TRICK OR TREAT? ASYMPTOTIC EXPANSIONS FOR SOME SEMIPARAMETRIC PROGRAM EVALUATION ESTIMATORS*

Hidehiko Ichimura
University College, London

Oliver Linton†
London School of Economics

July 20, 2001

Abstract

We investigate the performance of a class of semiparametric estimators of the treatment effect via asymptotic expansions. We derive approximations to the first three moments of the estimator that are valid to 'second order'. We use these approximations to define a method of bandwidth selection. We also propose a degrees of freedom like bias correction that improves the second order properties of the estimator but without requiring estimation of higher order derivatives of the unknown propensity score. We provide some numerical calibrations of the results.

Journal of Economic Literature Classification: C14

Keywords and phrases: Bandwidth Selection; Kernel Estimation; Semiparametric Estimation; Treatment Effect.

Preliminary and Incomplete

*We are grateful to the National Science Foundation and the Economic and Social Science Research Council for financial support.

†Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Tel. 0207 955-7864; Fax. 0207 831–1840; E-mail address: lintono@lse.ac.uk

1 Introduction

In a series of classic papers Tom [Rothenberg (1984,1985,1986)] introduced Edgeworth expansions to a broad audience. His treatment of the generalized least squares estimator (1984) in particular was immensely important because it dealt with an estimator of central important and the analysis was both deep and precise, but comprehensible. This is in contrast with some of the more frenzied publications about Edgeworth expansions that had hitherto appeared in econometrics journals. The use of Basu's theorem in that paper to establish the independence of the correction terms from the leading term is well known. The review paper (1984) was also very influential and highly cited.

It is our purpose here to present asymptotic expansions for a class of semiparametric estimators used in the treatment effects literature. We have argued elsewhere, Linton (1991,1995), that the first-order asymptotics of semiparametric procedures can be misleading and unhelpful. The limiting variance matrix of the semiparametric procedure Σ does not depend on the specific details of how the nonparametric function estimator $\widehat{g}$ is constructed, and thus sheds no light on how to implement this important part of the procedure. Specifically, bandwidth choice cannot be addressed by using the first-order theory alone. Also, the relative merits of alternative first-order equivalent implementations, e.g., one-step procedures, cannot be determined by the first-order theory alone. Finally, to show when bootstrap methods can provide asymptotic refinements for asymptotically pivotal statistics requires some knowledge of higher-order properties – see Horowitz (1995). This motivates the study of higher-order expansions. Carroll & Härdle (1989) was to our knowledge the first published paper that developed second-order mean squared error expansions for a semiparametric, i.e., smoothing-based but root-n consistent, procedure, in the context of a heteroskedastic linear regression. Härdle, Hart, Marron, & Tsybakov (1992) developed expansions for scalar average derivatives which was extended to the multivariate case, actually only the simpler situation of density-weighted average derivatives, by Härdle & Tsybakov (1993); these papers used the expansions to develop automatic bandwidth selection routines. This work was extended to the slightly more general case of density-weighted averages by Powell & Stoker (1996). In my PhD thesis [Linton (1991)] I developed expansions for a variety of semiparametric regression models including the partially linear model and the heteroskedastic linear regression model; some of this work was later published in Linton (1995, 1996a). The Linton (1995) paper work also provided some results on the optimality of the bandwidth selection procedures proposed therein. Xiao & Phillips (1996) worked out the same approximations for a time series regression model with serial correlation of unknown form; Xiao & Linton (1997) give the analysis for Bickel's (1982) adaptive estimator in the linear regression model; Linton & Xiao (1997) works out the approximations for the nonlinear least squares and profile likelihood estimators in a semiparametric binary choice model. Nishiyama & Robinson (2000) proved the validity of an

Edgeworth approximation to the distribution of the density weighted average derivative estimator. Linton (2000) derived an Edgeworth approximation to the distribution of the standardized estimator and a Wald statistic in a semiparametric instrumental variables model.

In this paper, we develop asymptotic expansions for an estimator of the treatment effect recently proposed in Hirano, Imbens, & Ridder (2000), henceforth HIR. Propensity Score matching is a nonexperimental method for estimating the average effect of social programs.[1] The method compares average outcomes of participants and nonparticipants conditioning on the propensity score value. When averaged over the propensity score, the average measures the average impact of a program if the conditioning on the observable variables makes the choice of the program conditionally mean independent from the potential outcomes. This methodology has received much attention recently in econometrics. While the method used often in practice uses the nearest match in either regressors or estimated propensity score to compare the treatment and the comparison groups, the asymptotic distribution theory for these methods have not been developed. The asymptotic distribution theory has been developed by Heckman, Ichimura & Todd (1998) for the kernel based matching method. HIR considers reweighting estimator that estimates the treatment effect as well. Both methods require choosing smoothing parameters but optimal methods to choosing the smoothing parameter have not been discussed. In this paper we consider optimal bandwidth selection for the reweighting estimator.

2 The Model and Estimator

We investigate a class of estimators for the treatment effect, studied by HIR. For each individual we observe $Z_i = (y_i, t_i, X_i)$, where $y_i = y_{1i} \cdot t_i + y_{0i} \cdot (1 - t_i)$ and

$$t_i = \begin{cases} 1 & \text{if treated} \\ 0 & \text{if untreated,} \end{cases}$$

while y_{1i} and y_{0i} are potential outcome for each individual i with and without the treatment and X_i is a vector of covariates. Actually, for convenience we will take X to be a scalar and to have a continuous density f bounded away from zero on its compact support. We will also assume that y_i possesses many finite moments. Define the propensity score

$$p(x) = \Pr[t_i = 1 | X_i = x] = E(t_i | X_i = x),$$

and marginal regressions $m_1(x) = E[y_{1i}|X_i = x]$ and $m_0(x) = E[y_{0i}|X_i = x]$.

[1] See Cochran (1968), Rosenbaum & Rubin (1983), and Heckman, Ichimura, & Todd (1998).

Identifying assumptions of the estimator are:

$$\begin{aligned} E[y_1|X_i, t_i &= 1] = E[y_1|X_i, t_i = 0] \\ E[y_0|X_i, t_i &= 1] = E[y_0|X_i, t_i = 0] \\ 0 &< p(X_i) < 1 \end{aligned}$$

with probability one in X_i. Clearly under these assumptions $E[y_{1i}|X_i, t_i = 1] = m_1(X_i)$ and $E[y_{0i}|X_i, t_i = 0] = m_0(X_i)$. Under these identifying assumptions,

$$\begin{aligned} g_1(x) &\equiv E[y_i \cdot t_i|X_i = x] = m_1(x) \cdot p(x), \text{ and} \\ g_0(x) &\equiv E[y_i \cdot (1 - t_i)|X_i = x] = m_0(x) \cdot (1 - p(x)). \end{aligned}$$

The average treatment effect parameter, θ_0, is defined thusly

$$\begin{aligned} \theta_0 &= E(y_{1i}) - E(y_{0i}) = E[m_1(X_i)] - E[m_0(X_i)] \\ &= E\left[\frac{g_1(X_i)}{p(X_i)}\right] - E\left[\frac{g_0(X_i)}{1 - p(X_i)}\right] = E\left[\frac{E(y_i \cdot t_i|X_i)}{p(X_i)}\right] - E\left[\frac{E(y_i \cdot (1 - t_i)|X_i)}{1 - p(X_i)}\right]. \end{aligned}$$

We consider the estimator $\widehat{\theta}$ of θ_0 that solves

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(Z_i, \theta, \widehat{p}(X_i)) = O_p(n^{-1}), \tag{1}$$

where

$$\Psi(Z_i, \theta, \widehat{p}(X_i)) = \frac{y_i \cdot t_i}{\widehat{p}(X_i)} - \frac{y_i \cdot (1 - t_i)}{1 - \widehat{p}(X_i)} - \theta \tag{2}$$

and $\widehat{p}(X_i) = \sum_{j=1}^{n} w_{ij} t_j$, where w_{ij} are smoothing weights that only depend on the covariates $X_1, \ldots, X_n$. HIR used series estimates. The bias correction method we propose below can also be applied to series estimates and indeed any linear smoother, but discussion of smoothing bias terms requires that we use kernel or local polynomial estimators. We will also adopt the leave-one-out paradigm that is used in many semiparametric estimates. To be specific then we shall take the following weights:

$$w_{ij} = \begin{cases} \dfrac{K\left(\frac{X_i - X_j}{h}\right)}{\sum_{\substack{l=1 \\ l \neq i}}^{n} K\left(\frac{X_i - X_l}{h}\right)} & j \neq i \\ \\ 0 & j = i, \end{cases}$$

where K is a probability density function symmetric about zero with support $[-1, 1]$, while $h = h(n)$ is a positive bandwidth sequence. We have taken the fixed bandwidth leave-one-out Nadaraya-Watson

kernel smoother as our estimator of the regression function. This is rather the Morris Minor of the smoothing world and possesses some not so attractive properties [Fan and Gijbels (1996)]. However, it is very convenient to work with technically.

3 Main Results

Define the standardized estimator $T = \sqrt{n}(\widehat{\theta} - \theta_0)$. We derive a stochastic expansion for T by Taylor expanding $\Psi(Z_i, \theta, \widehat{p}(X_i))$ around $\Psi(Z_i, \theta, p(X_i))$, thus obtaining a representation for T in terms of $\widehat{p}(X_i) - p(X_i)$ and the derivatives of Ψ with respect to p, which we denote by Ψ_p, Ψ_{pp} etc. We thereby obtain

$$T = T^* + R, \tag{3}$$

where T^* contains the leading terms of the expansion, and R is a remainder term that is $o_p(n^{-\alpha})$ in probability for some $\alpha > 0$. The magnitude $o_p(n^{-\alpha})$ is determined to ensure that our results in Theorems 1 and 2 below are sensible. The random variable T^* has finite moments to various orders and indeed it is a polynomial function of certain U-statistics. We shall calculate the moments of T^* and interpret them as if they were the moments of T. This methodology has a long tradition of application in econometrics following Nagar (1959). When $\sup_n E[T^2] < \infty$, we might reasonably expect that $E[T^2] = E[T^{*2}] + o(n^{-\alpha})$, but see Srinavasan (1970) for a cautionary tale in this regard. Unfortunately, in our case T does not necessarily have uniformly bounded moments. In this case, some additional justification for examining the moments of the truncated statistic must be given. With some additional work and regularity conditions it is possible to establish the stronger regularity that T and T^* have the same distribution to order $n^{-\alpha}$, which requires some restrictions on the tails of R, see the discussion in Rothenberg (1984). In this case our moment approximations can be interpreted as the moments of the approximating distribution.

HIR showed that $\sqrt{n}(\widehat{\theta} - \theta_0)$ is asymptotically normal with finite variance

$$v_0 = E\left[\left(\Psi(Z_i; \theta_0, p(X_i)) + s_p(X_i)\varepsilon_i\right)^2\right], \tag{4}$$

where $\varepsilon_j = t_j - p(X_j)$ and

$$s_p(x) = E[\Psi_p(Z_i; \theta_0, p(X_i))|X_i = x] = -\left[\frac{g_1(x)}{p^2(x)} + \frac{g_0(x)}{(1-p(x))^2}\right].$$

They also established that this estimator is semiparametrically efficient, i.e., it has the smallest asymptotic variance amongst the class of all feasible estimators.

We investigate the next order terms in the expansion of $\widehat{\theta}$. The two largest second order terms in T^* are both biases and are

$$O_p(h^2\sqrt{n}) + O_p(n^{-1/2}h^{-1}). \tag{5}$$

There are also mean zero random variables of order h^2 and order $n^{-1/2}h^{-1/2}$. However, according to the criterion of mean squared error, these stochastic terms are dominated by the bias terms, and the optimal thing to do is to minimize the size of (5) by choosing h appropriately. The optimal bandwidth is of order $h \asymp n^{-1/3}$ in which case both terms in (5) are the same magnitude, and indeed of order $n^{-1/6}$. Thus, the second order terms are very large and are mostly bias related. This suggests that the usual asymptotic approximation may not be very well located. We shall next assume that a bandwidth of the optimal order $h \asymp n^{-1/3}$ has been chosen so as to simplify the discussion of the results. Define the function

$$\beta(x) = \frac{(p \cdot f)''(x) - p(x)f''(x)}{f(x)}$$

and let

$$s_{pp}(x) = E[\Psi_{pp}(Z_i; \theta_0, p(X_i))|X_i = x] = \left[\frac{g_1(x)}{p(x)^3} - \frac{g_0(x)}{(1-p(x))^3}\right].$$

THEOREM 1. *Under some regularity conditions, as $n \to \infty$, $R = o_p(n^{-1/3})$ in (3) and:*

$$\begin{aligned} E(T^*) &\simeq \sqrt{n}h^2 b_1 + \frac{1}{\sqrt{n}h} b_2 + o(n^{-1/3}) \\ \text{var}(T^*) &\simeq v_0 + O(h^2) + O(n^{-1}h^{-1}) + o(n^{-1/3}), \end{aligned}$$

where

$$b_1 = \mu_2(K) E\left[s_p(X_i)\beta(X_i)\right] \quad ; \quad b_2 = \|K\|^2 E\left[s_{pp}(X_i)\frac{p(X_i)(1-p(X_i))}{2f(X_i)}\right].$$

and $\mu_2(K) = \int u^2 K(u)du/2$ and $\|K\|^2 = \int K(u)^2 du$.

The smoothing bias term b_1 can take either sign, since it depends on the covariance between the smoothing bias quantity $\beta(X)$ and the conditional expectation $s_p(X)$. The term b_2 can also take either sign depending on the sign of $s_{pp}(x)$. The correction term in the variance is clearly of smaller order than the squared bias no matter what bandwidth is chosen. If we define optimal bandwidth h_{opt} as one that minimizes the asymptotic mean squared error of the estimator, then the above result indicates that it suffices to minimize the size of the bias. Note that if the biases have opposite signs then the optimal bandwidth is going to set $\sqrt{n}h^2 b_1 + \frac{1}{\sqrt{n}h} b_2 = 0$ and this second order bias will then

be of smaller order. Otherwise, the optimal bandwidth will minimize this second order bias and there will be an interior solution to the optimization problem that can be found by calculus. Therefore,

$$h_{opt} = \begin{cases} \left(\frac{-b_2}{b_1}\right)^{1/3} n^{-1/3} & \text{if } \operatorname{sign}(b_2) \neq \operatorname{sign}(b_1) \\ \\ \left(\frac{b_2}{2b_1}\right)^{1/3} n^{-1/3} & \text{if } \operatorname{sign}(b_2) = \operatorname{sign}(b_1). \end{cases}$$

In some semiparametric estimators it has been shown that by using leave-one-out estimators and other devices one can eliminate the degrees of freedom bias terms of order $n^{-1/2}h^{-1}$, see for example Linton (1995). Indeed, we have used leave-one-out estimator here. Unfortunately, it has not completely eliminated the degrees of freedom bias. Instead, we define an explicit bias correction method and show that it does indeed 'knock' this term out. Specifically, we define the bias-corrected estimator

$$\widehat{\theta}^{bc} = \widehat{\theta} - \widehat{b}, \tag{6}$$

where

$$\widehat{b} = \frac{1}{n^3h^2} j \neq i \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\frac{y_i \cdot t_i}{\widehat{p}(X_i)^3} - \frac{y_i \cdot (1-t_i)}{[1-\widehat{p}(X_i)]^3} \right] \frac{1}{\widehat{f}^2(X_i)} K^2 \left(\frac{X_i - X_j}{h} \right) \widehat{\varepsilon}_j^2,$$

where $\widehat{\varepsilon}_j = t_j - \widehat{p}(X_j)$. This bias correction is similar conceptually to using $n-1$ instead of n in estimating a population variance; significantly, in this context we do not need to estimate higher derivatives of the unknown functions, and it follows that the sampling properties of this bias estimator should be relatively good. Can also do this multiplicatively?

The stochastic expansion for $\widehat{\theta}^{bc}$ is the same as that for $\widehat{\theta}$ except for the additional bias correcting term $\widehat{b}$. On computing the moments of the leading terms of this expansion however we find that the bias term b_2 has been eliminated; we therefore end up with a better trade-off in the mean squared error of this estimator. The largest terms are a squared bias of order h^4n and a variance of order $n^{-1}h^{-1}$. This trade-off leads to an optimal bandwidth $h \propto n^{-2/5}$ and mean squared error of $n^{-3/5}$. Let

$$\zeta_i = \Psi_p(Z_i; \theta_0, p(X_i)) - E[\Psi_p(Z_i; \theta_0, p(X_i))|X_i].$$

Let now $T = \sqrt{n}(\widehat{\theta}^{bc} - \theta_0)$ and obtain $T = T^* + R$ as in (3).

THEOREM 2. *Under some regularity conditions, as $n \to \infty$, $R = o_p(n^{-3/5})$ in (3) and:*

$$\begin{aligned} E(T^*) &\simeq \sqrt{n}h^2 b_1 + o(n^{-3/5}) \\ \operatorname{var}(T^*) &\simeq v_0 + \frac{1}{nh} v_1 + o(n^{-3/5}), \end{aligned}$$

where

$$\begin{aligned} v_1 &= \|K\|^2 \times \left\{ E\left[\frac{E(\varepsilon_j^2|X_j)E(\zeta_j^2|X_j)}{f(X_j)}\right] + E\left[\frac{E^2(\varepsilon_j\zeta_j|X_j)}{f(X_j)}\right]\right\} \\ &\quad + \|K*K\|^2 \times E\left[\frac{s_{pp}^2(X_j)E^2(\varepsilon_j^2|X_j)}{4f(X_j)}\right] \\ &\quad + 2\langle K, K*K\rangle \times E\left[\frac{s_{pp}(X_j)E(\varepsilon_j^2|X_j)E(\varepsilon_j\zeta_j|X_j)}{f(X_j)}\right], \end{aligned}$$

where $(K*K)(t) = \int K(t)K(t-u)du$ *is the convolution of* K *with itself and* $\langle f, g\rangle = \int f(t)g(t)dt$.

This shows that the bias correction can lead to improved mean squared error properties.[2] The optimal bandwidth is now

$$h_{opt} = \left(\frac{v_1}{4b_1^2}\right)^{1/5} n^{-2/5},$$

since b_1^2, v_1 are both non-negative. This bandwidth is smaller in magnitude than is optimal for the raw estimator $\widehat{\theta}$.

We have just presented results concerning the moments of the estimators, but this can also be extended to distributional approximations. In fact, to the relevant order $\widehat{\theta}$ is normally distributed, i.e.,

$$\Pr\left[\sqrt{n}(\widehat{\theta} - \theta_0) \leq x\right] = \Phi\left(\frac{x - \sqrt{n}h^2 b_1 + \frac{1}{\sqrt{n}h}b_2}{v_0}\right) + o(n^{-1/3}).$$

The approximation for $\sqrt{n}(\widehat{\theta}^{bc} - \theta_0)$ is more complicated because if we require an error rate consistent with our mean squared error [i.e., of order $n^{-3/5}$] then we will have to include the skewness terms of order $n^{-1/2}$;[3] in this case the approximate distribution is not normal in general but can be expressed in terms of the Edgeworth signed measures and the first three cumulant approximations. See Linton (2000) for a computation of this type.

Finally, we remark that the standard errors also depend on $\widehat{p}(\cdot)$ and there are similar concerns about the small sample properties of these quantities. These standard errors also suffer from a

[2] We are happy to report that this finding is in agreement with Rothenberg (1984, p909)

> *"This suggests that correction for bias may be more important than second order efficiency consideration when choosing among estimators"*

[3] In fact,

$$E[\{T^* - E(T^*)\}^3] \simeq O(n^{-1/2}).$$

degrees of freedom bias problem, which can be corrected in the same way as we have done for the estimator of θ.

4 Some Numerical Results

For comparison we present the optimal rates associated with a variety of semiparametric models that have been studied before. These are all for the univariate case with second order kernels or similar method.

TABLE 1

Rates of Convergence for Bandwidth and Mean Squared Error Correction

Model	Optimal Bandwidth	Optimal MSE Correction
1. Average Derivative	$n^{-2/7}$	$n^{-1/7}$
2. Variance Estimation	$n^{-1/5}$	$n^{-3/5}$
3. Partially Linear Model	$n^{-2/9}$	$n^{-7/9}$
4. Heteroskedastic Linear Regression	$n^{-1/5}$	$n^{-4/5}$
5. Variance a Function of Mean	$n^{-2/11}$	$n^{-5/11}$
6. Symmetric Location	$n^{-1/7}$	$n^{-4/7}$
7. HIR	$n^{-1/3}$	$n^{-1/3}$
8. HIR with Bias Correction	$n^{-2/5}$	$n^{-3/5}$

Notes. Models 2-6 are given in Linton (1991, Chapter 3). The result for Model 1 is taken from Härdle, Hart, Marron, & Tsybakov (1992).

The optimal bandwidth for nonparametric regression is of order $n^{-1/5}$ and has a consequent MSE of order $n^{-4/5}$. Table 1 shows that there is quite a variety of magnitudes for the optimal bandwidth in semiparametric estimation problems; sometimes the optimal bandwidth is bigger but usually it is smaller than the optimal rates for nonparametric estimation. These different rates reflect different magnitudes for bias and variance in these semiparametric functionals.

We next investigate the magnitudes of the second order effects in Theorems 1 and 2 and the optimal bandwidth size.

Design 1

$$\begin{aligned} X &\sim U[-0.5, 0.5] \\ m_0(x) &= \alpha_0 + \alpha_1 x \\ m_1(x) &= \tau + m_0(x) \\ y_0 &= m_0(x) + \eta,\ \eta \sim N(0, \sigma_\eta^2) \\ y_1 &= y_0 + \tau \\ t &= 1\left(\beta_0 + \beta_1 x + \delta > 0\right),\ \delta \sim N(0, \sigma_\delta^2). \end{aligned}$$

We find that

5 Conclusions and some not so deep thoughts

Our asymptotic expansions revealed some important facts about the HIR estimator. The main thing is that its properties are dominated by bias: one bias term is related to the curvature of the function p and the covariate density f [smoothing bias], and the second bias term is what we have called a degrees of freedom bias. The magnitude of the bias terms can be quite large and their signs are unknown in general. We proposed a simple bias correction that eliminates the degrees of freedom bias term, thereby permitting a smaller bandwidth and consequently a better mean squared error correction.

6 Appendix

We derive the stochastic expansion for the more general case where the estimator is only implicitly defined and where there may be multiple functional components. Therefore, we adopt a slightly different notation. Suppose that we observe data $\{Z_i\}_{i=1}^n$ partitioned as $Z_i = (Y_i, X_i)$, where the dependent variables are Y_i and the regressors are X_i. Suppose for simplicity that θ_0 is a scalar unknown parameter. We define our estimator $\widehat{\theta}$ to be any approximate zero of an estimated moment condition, i.e., any sequence that satisfies

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(Z_i; \widehat{\theta}, \widehat{g}(X_i)) = O_p(n^{-1}), \tag{7}$$

where Ψ is a given score function possessing various regularity conditions drawn on below, while $\widehat{g}$ is a nonparametric estimate of the unknown function g, where g is a regression function, hazard function, density function or similar object. This is a standard class of semiparametric estimators.

We will use subscripts to denote derivatives, so that $\Psi_\theta(z,\theta,g)$ is the derivative of $\Psi(z,\theta,g)$ with respect to θ and Ψ_g is the derivative of $\Psi(z,\theta,g)$ with respect to the scalar g. We will suppose that it has already been shown that $\widehat{\theta}-\theta_0 = O_p(n^{-1/2})$, and that the derivatives of Ψ satisfy the following:

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Z_i;\theta_0,\widehat{g}(X_i)) &= O_p(1) \quad ; \quad \frac{1}{n}\sum_{i=1}^{n}\Psi_\theta(Z_i;\theta_0,\widehat{g}(X_i)) = O_p(1) \quad ; \\
\frac{1}{n}\sum_{i=1}^{n}\Psi_{\theta\theta}(Z_i;\theta_0,\widehat{g}(X_i)) &= O_p(1)\sup_{|\theta-\theta_0|\leq c/\sqrt{n}}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi_{\theta\theta\theta}(Z_i;\theta,\widehat{g}(X_i))\right| = O_p(1).
\end{aligned}$$

Sufficient conditions for these results can be found in numerous places for a variety of estimators $\widehat{g}$ and functions Ψ. See for example Andrews (1994), Newey & McFadden (1994), Bickel, Klaassen, Ritov, & Wellner (1993) etc. Linton (1996) develops higher order asymptotic expansions for similar semiparametric estimators for a specific class of nonparametric estimators. We requires some further assumptions regarding the properties of the nonparametric estimator and on the derivatives of Ψ with respect to g. We will assume at least that we have a uniform rate of convergence on $\widehat{g}$. Specifically, we suppose that

$$||\widehat{g}-g||_\infty = \sup_{x\in int(\mathcal{X})}|\widehat{g}(x)-g(x)| = o_p(n^{-1/4}). \tag{8}$$

Sufficient conditions for this can be found in Masry (1996). We also suppose that

$$\sup_{|t-g(X_i)|\leq\epsilon}|\Psi_{ggggg}(Z_i;\theta_0,t)|\,,\ \sup_{|t-g(X_i)|\leq\epsilon}|\Psi_{\theta ggg}(Z_i;\theta_0,t)| \leq d(Z_i)$$

with $Ed(Z_i)<\infty$.

Let $A = n^{-1/2}\sum_{i=1}^{n}\Psi(Z_i;\theta_0,\widehat{g}(X_i))$, $B = n^{-1}\sum_{i=1}^{n}\Psi_\theta(Z_i;\theta_0,\widehat{g}(X_i))$, and $C = n^{-1}\sum_{i=1}^{n}\Psi_{\theta\theta}(Z_i;\theta_0,\widehat{g}(X_i))$. Then by a Taylor series expansion it follows that

$$0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Z_i;\widehat{\theta},\widehat{g}(X_i)) = A + B\sqrt{n}(\widehat{\theta}-\theta_0) + \frac{1}{2\sqrt{n}}C[\sqrt{n}(\widehat{\theta}-\theta_0)]^2 + O_p(n^{-1}).$$

Using standard techniques [Bhattacharya & Ghosh (1978)] we can invert this expansion to obtain

$$\sqrt{n}(\widehat{\theta}-\theta_0) = -B^{-1}A - \frac{B^{-3}CA^2}{2\sqrt{n}} + O_p(n^{-1}).$$

This expresses the standardized estimator as a simple function of sample averages of the parameter derivatives of the moment condition, i.e., $\sqrt{n}(\widehat{\theta}-\theta_0) \simeq r(A,B,C)$ for the given function r.

By a Taylor expansion in g, we have:

$$\begin{aligned} A &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Z_i;\theta_0,g(X_i)) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_g(Z_i;\theta_0,g(X_i))(\widehat{g}(X_i)-g(X_i)) \\ &+\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\Psi_{gg}(Z_i;\theta_0,g(X_i))(\widehat{g}(X_i)-g(X_i))^2 + \frac{1}{6\sqrt{n}}\sum_{i=1}^{n}\Psi_{ggg}(Z_i;\theta_0,g(X_i))(\widehat{g}(X_i)-g(X_i))^3 \\ &+\frac{1}{24\sqrt{n}}\sum_{i=1}^{n}\Psi_{gggg}(Z_i;\theta_0,g(X_i))(\widehat{g}(X_i)-g(X_i))^4 + o_p(n^{-3/4}) \\ &\equiv \sum_{j=0}^{4} A_j + o_p(n^{-3/4}), \end{aligned} \tag{9}$$

$$\begin{aligned} B &= E\Psi_\theta(Z_i;\theta_0,g(X_i)) + \frac{1}{n}\sum_{i=1}^{n}\Psi_\theta(Z_i;\theta_0,g(X_i)) - E\Psi_\theta(Z_i;\theta_0,g(X_i)) \\ &+\frac{1}{n}\sum_{i=1}^{n}\Psi_{\theta g}(Z_i;\theta_0,g(X_i))(\widehat{g}(X_i)-g(X_i)) + \frac{1}{2n}\sum_{i=1}^{n}\Psi_{\theta gg}(Z_i;\theta_0,g(X_i))(\widehat{g}(X_i)-g(X_i))^2 + o_p(n^{-3/4}) \\ &\equiv \sum_{j=0}^{3} B_j + o_p(n^{-3/4}). \end{aligned} \tag{10}$$

Similarly,

$$C = C_0 + o_p(n^{-1/4}), \tag{11}$$

where $C_0 = E\Psi_{\theta\theta}(Z_i;\theta_0,g(X_i))$. It follows from the uniform rate on $\widehat{g}$ that $A_2 = o_p(1)$, $A_3 = o_p(n^{-1/4})$, $A_4 = o_p(n^{-1/2})$, $B_2 = o_p(n^{-1/4})$ and $B_3 = o_p(n^{-1/2})$. By standard arguments $B_0 = O(1)$, $A_0 = O_p(1)$, and $B_1 = O_p(n^{-1/2})$. There is also a well known argument that shows

$$A_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_g(Z_i;\theta_0,g(X_i))(\widehat{g}(X_i)-g(X_i)) = O_p(1) + \text{bias term}.$$

See for example Newey (1995). Provided the bias term is made small, $A_1 = O_p(1)$. We should also point out that in some cases some of these terms disappear. For example, it can happen that $E[\Psi_g(Z_i;\theta_0,g(X_i))|X_i] = 0$, in which case A_1 is of smaller order. Other terms can also drop out as we will see.

We will now combine the information we have acquired about the individual terms in the expansion to simplify it further. We have shown that

$$T = T^* + o_p(n^{-3/4}),$$

where

$$T^* = -B_0^{-1}\left(\sum_{j=0}^{4} A_j\right) + B_0^{-2}\left(\sum_{j=1}^{3} B_j\right)\left(\sum_{j=0}^{4} A_j\right) - B_0^{-3}A_0\left(\sum_{j=1}^{3} B_j\right)^2 - \frac{B_0^{-3}C_0A_0^2}{\sqrt{n}}. \tag{12}$$

In fact, some of these terms are clearly redundant like $B_0^{-2}B_3\left(\sum_{j=3}^{4} A_j\right)$ etc. Furthermore, we will typically only be interested in retaining terms that contribute most to the mean squared error, which would entail dropping many further terms.

We now turn to the specific case of this paper, in which $\Psi_\theta(Z_i, \theta, \widehat{p}(X_i)) \equiv -1$ and so $\Psi_{\theta\theta}(Z_i, \theta, \widehat{p}(X_i))$ etc. and $\Psi_{\theta p}(Z_i, \theta, \widehat{p}(X_i))$ etc. are all identically zero, i.e., the expansion terminates. Therefore, the full expansion in our case is

$$\begin{aligned}
\sqrt{n}(\widehat{\theta} - \theta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Z_i;\theta_0,p(X_i)) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi_p(Z_i;\theta_0,p(X_i))(\widehat{p}(X_i)-p(X_i)) \\
&+\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\Psi_{pp}(Z_i;\theta_0,p(X_i))(\widehat{p}(X_i)-p(X_i))^2 \\
&+\frac{1}{6\sqrt{n}}\sum_{i=1}^{n}\Psi_{ppp}(Z_i;\theta_0,p(X_i))(\widehat{p}(X_i)-p(X_i))^3 \\
&+\frac{1}{24\sqrt{n}}\sum_{i=1}^{n}\Psi_{pppp}(Z_i;\theta_0,p(X_i))(\widehat{p}(X_i)-p(X_i))^4 + o_p(n^{-3/4}).
\end{aligned}$$

Define

$$\xi_i = \Psi_{pp}(Z_i;\theta_0,p(X_i)) - E[\Psi_{pp}(Z_i;\theta_0,p(X_i))|X_i],$$

which are i.i.d. error terms that are conditional mean zero given X_j. We have

$$\begin{aligned}
\Psi_p(Z_i;\theta_0,p(X_i)) &= -\frac{y_i\cdot t_i}{p(X_i)^2} - \frac{y_i\cdot(1-t_i)}{[1-p(X_i)]^2} \\
\Psi_{pp}(Z_i;\theta_0,p(X_i)) &= 2\frac{y_i\cdot t_i}{p(X_i)^3} - 2\frac{y_i\cdot(1-t_i)}{[1-p(X_i)]^3}.
\end{aligned}$$

Then, we can write

$$\begin{aligned}
\sqrt{n}(\widehat{\theta} - \theta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Z_i;\theta_0,p(X_i)) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_p(X_i)(\widehat{p}(X_i)-p(X_i)) \\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_i\cdot(\widehat{p}(X_i)-p(X_i)) + \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}s_{pp}(X_i)(\widehat{p}(X_i)-p(X_i))^2 \\
&+\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\xi_i\cdot(\widehat{p}(X_i)-p(X_i))^2 + o_p(n^{-1/2}).
\end{aligned}$$

We use the decomposition

$$\widehat{p}(X_i) - p(X_i) = \sum_{j \neq i} w_{ij}\varepsilon_j + \beta_n(X_i),$$

where w_{ij} are the smoothing weights that just depend on the covariates $X_1, \ldots, X_n$, while

$$\beta_n(X_i) = E[\widehat{p}(X_i)|X_1, \ldots, X_n] - p(X_i)$$

is the conditional smoothing bias that also just depends on the covariates $X_1, \ldots, X_n$.

We then write

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n s_p(X_i)(\widehat{p}(X_i) - p(X_i)) \\ = \; & \frac{1}{\sqrt{n}} \sum_{j=1}^n s_p(X_j)\varepsilon_j + \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j \left[\sum_{i \neq j} w_{ij} s_p(X_i) - s_p(X_j) \right] + \frac{1}{\sqrt{n}} \sum_{i=1}^n s_p(X_i)\beta_n(X_i), \end{aligned} \tag{13}$$

where the first term is $O_p(1)$ and jointly asymptotically normal with the leading term $\frac{1}{\sqrt{n}} \sum_{i=1}^n \Psi(Z_i; \theta_0, p(X_i))$, the second term is mean zero and has variance of the same magnitude as $E[\sum_{i \neq j} w_{ij} s_p(X_i) - s_p(X_j)]^2$, this we expect to be $O(h^4)$. In fact, we have

$$\begin{aligned} \sum_{i \neq j} w_{ij} s_p(X_i) - s_p(X_j) \; & \simeq \; \frac{1}{nh} \sum_{i \neq j} K\left(\frac{X_i - X_j}{h} \right) \frac{s_p(X_i)}{E\widehat{f}(X_i)} - s_p(X_j) \\ & \simeq \; \int K\left(\frac{X - X_j}{h} \right) s_p(X) \frac{f(X)}{E\widehat{f}(X)} dX - s_p(X_j) \\ & = \; O(h^2). \end{aligned}$$

The third term in (13) is a bias term with magnitude $h^2\sqrt{n}$ and variance also h^4.

We next turn to the term

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_i \cdot (\widehat{p}(X_i) - p(X_i)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_i \sum_{j \neq i} w_{ij}\varepsilon_j + \frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_i \beta_n(X_i),$$

where the first term is a second order degenerate U-statistic that has mean zero and variance of order $n^{-1}h^{-1}$; it is also uncorrelated with the leading term. The second term is mean zero and $O_p(h^2)$.

We next turn our attention to the term

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{pp}(X_i)(\widehat{p}(X_i)-p(X_i))^2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{pp}(X_i)(\sum_{j\neq i} w_{ij}\varepsilon_j + O_p(h^2))^2 \\ &\simeq \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{pp}(X_i)\sum_{j\neq i} w_{ij}^2 E(\varepsilon_j^2|X_j) \\ &\quad +\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{pp}(X_i)\sum_{j\neq i} w_{ij}^2[\varepsilon_j^2 - E(\varepsilon_j^2|X_j)] \\ &\quad +\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{pp}(X_i) j\neq l \sum_{j\neq i}\sum_{l\neq i} w_{ij}w_{il}\varepsilon_j\varepsilon_l.\end{aligned}$$

The first term is not mean zero and is of order $n^{-1/2}h^{-1}$ in probability and is the dominant term; the second term is mean zero and of order $n^{-1}h^{-1}$ in probability. The third term is mean zero and actually $O_p(n^{-1/2}h^{-1/2})$. We can rewrite this term

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{pp}(X_i) j\neq l \sum_{j\neq i}\sum_{l\neq i} w_{ij}w_{il}\varepsilon_j\varepsilon_l &= j\neq l\sum\sum\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{pp}(X_i)w_{ij}w_{il}\right)\varepsilon_j\varepsilon_l \\ &\simeq \frac{1}{n\sqrt{n}h} j\neq l\sum\sum (K*K)\left(\frac{X_j-X_l}{h}\right)\frac{s_{pp}(X_j)}{f(X_j)}\varepsilon_j\varepsilon_l.\end{aligned}$$

Finally, it is easy to see that

$$\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\xi_i\cdot(\widehat{p}(X_i)-p(X_i))^2 = O_p(h^4+n^{-1}h^{-1}).$$

Specifically, we can suppose without loss of generality that ξ_i is independent of ζ_i and so this term is mean zero and has variance

$$\frac{1}{4n}\sum_{i=1}^{n} E(\xi_i^2)\cdot E[(\widehat{p}(X_i)-p(X_i))^4],$$

which has the order as stated.

Let

$$\begin{aligned}M_n(X_i) &= E[(\widehat{p}(X_i)-p(X_i))^2|X_1,\ldots,X_n] \\ &\simeq \frac{1}{nh}||K||^2\frac{p(X_i)(1-p(X_i))}{f(X_i)} + \frac{h^4}{4}\mu_2^2(K)\beta^2(X_i).\end{aligned}$$

In conclusion we have

$$\begin{aligned}\sqrt{n}(\widehat{\theta}-\theta_0) \;\simeq\; & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Z_i;\theta_0,p(X_i))+s_p(X_i)\varepsilon_i \quad [=O_p(1)] \\ & +\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j\left[\sum_{i\neq j}w_{ij}s_p(X_i)-s_p(X_j)\right] \quad \left[=O_p(h^2)\right] \\ & +i\neq j\sum\sum\varphi_n(Z_i,Z_j) \quad \left[=O_p(n^{-1/2}h^{-1/2})\right] \\ & +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_p(X_i)\beta_n(X_i) \quad \left[=O_p(h^2\sqrt{n})\right] \\ & +\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}s_{pp}(X_i)M_n(X_i) \quad \left[=O_p(h^4\sqrt{n})+O_p(n^{-1/2}h^{-1})\right],\end{aligned}$$

where

$$\varphi_n(Z_i,Z_j)=\frac{1}{nh\sqrt{n}}\frac{1}{f(X_i)}\left[K\left(\frac{X_i-X_j}{h}\right)\zeta_i\varepsilon_j+\frac{1}{2}(K*K)\left(\frac{X_i-X_j}{h}\right)s_{pp}(X_i)\varepsilon_i\varepsilon_j\right].$$

Clearly, $E[\varphi_n(Z_i,Z_j)|Z_i]=E[\varphi_n(Z_i,Z_j)|Z_j]=0$. The first three lines contains mean zero and indeed asymptotically normal terms, while the fourth and fifth lines contain non-mean zero biases. ■

Define the quantity

$$\widetilde{b}=\frac{1}{n^3h^2}j\neq i\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{y_i\cdot t_i}{p(X_i)^3}-\frac{y_i\cdot t_i}{[1-p(X_i)]^3}\right]\frac{1}{f^2(X_i)}K^2\left(\frac{X_i-X_j}{h}\right)\varepsilon_j^2$$

whose expectation $E(\widetilde{b})$ is approximately equal to b_2/nh. Then it can be shown that

$$\frac{\widehat{b}-E(\widetilde{b})}{E(\widetilde{b})}=O_p(\sqrt{\frac{\log n}{nh}}+h^2).$$

This means that

$$\sqrt{n}(\widehat{\theta}^{bc}-\theta_0)=\sqrt{n}(\widehat{\theta}-\theta_0)-\frac{b_2}{\sqrt{nh}}(1+O_p(\sqrt{\frac{\log n}{nh}}+h^2)).$$

■

REFERENCES

Bhattacharya, R. N. and J. K. Ghosh (1978): "On the Validity of the Formal Edgeworth Expansion," Annals of Statistics 6, 434–451.

Carroll, R.J., and W. Härdle, (1989): "Second Order Effects in Semiparametric Weighted Least Squares Regression." Statistics, 2, 179-186.

Cochran, W.G. (1968) "The effectiveness of adjustment by subclassification in removing bias in observational studies." Biometrics, 24, 295–313.

Fan, J., and I. Gijbels (1996): Local Polynomial Modelling and Its Applications Chapman and Hall.

Härdle, W., J. Hart, J. S. Marron, and A. B. Tsybakov (1992): "Bandwidth Choice for Average Derivative Estimation," Journal of the American Statistical Association, 87, 218-226.

Härdle, W., and A. B. Tsybakov (1993): "How sensitive are Average Derivatives," Journal of Econometrics, 58, 31-48.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998) "Characterization of Selection Bias Using Experimental Data" Econometrica, 66, 1017–1098.

Heckman, J., H. Ichimura, and P. Todd (1998) "Matching as an Econometric Estimator" Review of Economic Studies, 65, 261–294.

Hirano, K., G. Imbens, G. Ridder, (2000), "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," NBER Technical Working Paper 251.

Hsieh, D.A., and C.F. Manski (1987): "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression." Annals of Statistics, 15, 541-551.

Linton, O.B. (1991): "Edgeworth Approximation in Semiparametric Regression Models," PhD Thesis, Department of Economics, University of California at Berkeley.

Linton, O.B. (1995): "Second Order Approximation in the Partially Linear Regression Model," Econometrica 63, 1079-1112.

Linton, O.B. (1996a): "Second order approximation in a linear regression with heteroskedasticity of unknown form," Econometric Reviews 15, 1-32.

Linton, O.B. (1996b): "Edgeworth Approximation for MINPIN Estimators in Semiparametric Regression Models." Econometric Theory 12, 30-60.

Linton, O.B. (1997): "Second-Order approximation for semiparametric instrumental variable estimators and test statistics." Cowles Foundation Discussion Paper no 1151. Forthcoming in Journal of Econometrics.

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

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

Nagar, A.L. (1959): "The bias and moment matrix of the general k-class estimator of the parameters in simultaneous equations," Econometrica 27, 575-595.

Nishiyama, Y., and Robinson, P. M. (2000): "Edgeworth expansions for semiparametric averaged derivatives," Econometrica 68, 931-980.

Robinson, P. M. (1995): "The normal approximation for semiparametric averaged derivatives," Econometrica 63, 667-680.

Rosenbaum, P. and Rubin, D.B. (1983) "The central role of the propensity score in observational studies for causal effects." Biometrika, 70, pp. 41-55.

Rothenberg, T., (1984a): "Approximating the Distributions of Econometric Estimators and Test Statistics." Ch.14 in: Handbook of Econometrics, vol 2, ed. Z. Griliches and M.Intriligator. North Holland.

Rothenberg, T., (1984b): "Approximate Normality of Generalized Least Squares Estimates." Econometrica, 52, 811-825.

Rothenberg, T., (1984c): "Hypothesis Testing in Linear Models when the Error Covariance Matrix is Nonscalar." Econometrica, 52, 827-842.

Rothenberg, T., (1988): "Approximate Power Functions for some Robust Tests of Regression Coefficients." Econometrica, 56, 997-1019.

Xiao, Z. and O.B. Linton (1997): "Second order approximation for an adaptive estimator in a linear regression." Forthcoming in Econometric Theory.

Xiao, Z. and P.C.B. Phillips (1996): "Higher order approximation for a frequency domain regression estimator," Journal of Econometrics, 86, 297-336.