x)\}$ by $\delta_n\beta(v;x)$ for some bounded and continuous function $\beta(v;x)$ and deterministic sequence $\delta_n \to 0$ as $n \to \infty$: We can also replace the weights $w_{ni}(v;x)$ in $\sum_{i=1}^n w_{ni}(v;x)\varepsilon_i$ by some approximation $\tilde{w}_{ni}(v;x)$ that depends only on $(X_i;V_i)$. We arrive at the expansion

$$\hat{m}(v;x) - m(v;x) = \sum_{i=1} \tilde{w}_{ni}(v;x)\varepsilon_i + \delta_n\beta(v;x) + R_n(v;x); \tag{3}$$

where $R_n(v;x)$ is a remainder term that contains the various approximation errors described above. We next state conditions under which $\hat{m}(v;x)$ is asymptotically normal.

Theorem 2. Suppose that: (i) $0 < \underline{\sigma}^2 \le \sigma_i^2 \le \overline{\sigma}^2 < \infty$; (ii) $R_n(v;x)/\min\{\alpha_n;\delta_n\} \to^p 0$; where $\alpha_n = 1/\sqrt{\sum_{i=1}^n \tilde{w}_{ni}^2(v;x)} \to^p 0$; and (iii) $\max_{1\le i\le n} \tilde{w}_{ni}^2(v;x)/\sum_{i=1}^n \tilde{w}_{ni}^2(v;x) \to^p 0$: Then

$$\frac{\hat{m}(v;x) - m(v;x) - \delta_n\beta(v;x)}{\sqrt{\sum_{i=1}^n \tilde{w}_{ni}^2(v;x)\sigma_i^2}} \xrightarrow{d} N(0;1):$$

This result is a standard application of the Lindeberg-Feller central limit theorem. The magnitude of the bias term, δ_n; depends on the method used and on the smoothness of m (and perhaps also on the smoothness of the covariate density). The magnitude of α_n depends on the estimation method and on the covariate density in general. The optimal rate in the central limit theorem is achieved when α_n and δ_n are the same magnitude.

By appropriately redefining y and m, Theorem 2 can be immediately applied to yield the limiting distribution for $\hat{\mu}_{1r}(x)$.

Now consider estimating $\mu_r(x)$ using $\hat{\mu}_{2r}(x)$, which is equivalent to

$$\hat{\mu}_{2r}(x) = r(\cdot,x) + \int_{\alpha_0(x)}^{\alpha_1(x)} r'(v,x)[\hat{m}(v,x) - 1(v < \cdot)]dv$$

where $\hat{m}(v,x)$ is defined in (1). This estimator 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 Tjøstheim and Auestad (1994), except that the integrating measure λ, where $d\lambda(v) = -r'(v,x)1(\alpha_0(x) \le v \le \alpha_1(x))dv$, 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 been worked out for specific smoothing methods like kernels or nearest neighbors. We give a derivation at the higher level of generality given by the definition (1). Then,

$$\begin{aligned}\hat{\mu}_{2r}(x) - \mu_r(x) &= \int_{\alpha_0(x)}^{\alpha_1(x)} r'(v,x)[\hat{m}(v,x) - m(v,x)]dv \\ &= \sum_{i=1} \bar{w}_{ni}(x)\varepsilon_i + \delta_n\bar{\beta}(x) + \bar{R}_n(x),\end{aligned}$$

where: $\bar{w}_{ni}(x) = \int w_{ni}(v,x)d\lambda(v)$, $\bar{\beta}(x) = \int \beta(v,x)d\lambda(v)$, and $\bar{R}_n(x) = \int R_n(v,x)d\lambda(v)$. We next state conditions under which $\hat{\mu}_{2r}(x)$ is asymptotically normal.

Theorem 3. Suppose that condition (i) from Theorem 2 is true, and that: (i) $\bar{R}_n(x)/\min\{\bar{\alpha}_n, \delta_n\} \xrightarrow{p} 0$, where $\bar{\alpha}_n = 1/\sqrt{\sum_{i=1}^n \bar{w}_{ni}^2(x)} \xrightarrow{p} 0$, and (ii) $\max_{1\le i\le n} \bar{w}_{ni}^2(x)/\sum_{i=1}^n \bar{w}_{ni}^2(x) \xrightarrow{p} 0$. Then,

$$\frac{\hat{\mu}_{2r}(x) - \mu_r(x) - \delta_n\bar{\beta}(x)}{\sqrt{\sum_{i=1}^n \bar{w}_{ni}^2(x)\sigma_i^2}} \xrightarrow{d} N(0,1). \qquad (4)$$

Note that the magnitude of the bias, δ_n, is the same for $\hat{\mu}_{2r}(x)$ as for $\hat{m}(v,x)$. However, the magnitude of the asymptotic variance of $\hat{\mu}_{2r}(x)$, which is $\bar{\alpha}_n^2$, can be expected to be of smaller magnitude than α_n^2 [i.e., the asymptotic variance of $\hat{m}(v,x)$] by virtue of the integration. This specially affects the verification of condition 3(i) because we must make the remainder term in Theorem 3 of smaller order than those in Theorem 2. In the next subsection we verify the conditions of Theorem 3 for a kernel estimator that falls in the class defined by (1). The optimal rate will balance $\bar{\alpha}_n$ with δ_n.

6.1.1 Verification of Conditions for Kernels

The Nadaraya-Watson kernel estimator has weights

$$w_{ni}(v,x) = \frac{k\left(\frac{v-V_i}{b}\right)K\left(\frac{x-X_i}{b}\right)}{\sum_{i=1}^n k\left(\frac{v-V_i}{b}\right)K\left(\frac{x-X_i}{b}\right)}$$

in (1), where k is a kernel function and $K(t) = \prod_{j=1}^{d} k(t_j)$. Define $\mu_2(k) = \int t^2 k(t)dt$. Let ∇, ∇^2 denote the first and second derivative operators.

Theorem 4. Suppose that assumptions B1 and B2 in the appendix hold and that the bandwidth sequence $b = b(n)$ satisfies $b \to 0$ and $nb^{d+2}/\log n \to \infty$. Then, (4) holds with $\delta_n = b^2$,

$$\beta(v;x) = \frac{\mu_2(k)}{2} \operatorname{tr}(\nabla^2 m(v;x) + \nabla m(v;x) \nabla s(v;x)), \tag{5}$$

where $s(v;x) = \log f_{V;X}(v;x)$, while

$$nb^d \left(\sum_{i=1} \overline{w}_{ni}^2(x)\sigma_i^2 \right) \xrightarrow{p} \|K\|^2 \int_{\alpha_0(x)}^{\alpha_1(x)} \sigma^2(v;x) \left(\frac{r'(v;x)}{f_{V;X}(v;x)} \right)^2 f_{V;X}(v;x)dv \equiv \Omega(x).$$

Thus $\hat{\mu}_{2r}(x)$ is asymptotically normal with mean $\mu_r(x) + b^2\overline{\beta}(x)$ and variance $n^{-1}b^{-d}\Omega(x)$.

6.2 Semiparametric Estimators

6.2.1 Estimation of β_0

Consider briefly the conditional mean estimator $\mu(X_i;\hat{\beta})$ where

$$\hat{\beta} = \arg\min \frac{1}{n} \sum_{i=1} \left(\frac{Y_i - 1(V_i < 0)}{\hat{h}(V_i \mid X_i)} - \mu(X_i;\beta) \right)^2 \hat{\tau}(Z_i);$$

If $\hat{h}(V_i \mid X_i)$ is replaced with a known design density $h(V_i \mid X_i)$ in the above, then the limiting distribution is given by ordinary weighted nonlinear least squares. The root n limiting distribution of $\hat{\beta}$ using an asymptotically trimmed kernel estimator of $\hat{h}(V_i \mid X_i)$ is given by Lewbel (2000) (for the case where μ is linear). If V is independent of X, then $\hat{h}(V_i \mid X_i)$ can be replaced with

$$\hat{h}(V_i) = \frac{1}{nb} \sum_{j=1} k \left(\frac{V_i - V_j}{b} \right).$$

and the resulting estimator $\hat{\beta}$ is a special case of the general theory covered by Andrews (1994, Theorem 1) and Newey and McFadden (1994, section 8).

6.2.2 Estimation of θ_0

Consider first the estimator $\hat{\theta}$ defined by

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \min_{\alpha} \frac{1}{n} \sum_{i=1} \left(\frac{Y_i - 1(V_i < 0)}{\hat{h}(V_i \mid X_i)} - \alpha - g(X_i;\theta) \right)^2 \hat{\tau}(Z_i);$$

The results of the previous section can be immediately applied taking $\beta = (\alpha;\theta)$ since $\mu(X;\beta) = \alpha - g(X;\theta)$.

To obtain a semiparametrically efficient estimator of θ_0 we apply one-step estimation to Klein and Spady (1993), starting from this initial root-n consistent $\tilde{\theta}$. Define $\Gamma_i(\theta) = E[Y \mid g(X;\theta) + V = g(X_i;\theta) + V_i] = G^*[g(X_i;\theta) + V_i]$, and

$$\hat{\Gamma}_i(\theta) = \frac{\sum_{j=1}^n Y_j k\left(\frac{g(X_i;\theta)+V_i - g(X_j;\theta) - V_j}{b}\right)}{\sum_{j=1}^n k\left(\frac{g(X_i;\theta)+V_i - g(X_j;\theta) - V_j}{b}\right)}$$

for every θ. Let

$$\hat{Q}(\theta) = \frac{1}{n}\sum_{i=1}^n Y_i \ln[\hat{\Gamma}_i(\theta)] + (1 - Y_i)\ln[1 - \hat{\Gamma}_i(\theta)].$$

The semiparametrically efficient Klein and Spady estimator is $\bar{\theta} = \arg\sup_{\theta\in\Theta} \hat{Q}(\theta)$, which satisfies the first order conditions $\hat{Q}'(\bar{\theta}) = 0$. Given the well known problems with computing the Klein and Spady estimator we instead propose the one-step estimator

$$\hat{\theta} = \tilde{\theta} - [\hat{H}(\tilde{\theta})]^{-1}\hat{Q}'(\tilde{\theta}); \tag{6}$$

where

$$\hat{Q}'(\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i}{\hat{\Gamma}_i(\tilde{\theta})} - \frac{1 - Y_i}{1 - \hat{\Gamma}_i(\tilde{\theta})}\right)\hat{\Gamma}_i'(\tilde{\theta})$$

and

$$\hat{H}(\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^n \left(\frac{-1}{\hat{\Gamma}_i(\tilde{\theta})} + \frac{-1}{1 - \hat{\Gamma}_i(\tilde{\theta})}\right)\hat{\Gamma}_i'(\tilde{\theta})\hat{\Gamma}_i'(\tilde{\theta})';$$

and in which $\hat{\Gamma}_i'(\theta) = \partial\hat{\Gamma}_i(\theta)/\partial\theta$. The initial condition $\tilde{\theta}$ is any root-N consistent estimator of θ. Two-step estimation in semiparametric models have been examined in some detail in the monograph Bickel, Klaassen, Ritov and Wellner (1993) and the references therein. Under some regularity conditions, it can be shown that $\hat{\theta}$ has the same asymptotic distribution as $\bar{\theta}$.

6.2.3 Estimation of μ with estimated θ

Here we state the asymptotic properties of the conditional moment estimators based on Corollary 1. Define

$$\hat{\mu}^*_{3r}(x;\hat{\theta}) = r[g(x;\hat{\theta});x] + \frac{1}{n}\sum_{i=1}^n \frac{r'[g(x;\hat{\theta}) - \hat{\upsilon}_i;x][Y_i - 1(\hat{\upsilon}_i > 0)]}{\hat{\psi}(\hat{\upsilon}_i)}$$

$$\hat{\mu}^*_{4r}(x;\hat{\theta}) = r[g(x;\hat{\theta});x] + \frac{1}{n}\sum_{i=1}^n \frac{r'[g(x;\hat{\theta}) - \hat{\upsilon}_i;x][Y_i - 1(\hat{\upsilon}_i > 0)]}{\tilde{\psi}(\hat{\upsilon}_i)};$$

where $\hat{\upsilon}_i = g(X_i;\hat{\theta}) - V_i$ and

$$\hat{\psi}(\hat{\upsilon}_i) = \frac{1}{n}\sum_{j=1}^n h[g(X_j;\hat{\theta}) - \hat{\upsilon}_i] \quad ; \quad \tilde{\psi}(\hat{\upsilon}_i) = \frac{1}{nb}\sum_{j=1}^n k\left(\frac{\hat{\upsilon}_i - \hat{\upsilon}_j}{b}\right).$$

We shall suppose that

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

for some function m that is mean zero and has finite variance. The estimators $\widehat{\mu}_{3r}(x)$ and $\widehat{\mu}_{4r}(x)$ are the special cases of $\widehat{\mu}^*_{3r}(x;\widehat{\theta})$ and $\widehat{\mu}^*_{4r}(x;\widehat{\theta})$ in which θ is known, and so correspond to the case of m being identically zero.

For each $\theta\in\Theta$ and $x\in\mathcal{X}$ let

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

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

$$\Gamma_F = \left(\frac{\partial}{\partial\theta}E\left[f_1(Z_i;\theta)\right]\Big|_{\theta=\theta_0}\right);$$

where $U_i(\theta)=g(X_i;\theta)-V_i$. The quantities f_0, f_1 and Γ_F depend on x but we have suppressed this notationally. Note also that $Ef_1(Z_i;\theta_0)=\mu_r(x)$. Finally, let

$$\gamma_i=\frac{\partial g}{\partial\theta}(X_i;\theta_0)-E\left[\frac{\partial g}{\partial\theta}(X_i;\theta_0)\right]:$$

Theorem 5. Suppose that Assumptions C1-C3 in the Appendix hold. Then, as $n\to\infty$;

$$\sqrt{n}[\widehat{\mu}^*_{3r}(x;\widehat{\theta})-\mu_r(x;\theta_0)]=\frac{1}{\sqrt{n}}\sum_{j=1}\eta_j+o_p(1); \tag{7}$$

where $\eta_j=\eta_{1j}+\eta_{2j}+\eta_{3j}$; with:

$$\eta_{1j} = f_1(Z_j;\theta_0)-Ef_1(Z_j;\theta_0)$$

$$\eta_{2j} = \left(\Gamma_F-E\left[f_0(Z_i;\theta_0)\frac{\psi'(U_i)}{\psi(U_i)}\gamma_i\right]\right)m(Z_j;\theta_0)$$

$$\eta_{3j} = -E\left[f_0(Z_i;\theta_0)\frac{h[g(X_j;\theta_0)-U_i]-\psi(U_i)}{\psi(U_i)}\Big|X_j\right]:$$

This implies that $\sqrt{n}[\widehat{\mu}^*_{3r}(x;\widehat{\theta})-\mu_r(x;\theta_0)]$ is asymptotically normal with mean zero and variance $\Sigma_\eta=\mathrm{var}(\eta_j)$ by the Lindeberg-Feller central limit theorem: 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 result can be extended to a functional central limit theorem in x because the $o_p(1)$ term in (7) is uniform in $x\in\mathcal{X}$ and the stochastic process $\eta_j(x)$ is tight in x due to the smoothness properties of r:

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

$$\gamma_i^* = \frac{\partial g}{\partial \theta'}(X_i;\theta_0) - E\left[\frac{\partial g}{\partial \theta'}(X_i;\theta_0)|U_i\right];$$

where $\phi_u(U_i) = E[\partial g(X_i;\theta_0)/\partial\theta|U_i]$.

Theorem 6. Suppose that assumptions B1,B2 and C1-C4 in the Appendix hold. Then

$$\sqrt{n}[\widehat{\mu}^*_{4r}(x;\widehat{\theta}) - \mu_r(x)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta_i^* + o_p(1);$$

where $\eta_j^* = \eta_{1j}^* + \eta_{2j}^* + \eta_{3j}^*$; with $\eta_{1j}^* = \eta_{1j}$; while

$$\eta_{2j}^* = \left(\Gamma_F - E\left[f_0(Z_i;\theta_0)\left(\frac{\psi'(U_i)}{\psi(U_i)}\gamma_i^* - \phi_u'(U_i)\right)\right]\right) m(Z_j;\theta_0)$$

$$\eta_{3j}^* = -\frac{r'[g(x;\theta_0) - U_j;x]}{\psi(U_j)}(Y_j - E[Y_j|U_j]).$$

This implies that $\sqrt{n}[\widehat{\mu}^*_{4r}(x;\widehat{\theta}) - \mu_r(x;\theta_0)]$ is asymptotically normal with mean zero and variance $\Sigma^*_\eta = \mathrm{var}(\eta_j^*)$. 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.

Remarks

1. Standard errors can be constructed by substituting population quantities by estimated ones. That is, let $\widehat{\Sigma}_\eta = n^{-1}\sum_{j=1}^{n}\widehat{\eta}_j^2$, where $\widehat{\eta}_j = \widehat{\eta}_{1j} + \widehat{\eta}_{2j} + \widehat{\eta}_{3j}$ and for example

$$\widehat{\eta}_{1j} = \frac{r'[g(x;\widehat{\theta}) - \widehat{U}_i;x][Y_i - 1(\widehat{U}_i > 0)]}{\widehat{\psi}(\widehat{U}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{r'[g(x;\widehat{\theta}) - \widehat{U}_i;x][Y_i - 1(\widehat{U}_i > 0)]}{\widehat{\psi}(\widehat{U}_i)}.$$

To construct $\widehat{\eta}_{2j}$ we replace $E[f_0(Z_i;\theta_0)(\psi'(U_i)/\psi(U_i))\gamma_i]$ in η_{2j} by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r'[g(x;\widehat{\theta}) - \widehat{U}_i;x][Y_i - 1(\widehat{U}_i > 0)]}{\widehat{\psi}(\widehat{U}_i)}\frac{\widehat{\psi}'(\widehat{U}_i)}{\widehat{\psi}(\widehat{U}_i)}\left(\frac{\partial g}{\partial \theta}(X_i;\widehat{\theta}) - \frac{1}{n}\sum_{j=1}^{n}\frac{\partial g}{\partial \theta}(X_j;\widehat{\theta})\right).$$

For the quantity Γ_F we must use numerical derivatives, i.e.,

$$\widehat{\Gamma}_F = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{f}_1(Z_i;\widehat{\theta} + \delta e_k) - \widehat{f}_1(Z_i;\widehat{\theta})}{\delta},$$

where e_k is the elementary vector in direction k and δ is a small number, while

$$\widehat{f}_1(Z_i;\widehat{\theta}) = r[g(x;\widehat{\theta});x] + \frac{r'[g(x;\widehat{\theta}) - U_i(\widehat{\theta});x][Y_i - 1(U_i(\widehat{\theta}) > 0)]}{\widehat{\psi}(U_i(\widehat{\theta}))}.$$

For the conditional expectation in η_{3j} we should use a kernel regression smoother on the estimated quantities.

2. Regarding efficiency, it is not possible to provide a ranking of the two estimators $\widehat{\mu}^*_{3r}(x;\widehat{\theta})$ and $\widehat{\mu}^*_{4r}(x;\widehat{\theta})$ uniformly throughout the 'parameter space'. However, one step in that direction might be to use a semiparametrically efficient estimator of θ_0. It may be possible to develop an efficiency bound for estimation of the function $\mu_r(\cdot)$ 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.

6.2.4 Quantile estimators

The distribution theory is trivial. The estimator $\widehat{w}_q(x) = \widehat{G}^{-1}(q \mid x)$ has the distribution theory for standard conditional quantile estimators. The distribution theory for $\widehat{w}_q(x) = g(x;\widehat{\theta}) - \widehat{G}^{*-1}(1-q)$ is the same as the distribution theory for $\widetilde{w}_q(x) = g(x;\theta_0) - \widetilde{G}^{*-1}(1-q)$, where

$$\widetilde{G}^*(u) = \widehat{E}(Y \mid U = u),$$

which is again basically 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.

7 Conclusions

We have provided some estimators of conditional moments and quantiles of the latent W. We have for convenience assumed throughout that the support of V (which must contain the support of W) is bounded. Most of the results in this paper should extend to the infinite support case, although some of the estimators may then require asymptotic trimming to deal with issues arising from division by a density estimate when the true density is not bounded away from zero.

The precision of these estimators depends in part on the density h. When designing experiments one may wish to choose h to maximize efficiency based on the variance estimators.

8 Appendix

8.1 Regularity Conditions

We first state some regularity conditions that are needed for the nonparametric estimation of h.

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 constant c.

Assumption B.2. The variables $(V; X)$ are continuously distributed with Lebesgue density $f_{V;X}(v; x)$ that satisfies $\inf_{\alpha_0(x) \le v \le \alpha_1(x)} f_{V;X}(v; x) > 0$. Furthermore, m and $f_{V;X}$ are twice continuously differentiable for all v with $\alpha_0(x) \le v \le \alpha_1(x)$. The set $[\alpha_0(x); \alpha_1(x)] \times \{x\}$ is strictly contained in the support of $(V; X)$.

We also need some conditions on the estimator and on the regression functions and densities.

Assumption C.1. Suppose that

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1} m(Z_i; \theta_0) + o_p(1)$$

for some function m such that $E[m(Z_i; \theta_0)] = 0$ and $\Omega = E[m(Z_i; \theta_0)m(Z_i; \theta_0)'] < \infty$.

Assumption C.2. The function g is twice continuously differentiable in θ and

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

with $Ed_1(X_i) < \infty$ and $Ed_2(X_i) < \infty$.

Assumption C.3. The density function h is continuous and is strictly positive on its support and is twice 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 $nb^6 \to \infty$.

8.2 Distribution Theory for Nonparametric Estimators

Proof of Theorem 4. Under Assumptions B1 and B2 the expansion (3) holds with $\delta_n = b^2$ and $\beta(v; x)$ is as stated in (5), the weights

$$w_{ni}(v; x) = \frac{1}{f_{V;X}(v; x)} \frac{1}{nb^{d+1}} K\left(\frac{v - V_i}{b}\right) K\left(\frac{x - X_i}{b}\right);$$

while the remainder term satisfies

$$\sup_{\alpha_0(x) \le v \le \alpha_1(x)} |R_n(v; x)| = O_p\left(\frac{\sqrt{\log n}}{nb^{d+1}}\right) + o_p(b^2):$$

See for example Masry (1996a, 1996b).

Provided $nb^{d+2}/\log n \to \infty$; condition (i) of Theorem 3 is satisfied because $\hat{\alpha}_n = O_p(1/\sqrt{nb^d})$ as we now show. We have

$$\left|\overline{w}_{ni}(x) - \frac{-1}{nb^d}K\left(\frac{x-X_i}{b}\right)\frac{r'(V_i;x)}{f_{V;X}(V_i;x)}\right|$$

$$= \left|\frac{1}{nb^d}K\left(\frac{x-X_i}{b}\right)\left[\int_{\alpha_0(x)}^{\alpha_1(x)}\frac{r'(v;x)}{f_{V;X}(v;x)}\frac{1}{b}K\left(\frac{v-V_i}{b}\right)dv - \frac{r'(V_i;x)}{f_{V;X}(V_i;x)}\right]\right|$$

$$= \left|\frac{1}{nb^d}K\left(\frac{x-X_i}{b}\right)\int_{\alpha_0(x)}^{\alpha_1(x)}\left[\frac{r'(v;x)}{f_{V;X}(v;x)} - \frac{r'(V_i;x)}{f_{V;X}(V_i;x)}\right]\frac{1}{b}K\left(\frac{v-V_i}{b}\right)dv\right| \text{ for large } n$$

$$= O(b^2)$$

by a change of variables and dominated convergence argument that is in wide use in nonparametrics (see, e.g., Newey and McFadden 1994 section 8). It works in this case because the set $[\alpha_0(x);\alpha_1(x)]$ is contained in the support of V and the conditions on K etc. Therefore, the asymptotic variance of $\hat{\theta}_{2r}(x)$ is

$$\sum_{i=1}^{n}\overline{w}^2_{ni}(x)\sigma_i^2 \simeq \frac{1}{n^2b^{2d}}\sum_{i=1}^{n}K^2\left(\frac{x-X_i}{b}\right)\left(\frac{r'(V_i;x)}{f_{V;X}(V_i;x)}\right)^2\sigma_i^2 = O_p(n^{-1}b^{-d});$$

as follows from Markov's inequality. Therefore, $\hat{\alpha}_n = O_p(1/\sqrt{nb^d})$ as required. In fact, $\sum_{i=1}^{n}\overline{w}^2_{ni}(x)\sigma_i^2$ satisfies a law of large numbers and is approximately

$$\frac{1}{nb^d}E\left[\frac{1}{b^{d_x}}K^2\left(\frac{x-X_i}{b}\right)\left(\frac{r'(V_i;x)}{f_{V;X}(V_i;x)}\right)^2\sigma^2(V_i;X_i)\right]$$

$$\simeq \frac{1}{nb^d}\|K\|^2\int\sigma^2(v;x)\left(\frac{r'(v;x)}{f_{V;X}(v;x)}\right)^2 f_{V;X}(v;x)dv;$$

where $\sigma^2(V_i;X_i) = \sigma_i^2$; by a change of variables and dominated convergence. Furthermore, condition (ii) of Theorem 3 is satisfied by the arguments used in Gozalo and Linton (1999, Lemma CLT). ■

8.3 Distribution Theory for Semiparametric Quantities

Let E_i denote expectation conditional on Z_i:

Proof of Theorem 5. Recall that

$$\hat{\theta}^{*}_{3r}(x;\hat{\beta}) = r[g(x;\hat{\beta});x] + \frac{1}{n}\sum_{i=1}^{n}\frac{r'[g(x;\hat{\beta}) - \hat{U}_i;x][Y_i - 1(\hat{U}_i > 0)]}{\hat{\lambda}(\hat{U}_i)};$$

where $\hat{U}_i = g(X_i;\hat{\beta}) - V_i$ and

$$\hat{\lambda}(\hat{U}_i) = \frac{1}{n}\sum_{j=1}^{n}h[g(X_j;\hat{\beta}) - \hat{U}_i]:$$

By a geometric series expansion we can write

$$
\begin{aligned}
\hat{r}^*_{3r}(x;\hat\theta) = {}& \frac{1}{n}\sum_{i=1}^n f_1(Z_i;\hat\theta) - \frac{1}{n}\sum_{i=1}^n f_2(Z_i;\theta_0)[\hat\psi(\hat U_i) - \psi(U_i)] \\
& - \frac{1}{n}\sum_{i=1}^n [f_2(Z_i;\hat\theta) - f_2(Z_i;\theta_0)][\hat\psi(\hat U_i) - \psi(U_i)] \\
& + \frac{1}{n}\sum_{i=1}^n \frac{r'[g(x;\hat\theta) - \hat U_i;x][Y_i - 1(\hat U_i > 0)]}{\psi^2(U_i)\hat\psi(\hat U_i)}[\hat\psi(\hat U_i) - \psi(U_i)]^2;
\end{aligned}
$$

where

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

Leading Terms. We make use of Lemmas 1 and 2 given below. Lemma 1 implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n [f_1(Z_i;\hat\theta) - Ef_1(Z_i;\theta_0)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\Gamma_F m(Z_i;\theta_0) + [f_1(Z_i;\theta_0) - Ef_1(Z_i;\theta_0)]\} + o_p(1). \qquad (8)$$

Furthermore, by Lemma 2

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

where $L(Z_i;Z_j) = \xi_j(U_i) + \gamma(Z_i)m(Z_j;\theta_0)$; and

$$\xi_j(u) = h[g(X_j;\theta_0) - u] - E(h[g(X_j;\theta_0) - u])$$

$$\gamma(Z_i) = \psi'(U_i)\left(\frac{\partial g}{\partial\theta}(X_i;\theta_0) - E\left[\frac{\partial g}{\partial\theta}(X_i;\theta_0)\right]\right).$$

Then

$$
\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n f_2(Z_i;\theta_0)L(Z_i;Z_j) &= \sum_{i=1}^n\sum_{j=1}^n \varphi_n(Z_i;Z_j) \\
&= \frac{1}{n}\sum_{j=1}^n \omega(Z_j) + o_p(n^{-1/2});
\end{aligned}
$$

where

$$\omega(Z_j) = E[f_2(Z_i;\theta_0)\gamma(Z_i)]m(Z_j;\theta_0) + E\left[f_2(Z_i;\theta_0)\xi_j(U_i)|Z_j\right]$$

by standard U-statistic theory. We have

$$E\left[f_2(Z_i;\theta_0)\xi_j(U_i)|Z_j\right] = E\left[f_2(Z_i;\theta_0)h[g(X_j;\theta_0)-U_i]|Z_j\right] - E\left[f_2(Z_i;\theta_0)\psi(U_i)\right]$$

$$= E\left[(f_1(Z_i;\theta_0)-r[g(x;\theta_0);x])\frac{h[g(X_j;\theta_0)-U_i]}{\psi(U_i)}|Z_j\right] - E\left[(f_1(Z_i;\theta_0)-r[g(x;\theta_0);x])\right]$$

$$= E\left[\frac{r'[g(x;\theta)-U_i;x][Y_i-1(U_i>0)]}{\psi(U_i)}\frac{h[g(X_j;\theta_0)-U_i]-\psi(U_i)}{\psi(U_i)}|Z_j\right]$$

$$E\left[f_2(Z_i;\theta_0)\Gamma(Z_i)\right] = E\left[f_2(Z_i;\theta_0)\psi(U_i)\frac{\psi'(U_i)}{\psi(U_i)}\left(\frac{\partial g}{\partial\theta}(X_i;\theta_0)-E\left[\frac{\partial g}{\partial\theta}(X_i;\theta_0)\right]\right)\right]$$

$$= E\left[\frac{r'[g(x;\theta_0)-U_i;x][Y_i-1(U_i>0)]}{\psi(U_i)}\frac{\psi'(U_i)}{\psi(U_i)}\gamma_i\right],$$

so that the leading terms are as stated.

Remainders. By the Cauchy-Schwarz inequality

$$\left|\frac{1}{n}\sum_{i=1}^{n}[f_2(Z_i;\hat\theta)-f_2(Z_i;\theta_0)][\hat\psi(\hat U_i)-\psi(U_i)]\right|$$

$$\le \left(\frac{1}{n}\sum_{i=1}^{n}[f_2(Z_i;\hat\theta)-f_2(Z_i;\theta_0)]^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}[\hat\psi(\hat U_i)-\psi(U_i)]^2\right)^{1/2}$$

$$= O_p(n^{-1})$$

from another application of Lemmas 1 and 2.

We have assumed that $\inf_{u\in U}\psi(u)>0$, which implies that

$$\min_{1\le i\le n}\psi(\hat U_i)=\inf_{u\in U}\psi(u)+O_p(n^{-1/2})$$

is bounded away from zero with probability tending to one. Therefore,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{r'[g(x;\hat\theta)-\hat U_i;x][Y_i-1(\hat U_i>0)]}{\psi^2(\hat U_i)\hat\psi(\hat U_i)}[\hat\psi(\hat U_i)-\psi(\hat U_i)]^2\right|$$

$$\le \frac{\sup_{u\in U}[\hat\psi(u)-\psi(u)]^2+O_p(n^{-1/2})}{\inf_{u\in U}\psi^2(u)\hat\psi(u)+O_p(n^{-1/2})}\frac{1}{n}\sum_{i=1}^{n}|r'[g(x;\hat\theta)-\hat U_i;x]|\cdot(|Y_i|+1)$$

$$= O_p(n^{-1}).$$

In conclusion,

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

as required. The asymptotic distribution of $\sqrt{n}[\hat{m}^*_{3r}(x;\hat\theta) - \mu_r(x;\theta_0)]$ follows from the central limit theorem for independent random variables.

■

Proof of Theorem 6. By a geometric series expansion we can write

$$\begin{aligned}
\hat{m}^*_{4r}(x;\hat\theta) &= \frac{1}{n}\sum_{i=1}^n f_1(Z_i;\hat\theta) - \frac{1}{n}\sum_{i=1}^n f_2(Z_i;\theta_0)[\hat\psi(\hat U_i) - \psi(U_i)] \\
&\quad - \frac{1}{n}\sum_{i=1}^n [f_2(Z_i;\hat\theta) - f_2(Z_i;\theta_0)] \times [\hat\psi(\hat U_i) - \psi(U_i)] \\
&\quad + \frac{1}{n}\sum_{i=1}^n \frac{r'[g(x;\hat\theta) - \hat U_i;x][Y_i - 1(\hat U_i > 0)]}{\psi^2(U_i)\hat\psi(\hat U_i)}[\hat\psi(\hat U_i) - \psi(U_i)]^2.
\end{aligned}$$

Leading Terms. We make use of Lemma 3 given below. The term $n^{-1}\sum_{i=1}^n f_1(Z_i;\hat\theta)$ has already been analyzed above. By Lemma 3 we have with probability tending to one

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n f_2(Z_i;\theta_0)\left[\hat\psi(\hat U_i) - \psi(U_i)] - \frac{1}{n}\sum_{j=1}^n L^*(Z_i;Z_j)\right]\right| &\le \left(\frac{1}{nb^3}\frac{1}{n}\sum_{i=1}^n |f_2(Z_i;\theta_0)| d(X_i)\right) \\
&= O_p(n^{-1}b^{-3}) \qquad (9)
\end{aligned}$$

for some function $d(.)$ and random variables $L^*(Z_i;Z_j) = b^{-1}k((U_i - U_j)/b) - \psi(U_i) + \Gamma^*(Z_i)\cdot m(Z_j;\theta_0)$, where

$$\Gamma^*(Z_i) = \psi'(U_i)\left[\frac{\partial g}{\partial\theta'}(X_i;\theta_0) - E\left[\frac{\partial g}{\partial\theta'}(X_i;\theta_0)|U_i\right]\right] - \psi(U_i)\phi'_u(U_i).$$

Under our bandwidth conditions, the right hand side of (9) is $o_p(n^{-1/2})$. Furthermore,

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

where

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

Note that $E_i\varphi_n(Z_i;Z_j) = 0$ but $E_j\varphi_n(Z_i;Z_j) \neq 0$. We write

$$T_{n4} = \frac{1}{n}\sum_{j=1}^n \psi(Z_j) + \sum_{\substack{i=1 \\ i\neq j}}^n\sum_{j=1}^n \omega_n(Z_i;Z_j),$$

where

$$\text{Ã}(Z_j) \;=\; n^2 E_j \,'_n(Z_i;Z_j)$$

$$!_n(Z_i;Z_j) \;=\; '_n(Z_i;Z_j) \;\text{¡}\; E_j \,'_n(Z_i;Z_j);$$

so that $!_n(Z_i;Z_j)$ is a degenerate kernel satisfying $E_i \,!_n(Z_i;Z_j) = E_j \,!_n(Z_i;Z_j) = 0$: We next compute $\text{Ã}_n(Z_j)$; using integration by parts, a change of variable, and dominated convergence we have

$$\text{Ã}(Z_j) = (f_2(Z_j;\text{µ}_0) \;\text{¡}\; E[f_2(Z_j;\text{µ}_0)jU_j])\,\text{Ã}(U_j) + E\,[f_2(Z_i;\text{µ}_0)\,\text{¡}^{\text{¤}}(Z_i)]\,\text{¢}\, m(Z_j;\text{µ}_0) + O_p(b^2):$$

Finally,

$$(f_2(Z_i;\text{µ}_0) \;\text{¡}\; E[f_2(Z_i;\text{µ}_0)jU_i])\,\text{Ã}(U_i) = \frac{r^0[g(x;\text{µ}_0) \;\text{¡}\; U_i;x]}{\text{Ã}(U_i)}(Y_i \;\text{¡}\; E[Y_ijU_i])$$

$$E[f_2(Z_i;\text{µ}_0)\,\text{¡}^{\text{¤}}(Z_i)] = E\;\text{"}\;\frac{r^0[g(x;\text{µ}_0) \;\text{¡}\; U_i;x]}{\text{Ã}(U_i)}[Y_i \;\text{¡}\; 1(U_i > 0)] \;\text{£}\; \text{Ã}\;\frac{\text{Ã}^0(U_i)}{\text{Ã}(U_i)}\,\text{°}^{\text{¤}}_i \;\text{¡}\; \text{Á}^0_u(U_i)\;!\#\;:$$

Remainder Terms. First,

$$\text{–}\frac{1}{n}\sum_{i=1}[f_2(Z_i;\hat{\text{µ}}) \;\text{¡}\; f_2(Z_i;\text{µ}_0)][\hat{\text{Ã}}(\hat{U}_i) \;\text{¡}\; \text{Ã}(U_i)]\text{–}$$
$$\cdot\;\text{Ã}\;\frac{1}{n}\sum_{i=1}[f_2(Z_i;\hat{\text{µ}}) \;\text{¡}\; f_2(Z_i;\text{µ}_0)]^2\;!^{1=2}\;\text{Ã}\;\frac{1}{n}\sum_{i=1}[\hat{\text{Ã}}(\hat{U}_i) \;\text{¡}\; \text{Ã}(U_i)]^2\;!^{1=2}$$
$$= \; o_p(n^{\text{¡}1=2}):$$

Second

$$\text{–}\frac{1}{n}\sum_{i=1}\frac{r^0[g(x;\hat{\text{µ}}) \;\text{¡}\; \hat{U}_i;x][Y_i \;\text{¡}\; 1(\hat{U}_i > 0)]}{\text{Ã}^2(U_i)\hat{\text{Ã}}(\hat{U}_i)}[\hat{\text{Ã}}(\hat{U}_i) \;\text{¡}\; \text{Ã}(U_i)]^2\text{–}$$
$$\cdot\;\frac{\sup_{u2U}[\hat{\text{Ã}}(u) \;\text{¡}\; \text{Ã}(u)]^2(1 + o_p(1))}{\inf_{u2U}\text{Ã}^2(u)\hat{\text{Ã}}(u) + o_p(1)}\frac{1}{n}\sum_{i=1} jr^0[g(x;\hat{\text{µ}}) \;\text{¡}\; \hat{U}_i;x]j\,\text{¢}\,(jY_ij + 1)$$
$$= \; o_p(n^{\text{¡}1=2}):$$

■

8.4 Subsidiary Results

Define

$$F_n(\text{µ}) = \frac{1}{n}\sum_{i=1} f(Z_i;\text{µ})$$

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 m

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1} m(Z_i; \theta_0) + o_p(1)$$

where $E[m(Z_i;\theta_0)] = 0$ and $\Omega = E[m(Z_i;\theta_0)m(Z_i;\theta_0)'] < \infty$.

(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

$$\sqrt{n}[F_n(\hat{\theta}) - F(\theta_0)] \Rightarrow N(0;V),$$

where

$$\begin{aligned} V &= \operatorname{var}[\Gamma_F m(Z_i;\theta_0) + f(Z_i;\theta_0)] \\ &= \Gamma_F \Omega \Gamma_F{}' + \operatorname{var}[f(Z_i;\theta_0)] + 2\Gamma_F E m(Z_i;\theta_0) f(Z_i;\theta_0). \end{aligned}$$

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

Proof: Since $\hat{\theta}$ is root-n consistent, there exists a sequence $\delta_n \to 0$ such that

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

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

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

and the result now follows from standard CLT arguments. ■

Lemma 2. As $n \to \infty$

$$\max_{1\le i\le n}\left|\hat{\psi}(\hat{U}_i)-\psi(U_i)-\frac{1}{n}\sum_{j=1}^{n}L(Z_i;Z_j)\right|=o_p(n^{-1/2}); \tag{10}$$

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

$$\xi_j(u)=h[g(X_j;\theta_0)-u]-E(h[g(X_j;\theta_0)-u])$$

$$\Gamma(Z_i)=\psi'(U_i)\left(\frac{\partial g}{\partial\theta}(X_i;\theta_0)-E\left[\frac{\partial g}{\partial\theta}(X_i;\theta_0)\right]\right):$$

Proof. We have for any $u;$

$$\begin{aligned}\hat{\psi}(u)-\psi(u) &= \frac{1}{n}\sum_{j=1}h[g(X_j;\hat{\theta})-u]-E(h[g(X_j;\theta_0)-u])\\ &= \frac{1}{n}\sum_{j=1}h[g(X_j;\theta_0)-u]-E(h[g(X_j;\theta_0)-u])\\ &\quad+\frac{1}{n}\sum_{j=1}h'[g(X_j;\theta_0)-u]\frac{\partial g}{\partial\theta}(X_j;\theta_0)(\hat{\theta}-\theta_0)+R_n(u);\end{aligned}$$

where

$$\begin{aligned}R_n(u) &= \frac{1}{2n}\sum_{j=1}h''[g(X_j;\bar{\theta})-u](\hat{\theta}-\theta_0)'\frac{\partial g}{\partial\theta}(X_j;\bar{\theta})\frac{\partial g}{\partial\theta'}(X_j;\bar{\theta})(\hat{\theta}-\theta_0)\\ &\quad+\frac{1}{2n}\sum_{j=1}h'[g(X_j;\bar{\theta})-u](\hat{\theta}-\theta_0)'\frac{\partial^2 g}{\partial\theta\partial\theta'}(X_j;\bar{\theta})(\hat{\theta}-\theta_0);\end{aligned}$$

where $\bar{\theta}$ are intermediate values between $\hat{\theta}$ and θ_0. With probability tending to one for a sequence $\delta_n \to 0$ we have by the Cauchy Schwarz inequality

$$\begin{aligned}|R_n(u)| &\le \|\hat{\theta}-\theta_0\|^2\frac{1}{2n}\sum_{j=1}\sup_{\|\theta-\theta_0\|\le\delta_n}|h''[g(X_j;\theta)-u]|\sup_{\|\theta-\theta_0\|\le\delta_n}\left\|\frac{\partial g}{\partial\theta}(X_j;\theta)\right\|\\ &\quad+\|\hat{\theta}-\theta_0\|^2\frac{1}{2n}\sum_{j=1}\sup_{\|\theta-\theta_0\|\le\delta_n}|h'[g(X_j;\theta)-u]|\sup_{\|\theta-\theta_0\|\le\delta_n}\left\|\frac{\partial^2 g}{\partial\theta\partial\theta'}(X_j;\theta)\right\|\\ &\le \|\hat{\theta}-\theta_0\|^2\cdot\left(\sup_t|h''(t)|\frac{1}{2n}\sum_{j=1}d_1(X_j)+\sup_t|h'(t)|\cdot\frac{1}{2n}\sum_{j=1}d_2(X_j)\right) \qquad (11)\\ &= O_p(n^{-1}):\end{aligned}$$

Since the right hand side of (11) does not depend on $u;$ this order is uniform in $u:$ Furthermore because h' is bounded and continuous, by a standard uniform law of large numbers

$$\sup_{u\in U}\left|\frac{1}{n}\sum_{j=1}h'[g(X_j;\theta_0)-u]\frac{\partial g}{\partial\theta}(X_j;\theta_0)-E\left[h'[g(X_j;\theta_0)-u]\frac{\partial g}{\partial\theta}(X_j;\theta_0)\right]\right|=o_p(1);$$

where $\mathcal{U}$ is the support of $U_i = g(X_i;\theta_0) - V_i$. Therefore,

$$\sup_{u\in\mathcal{U}}\left|\hat\psi(u) - \psi(u) - \frac{1}{n}\sum_{j=1}^{n}\xi_j(u) - J(u)(\hat\theta - \theta_0)\right| = o_p(n^{-1/2}); \tag{12}$$

where $\xi_j(u) = h[g(X_j;\theta_0) - u] - E(h[g(X_j;\theta_0) - u])$ are i.i.d. with mean zero and finite variance, and

$$J(u) = E\left[h'[g(X_j;\theta_0) - u]\frac{\partial g}{\partial \theta}(X_j;\theta_0)\right].$$

Because $\sqrt{n}(\hat\theta - \theta_0) = O_p(1)$ the supremum over $u \in \mathcal{U}$ is the same as a maximum over $\hat U_i$.

Furthermore, by a second order Taylor series expansion

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}\xi_j(\hat U_i) + J(\hat U_i)(\hat\theta - \theta_0) &= \frac{1}{n}\sum_{j=1}^{n}\xi_j(U_i) + \frac{1}{n}\sum_{j=1}^{n}\frac{\partial \xi_j}{\partial u}(U_i)\frac{\partial g}{\partial \theta}(X_j;\theta_0)(\hat\theta - \theta_0) \\
&\quad + J(U_i)(\hat\theta - \theta_0) + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{j=1}^{n}\xi_j(U_i) + E_i\left[\frac{\partial \xi_j}{\partial u}(U_i)\frac{\partial g}{\partial \theta}(X_j;\theta_0)\right](\hat\theta - \theta_0) \\
&\quad + J(U_i)(\hat\theta - \theta_0) + o_p(n^{-1/2});
\end{aligned}$$

where

$$\frac{\partial \xi_j}{\partial u}(u) = -h'[g(X_j;\theta_0) - u] + E(h'[g(X_j;\theta_0) - u]);$$

and the error term is bounded in the same way as above using the continuous second derivatives of h, g. That is, $\max_{1\le i\le n}|J(\hat U_i) - J(U_i)| = O_p(1)$ and

$$\sup_{\|\theta - \theta_0\| \le \delta_n}\max_{1\le i\le n}\left|\frac{\partial^2 \xi_j}{\partial u^2}(U_i(\theta))\right| \le d(Z_i)$$

with $Ed(Z_i) < \infty$. Note that

$$\begin{aligned}
J(U_i) + E_i\left[\frac{\partial \xi_j}{\partial u}(U_i)\frac{\partial g}{\partial \theta}(X_j;\theta_0)\right] &= E_i\left[h'[g(X_j;\theta_0) - U_i]\frac{\partial g}{\partial \theta}(X_j;\theta_0)\right] \\
&\quad + E_i\left[\{-h'[g(X_j;\theta_0) - U_i] + E_i(h'[g(X_j;\theta_0) - U_i])\}\frac{\partial g}{\partial \theta}(X_j;\theta_0)\right] \\
&= E_i(h'[g(X_j;\theta_0) - U_i])\,E\left(\frac{\partial g}{\partial \theta}(X_j;\theta_0)\right);
\end{aligned}$$

so that

$$\hat\psi(\hat U_i) - \psi(\hat U_i) = \frac{1}{n}\sum_{j=1}^{n}\xi_j(U_i) + \left(E_i[h'[g(X_j;\theta_0) - U_i]]\cdot E\left[\frac{\partial g}{\partial \theta}(X_j;\theta_0)\right]\right)\cdot(\hat\theta - \theta_0) + o_p(n^{-1/2}). \tag{13}$$

Finally,

$$\psi(\hat U_i) - \psi(U_i) = -E_i\left(h'[g(X_j;\theta_0) - U_i]\right)\cdot\frac{\partial g}{\partial\theta}(X_i;\theta_0)\cdot(\hat\theta-\theta_0) + o_p(n^{-1/2}). \qquad (14)$$

These results are uniform under some additional conditions. Combining (13) and (14) we obtain the result (10).

■

Lemma 3. We have with probability tending to one

$$\left|\hat\psi(\hat U_i) - \psi(U_i) - \frac{1}{n}\sum_{j=1}^{n} L^*(Z_i;Z_j)\right| \le \frac{k}{nb^3}d(X_i)$$

for some function d, where

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

$$\Gamma^*(Z_i) = \psi'(U_i)\left[\frac{\partial g}{\partial\theta'}(X_i;\theta_0) - E\left[\frac{\partial g}{\partial\theta'}(X_i;\theta_0)|U_i\right]\right] - \psi(U_i)\varphi_u'(U_i).$$

Proof. Making a second order Taylor series expansion we have $\hat\psi(\hat U_i) - \psi(U_i) = T_{ni} + R_{ni}$, where

$$T_{ni} = \bar\psi(U_i) - \psi(U_i) + \frac{1}{nb^2}\sum_{j=1}^{n} k'\left(\frac{U_i-U_j}{b}\right)\left[\frac{\partial g}{\partial\theta'}(X_i;\theta_0) - \frac{\partial g}{\partial\theta'}(X_j;\theta_0)\right](\hat\theta-\theta_0)$$

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

where θ^* are intermediate values between $\hat\theta$ and θ_0, and $U_i^* = U_i(\theta^*)$. We have with probability tending to one

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

by the Cauchy-Schwarz inequality. By the uniform convergence results of Masry (1996a,1996b):

$$\begin{aligned}\max_{1\le i\le n}\frac{1}{nb}\sum_{j=1}^{n}\left|k'\left(\frac{U_i-U_j}{b}\right)\right| &= O_p(1) \\ \max_{1\le i\le n}\frac{1}{nb}\sum_{j=1}^{n}\left|k'\left(\frac{U_i-U_j}{b}\right)\right|d_2(X_j) &= O_p(1);\end{aligned}$$

so that 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, if the functions d_j are bounded this translates into a uniform result

$$\max_{1\leq i\leq n} |R_{ni}| = O_p(n^{-1}b^{-3}).$$

Provided $nb^6 \to \infty$; this term is $o_p(n^{-1/2})$. With additional smoothness conditions on k this condition can be substantially weakened.

Furthermore, by Masry (1996a, 1996b)

$$\max_{1\leq i\leq n}\left|\frac{1}{nb^2}\sum_{j=1} k'\left(\frac{U_i - U_j}{b}\right) - E\left[\frac{1}{b^2}k'\left(\frac{U_i - U_j}{b}\right)|U_i\right]\right| = O_p\left(\sqrt{\frac{\log n}{nb^3}}\right)$$

$$\max_{1\leq i\leq n}\left|\frac{1}{nb^2}\sum_{j=1} k'\left(\frac{U_i - U_j}{b}\right)\frac{\partial g}{\partial \theta'}(X_j;\theta_0) - E\left[\frac{1}{b^2}k'\left(\frac{U_i - U_j}{b}\right)\phi_u(U_j)|U_i\right]\right| = O_p\left(\sqrt{\frac{\log n}{nb^3}}\right).$$

Also,

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

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. Therefore,

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

Finally, we have

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

by standard results for kernel estimates. ■

8.4.1 Stochastic Equicontinuity Results

We now show that condition (iii) of Lemma 1 is satisfied. Let $£_n(c) = fµ: {}^{p}\overline{n}\,\overline{µ ¡ µ^{0}}\, · cg$: Since ${}^{p}\overline{n}(\hat{µ} ¡ µ^{0}) = O_p(1)$; for all $²$ > 0 there exists a $c_²$ and an integer n_0 such that for all $n ¸ n_0$; $Pr[\hat{µ} 2 £_n(c_²)] ¸ 1 ¡ ²$: Define the stochastic process

$$º_n(µ) = \frac{1}{p\overline{n}} \sum_{i=1} f(Z_i; µ) ¡ E[f(Z_i; µ)]; \quad µ 2 £;$$

where

$$f(Z_i; µ) = r[g(x; µ); x] + \frac{r^{0}[g(x; µ) ¡ U_i(µ); x][Y_i ¡ 1(U_i(µ) > 0)]}{\tilde{A}(U_i)}$$

and define the pseudo-metric

$$½(µ; µ^{0}) = E \left[[f(Z_i; µ) ¡ f(Z_i; µ^{0})]^{2} \right];$$

on $£$: Under this metric, the parameter space $£$ is totally bounded. We are only interested in the behaviour of this process as $µ$ varies in the small set $£_n$: By writing $µ = µ^{0} + ° n^{¡1=2}$; we shall make a reparameterization to $º_n(°)$; where $° 2 ¡(c) ½ R^{p}$: We establish the following result:

$$\sup_{° 2 ¡} jº_n(°) ¡ º_n(0)j = o_p(1) \qquad (15)$$

To prove (15) it is sufficient to show a pointwise law of large numbers, e.g., $º_n(°) ¡ º_n(0) = o_p(1)$ for any $° 2 ¡$; and stochastic equicontinuity of the process $º_n$ at $° = 0$. 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 $º_n(°)$ is the same for all $° 2 ¡$ by the smoothness of the expected value in $°$. To complete the proof of (15) we shall use the following lemma, proved below, which states that $º_n$ is stochastically equicontinuous in $µ$. The difficulty in establishing the required equicontinuity arises solely because the function g inside U is nonlinear in $µ$:

Lemma SE. Under the above assumptions, the process $º_n(°)$ is stochastically equicontinuous, i.e., for all $²$ > 0 and $´$ > 0; there exists $±$ > 0 such that

$$\limsup_{n ! 1} Pr \left[\sup_{½(t_1; t_2) < ±} jº_n(t_1) ¡ º_n(t_2)j > ´ \right] < ²:$$

Proof of Lemma SE. By a second order Taylor series expansion of $g(Z_i; µ)$ around $g(Z_i; µ^{0})$:

$$g(Z_i; µ^{0} + ° n^{¡1=2}) = g(Z_i; µ^{0}) + \frac{1}{p\overline{n}} \sum_{k=1} \frac{@g}{@µ_k}(Z_i; µ^{0}) °_k + \frac{1}{n} \sum_{k=1} \sum_{r=1} \frac{@^{2}g}{@µ_k @µ_r}(Z_i; \overline{µ}) °_k °_r \qquad (16)$$

for some intermediate points $\bar{\mu}$: Define the linear approximation to $g(Z_i;\mu^0 + \circ n^{¡1=2})$;

$$T(Z_i;\circ) = g(Z_i;\mu^0) + \mathop{\mathsf{X}}_{k=1} \frac{@g}{@\mu_k}(Z_i;\mu^0)\circ_k$$

for any $\circ$: By assumption C2, for all $k;r;$ $\sup_{\mu 2 \pounds} j@^2 g(Z_i;\mu)=@\mu_k@\mu_r j^2 \cdot d(Z_i)$ with $Ed(Z_i) < 1$: Therefore, for all $\pm > 0$ there exists an $" > 0$ such that

$$\Pr\text{"}\frac{1}{p\!=\!n}\max_{i;k;r}\sup_{\mu 2 \pounds_n}\left|\frac{@^2 g}{@\mu_k@\mu_r}(Z_i;\mu)\right| > \text{"}\ \# \cdot n\mathop{\mathsf{X}}_{k;r}\Pr\text{"}\frac{1}{p\!=\!n}\sup_{\mu 2 \pounds_n}\left|\frac{@^2 g}{@\mu_k@\mu_r}(Z_i;\mu)\right| > \text{"}\ \#$$

$$\cdot \frac{\mathbf{P}_{k;r}E[d(Z_i)]}{\text{"}^2}$$

$$\cdot \pm$$

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

$$\max_{1\cdot i\cdot n}\left|\frac{1}{n}\mathop{\mathsf{X}}_{k=1}\mathop{\mathsf{X}}_{r=1}\frac{@^2 g}{@\mu_k@\mu_r}(Z_i;\bar{\mu})\circ_k\circ_r\right| \cdot \frac{\bar{¼}}{p\!=\!n}$$

for some $\bar{¼} < 1$: Define the stochastic process

$$\circ_{n1}(\circ;¼) = \frac{1}{p\!=\!n}\mathop{\mathsf{X}}_{i=1}\bar{f}(Z_i;\mu_0 + \circ n^{¡1=2};¼n^{¡1=2})\ ¡\ E\bar{f}(Z_i;\mu_0 + \circ n^{¡1=2};¼n^{¡1=2})$$

on $\circ$ 2 ¡ and $¼$ 2 ¦ = $[0;\bar{¼}]$; where

$$\bar{f}(Z_i;\mu_0 + \circ n^{¡1=2};¼n^{¡1=2})$$

$$= r[g(x;\mu_0 + \circ n^{¡1=2});x] + \frac{r^0[g(x;\mu_0 + \circ n^{¡1=2})\ ¡\ U_i(\mu_0 + \circ n^{¡1=2});x]}{\tilde{A}(U_i)}[Y_i\ ¡\ 1(T(Z_i;\circ n^{¡1=2}) + \frac{¼}{p\!=\!n} > 0)]$$

It suffices to show that $\circ_{n1}(\circ;¼)$ is stochastically equicontinuous in $\circ;¼$; and the deterministic centering term is of smaller order. The latter argument is a standard Taylor expansion. The argument for $\circ_{n1}(\circ;¼)$ is very similar to that contained in Sherman (1993) because we basically have a linear index structure in this part. One can apply Lemma 2.12 in Pakes and Pollard (1989). ■

References

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

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

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

[4] 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.

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

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

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

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

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

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

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

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

[13] 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.

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

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

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

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

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