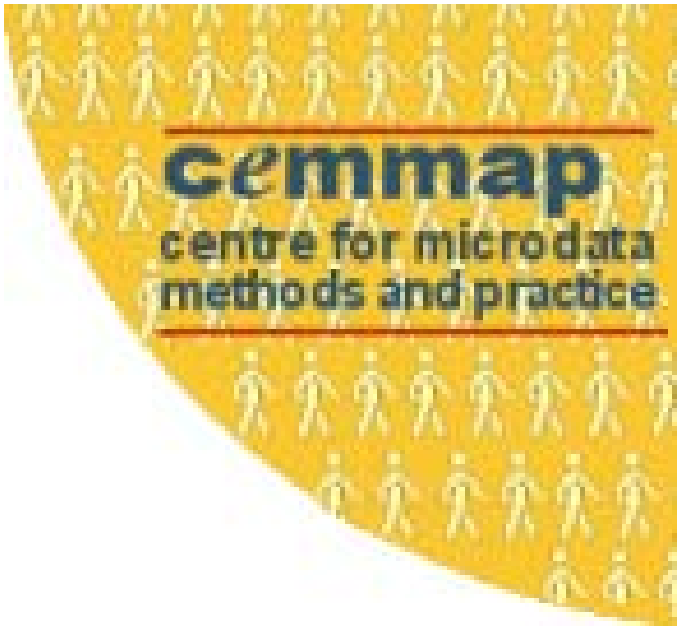

ASYMPTOTIC EXPANSIONS FOR SOME SEMIPARAMETRIC PROGRAM EVALUATION ESTIMATORS

Hidehiko Ichimura
Oliver Linton

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP04/01

ASYMPTOTIC EXPANSIONS FOR SOME SEMIPARAMETRIC PROGRAM EVALUATION ESTIMATORS*

Hidehiko Ichimura
University College, London

Oliver Linton†
London School of Economics

September 12, 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 two 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; Program Evaluation; Semiparametric Estimation; Treatment Effect.

1 Introduction

In a series of classic papers Tom [Rothenberg (1984abc,1988)] introduced Edgeworth expansions to a broad audience. His treatment of the generalized least squares estimator (1984b) in particular

*We would like to thank Tom Rothenberg and seminar participants for helpful comments and David Jacho-Chavez for research assistance. 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

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 a well-known example of his elegant work. The review paper (1984a) 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) and Heckman, Ichimura, Smith and Todd (1998), 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 the second author's PhD thesis[Linton (1991)], written under Tom's supervision, 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 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 (2001) 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 (1997) 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. Let y_{1i} and y_{0i} denote potential outcome for each individual i with and without 'the treatment'. 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 X_i is a vector of covariates or pre-treatment variables. 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. We are interested in the average treatment effect parameter

$$\tau_0 = E(y_{1i}) - E(y_{0i}).$$

We shall assume the following identifying conditions:

$$E[y_1|X_i, t_i = 1] = E[y_1|X_i, t_i = 0]$$

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

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

with probability one in X_i, where

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

is the propensity score. The first two assumptions are that treatment and potential outcome are mean independent given covariates; the final assumption is that there are at least some unobserved influences on the probability of receiving the treatment. See Rosenbaum and Rubin (1983) and Heckman et al. (1998). Clearly under these assumptions $E[y_{1i}|X_i, t_i = 1] = E[y_{1i}|X_i = x] = m_1(X_i)$ and $E[y_{0i}|X_i, t_i = 0] = E[y_{0i}|X_i = x] = m_0(X_i)$. Furthermore, the following observable regressions are related to the unobservable regressions:

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

It now follows that the average treatment effect parameter τ_0 satisfies

$$\begin{aligned} \tau_0 &= E(y_{1i}) - E(y_{0i}) = E[m_1(X_i)] - E[m_0(X_i)] \\ &= E\left[\frac{g_1(X_i)}{p(X_i)}\right] - E\left[\frac{g_0(X_i)}{1 - p(X_i)}\right] = E\left[\frac{E(y_i \cdot t_i|X_i)}{p(X_i)}\right] - E\left[\frac{E(y_i \cdot (1 - t_i)\,|X_i)}{1 - p(X_i)}\right] \\ &= E\left[\frac{y_i \cdot t_i}{p(X_i)}\right] - E\left[\frac{y_i \cdot (1 - t_i)}{1 - p(X_i)}\right], \end{aligned}$$

where the last line follows from the law of iterated expectations. The last line is the relation that HIR use to suggest an estimator. The HIR estimator is

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i \cdot t_i}{\widehat{p}(X_i)} - \frac{y_i \cdot (1 - t_i)}{1 - \widehat{p}(X_i)}\right],$$

where $\widehat{p}(X_i)$ was a nonparametric estimate of $p(X_i)$, in fact they chose series estimates.

We allow a slightly greater degree of generality; in particular, we consider the estimator $\widehat{\tau}$ of τ_0 to be any sequence that solves

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

where

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

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 $X_1, \ldots, X_n$. As we mentioned earlier, HIR used series estimates. The bias correction method we propose below can also be applied to series estimates and indeed to 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 we let the parameter vector $(\widehat{\alpha}_0(X_i), \widehat{\alpha}_1(X_i))$ minimize the criterion function

$$\sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right) \{t_j - \alpha_0 + \alpha_1(X_j - X_i)\}^2, \tag{3}$$

with respect to (α_0, α_1), where K is a differentiable probability density function symmetric about zero with support $[-1, 1]$, while $h = h(n)$ is a positive bandwidth sequence. Then let $\widehat{p}(X_i) = \widehat{\alpha}_0(X_i)$ and let w_{ij} be the corresponding smoothing weights. We have taken the fixed bandwidth leave-one-out local linear kernel smoother as our estimator of the regression function. This estimator is preferable to the local constant kernel estimator because of its superior bias properties both at interior and boundary regions, see Fan and Gijbels (1996).

3 Main Results

Define the standardized estimator $T = \sqrt{n}(\widehat{\tau} - \tau_0)$. HIR showed that

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \rho_i + o_p(1) = T_0 + o_p(1), \tag{4}$$

where $\rho_i = \Psi(Z_i; \tau_0, p(X_i)) + s_p(X_i)\varepsilon_i$, where $\varepsilon_j = t_j - p(X_j)$ and

$$s_p(x) = E[\Psi_p(Z_i; \tau_0, p(X_i))|X_i = x] = -\left[\frac{m_1(x)}{p(x)} + \frac{m_0(x)}{1 - p(x)}\right].$$

Here, the derivatives of Ψ with respect to p are denoted by Ψ_p, Ψ_{pp} etc. Therefore, T is asymptotically normal with finite variance

$$v_0 = E\left[\left(\Psi(Z_i; \tau_0, p(X_i)) + s_p(X_i)\varepsilon_i\right)^2\right]. \tag{5}$$

In fact, they rewrote the asymptotic variance in the more interpretable form

$$v_0 = \operatorname{var}\left[E\left(y_{1i} - y_{0i}|X_i\right)\right] + E\left[\frac{\operatorname{var}(y_{1i}|X_i)}{p(X_i)}\right] + E\left[\frac{\operatorname{var}(y_{0i}|X_i)}{1 - p(X_i)}\right].$$

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 are interested in the higher order properties of their estimator. We derive a stochastic expansion for T by Taylor expanding $\Psi(Z_i, \tau, \widehat{p}(X_i))$ around $\Psi(Z_i, \tau, p(X_i))$, thereby obtaining the representation

$$T = T_0 + T_1 + R = T^* + R, \tag{6}$$

where the leading term T_0 is as defined in (4), T_1 contains the 'second order' terms, while R is a remainder term that is of smaller order in probability. To be specific, we show that $R = o_p(n^{-\alpha})$ in probability for some $\alpha > 0$ that is determined by the order of magnitude of the bandwidth and of course by how many terms in the Taylor expansion we retain. 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 linear combination 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).[2] The two largest [in probability] second order terms in T^* are both non-zero mean terms and are

$$O_p(h^2\sqrt{n}) + O_p(n^{-1/2}h^{-1}). \tag{7}$$

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 (7) by choosing h appropriately. The optimal bandwidth is therefore of order $h \asymp n^{-1/3}$, in which case both terms in (7) are the same magnitude, and indeed are both 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 functions

$$\beta(x) = p''(x)$$

$$s_{pp}(x) = E[\Psi_{pp}(Z_i; \tau_0, p(X_i))|X_i = x] = 2\left[\frac{m_1(x)}{p(x)^2} - \frac{m_0(x)}{(1-p(x))^2}\right]$$

[2]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 (1984a). In this case our moment approximations can be interpreted as the moments of the approximating distribution.

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

THEOREM 1. *Under some regularity conditions, as $n \to \infty$, $R = o_p(n^{-1/3})$ in (6) and:*

$$\begin{aligned} E(T^*) &\simeq \sqrt{n}h^2 b_{n1} + \frac{1}{\sqrt{n}h} b_2 + o(n^{-1/3}) \\ \text{var}(T^*) &\simeq v_0 + o(n^{-1/3}), \end{aligned}$$

where b_{n1} is deterministic and satisfies $b_{n1} \to b_1$ with

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

The leading 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 $s_p(X)$. When p is a standard normal c.d.f., $p''(x) < 0$ for all x and the bias function is negative. The term b_2 can also take either sign depending on the sign of $s_{pp}(x)$. Suppose there is a constant treatment effect τ independent of X, that $p(x) = 1/2$ for all x, and that f is uniform with range one. Then $b_2 = \|K\|^2 \times \tau$, and the sign of b_2 is determined by the direction of the treatment effect. The correction term in the variance is clearly of smaller order than the squared bias no matter what bandwidth is chosen.

Define the asymptotic mean squared error of the estimator to be [apart from a factor of order n^{-1}]

$$AMSE(\widehat{\tau}) = E(T^{*2}) = \text{var}(T^*) + E^2(T^*), \tag{8}$$

and define an optimal bandwidth h_{opt} to be a sequence that minimizes $AMSE(\widehat{\tau})$. By Theorem 1,

$$AMSE(\widehat{\tau}) = v_0 + \left(\sqrt{n}h^2 b_1 + \frac{1}{\sqrt{n}h} b_2\right)^2 + o(n^{-1/3})$$

and it suffices to minimize the size of the term inside the brackets. 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. To summarize, we have

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

A feasible bandwidth selection method can be defined based on estimates of the quantities b_j, $j = 1, 2$, either nonparametric estimates or parametric estimates suggested from some sort of Silverman's rule of thumb idea.[3]

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 Hall and Marron (1987) and Linton (1995). Indeed, we have used a 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 and therefore permits a smaller bandwidth and a better AMSE. Specifically, we define the bias-corrected estimator

$$\widehat{\tau}^{bc} = \widehat{\tau} - \widehat{b}, \tag{9}$$

where

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

where $\widehat{\varepsilon}_j = t_j - \widehat{p}(X_j)$. Note that the way we have defined the bias correction can be applied to any linear smoother with weights w_{ij}. 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.

The stochastic expansion for $\widehat{\tau}^{bc}$ is the same as that for $\widehat{\tau}$ 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 $h^4 n$ 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

$$\zeta_i = \Psi_p(Z_i; \tau_0, p(X_i)) - E[\Psi_p(Z_i; \tau_0, p(X_i))|X_i]$$

$$\begin{aligned}(K * K)(t) &= \int K(t)K(t-u)du \\ \langle f, g\rangle &= \int f(t)g(t)dt.\end{aligned}$$

Let now $T = \sqrt{n}(\widehat{\tau}^{bc} - \tau_0)$ and obtain the stochastic expansion $T = T^* + R$ as in (6).

THEOREM 2. *Under some regularity conditions, as $n \to \infty$, $R = o_p(n^{-3/5})$ in (6) and:*

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

[3]This would require a model for m_j, p, and f.

where

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

This shows that the bias correction can lead to improved mean squared error properties.[4] Now

$$AMSE(\widehat{\tau}^{bc}) = v_0 + nh^4 b_1^2 + \frac{1}{nh}v_1 + o(n^{-3/5})$$

and the optimal bandwidth is

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

since b_1^2, v_1 are both non-negative. This bandwidth is smaller in magnitude than is optimal for the raw estimator $\widehat{\tau}$. A feasible bandwidth selection method can be defined based on estimates of the quantities b_1, v_1, either nonparametric estimates or parametric estimates suggested from some sort of Silverman's rule of thumb idea.

Jones and Sheather (1991) investigated again the situation of Hall and Marron (1987), namely that of estimating integrated squared density derivatives. They argued against doing the degrees of freedom bias correction in this case. Their reasoning was that the leading smoothing bias term was always negative and the degrees of freedom bias term was always positive, so that by a judicious choice of bandwidth one could knock both these terms out simultaneously. In the language of our problem we would end up with [assuming that $h \asymp n^{-1/3}$]

$$AMSE(\widehat{\tau}^{JS}) \simeq v_0 + \left(\sqrt{n}h^4 b_{11}\right)^2 + \frac{1}{nh}v_1,$$

say, where b_{11} is a higher order smoothing bias term [assuming that the underlying functions are smooth enough]. In this case, the correction term is of order $n^{-2/3}$, which is even smaller than the order $n^{-3/5}$ obtained with our degrees of freedom bias correction. The catch is that in our more complicated model, the signs of the two bias terms are not necessarily opposite and so the Jones

[4] We are happy to report that this finding is partly in agreement with Rothenberg (1984a, p909) who says:

> *"This suggests that correction for bias may be more important than second order efficiency consideration when choosing among estimators."*
>
> In our case, correction for bias improves mean squared error.

and Sheather method is not guaranteed to work. In any case, the Jones and Sheather method requires estimation of higher order derivatives of the regression function and is against the spirit of our approach.

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 $\widehat{\tau}$ is normally distributed, i.e.,

$$\Pr\left[\sqrt{n}(\widehat{\tau}-\tau_0)\leq x\right]=\Phi\left(\frac{x-\sqrt{n}h^2b_1+\frac{1}{\sqrt{n}h}b_2}{\sqrt{v_0}}\right)+o(n^{-1/3}).$$

The approximation for $\sqrt{n}(\widehat{\tau}^{bc}-\tau_0)$ 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 $n^{-1/2}$;[5] 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 (1997) 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 degrees of freedom bias problem, which can be corrected in the same way as we have done for the estimator of τ.

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

[5] In fact, in both cases

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

which is the same magnitude as in parametric models.

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 investigate the magnitudes of the second order effects in Theorems 1 and 2 and the optimal bandwidth size. We compute the theoretically optimal bandwidths and mean squared errors for the following model.

Design

$$\begin{aligned} X &\sim U[-0.5, 0.5] \\ m_0(x) &= x \\ m_1(x) &= \tau + m_0(x) \\ y_0 &= m_0(x) + \eta \\ y_1 &= y_0 + \tau \\ t &= 1\left(\beta x + \delta > 0\right), \end{aligned}$$

where $\eta, \delta \sim N(0,1)$ and are mutually independent. We vary the parameters τ and β with $\tau \in \{-2,-1,0,1,2\}$ and $\beta \in \{1,2,3\}$.[6]

Note that v_0 changes substantially with β and less with τ. For example, when $(\beta,\tau) = (1,-2)$, $v_0 = 4.28$, while when $(\beta,\tau) = (1,+2)$, $v_0 = 4.31$. However, when $(\beta,\tau) = (3,-2)$, $v_0 = 29.57$ and when $(\beta,\tau) = (3,+2)$, $v_0 = 42.86$. By contrast, b_1 and b_2 are quite small in absolute terms. For

[6]The regression R^2 of $\beta x + \delta$ on x is $R^2 = 2\beta^2/(2\beta^2+3)$ and of y_j on x is given by the same formula with $\beta = 1$, i.e., $R^2 = 0.4$.

$(\beta, \tau) = (1, 2)$, $(b_1, b_2) = (0.031, 0.484)$, while for $(\beta, \tau) = (3, -2)$, $(b_1, b_2) = (12.88, -3.59)$. In most cases b_1 and b_2 have opposite signs. The constant v_1 is very large when $\beta = 3$. When $(\beta, \tau) = (1, 0)$, $v_1 = 0.5$, while when $(\beta, \tau) = (3, -2)$, $v_1 = 110.36$.

We report the relative root mean squared error against bandwidth $[RRMSE = \sqrt{AMSE/v_0}]$ in the figures below for a sample size of $n = 100, n = 1000$, and $n = 10,000$. The solid line is for the raw estimator and the dashed line is for the bias corrected estimator.

Figures 1-3 here

The effects of bandwidth on performance are quite clear from these pictures. As discussed earlier there is a bandwidth for which the RRMSE of $\widehat{\tau}$ is exactly equal to one, but this never happens for $\widehat{\tau}^{bc}$. This gives the misleading impression that the un-corrected estimator is better. In reality, we should superimpose additional terms on the RRMSE expansion of $\widehat{\tau}$ in order to achieve the same accuracy as in the expansion for $\widehat{\tau}^{bc}$. Basically, we should not take the constants too seriously. It is clear from the pictures that $\widehat{\tau}^{bc}$ has better RRMSE when h is small, but that the estimators have similar RRMSE when h is large.

Note that the Silverman's rule of thumb bandwidth [for Gaussian kernels] $h_{rot} = 1.06\sigma n^{-1/5}$ is for $n = 100$, $h_{rot} = 0.35$, $n = 200$, $h_{rot} = 0.30$, $n = 500$, $h_{rot} = 0.25$, $n = 1000$, $h_{rot} = 0.22$, and $n = 10,000$, $h_{rot} = 0.14$. Another common bandwidth choice is just $0.2 \times range$, which in this case would result in $h = 0.2$. For the small sample size these bandwidths rarely do dreadfully, but such large bandwidths can have disastrous effects in the larger samples.

5 Conclusions

Our asymptotic expansions revealed some 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

Sufficient conditions for our results can be found in numerous places for a variety of estimators $\widehat{p}$ and functions Ψ. See for example Andrews (1994), Newey & McFadden (1994), Bickel, Klaassen, Ritov, & Wellner (1993) etc. Linton (1996b) develops higher order asymptotic expansions for a general

class of semiparametric estimators. We will require smoothness conditions on p, f. We require that both p and f be bounded away from zero on the compact support of X. We also need some moment conditions on y_{ji}. The conditions should imply at least that

$$\sup_{x \in C} |\widehat{p}(x) - p(x)| = o_p(n^{-1/4}), \tag{10}$$

where C is the support of X. Sufficient conditions for this can be found in Masry (1996a,b), who actually shows that

$$\sup_{x \in C} |\widehat{p}(x) - p(x)| = O_p(h^2) + O_p\left(\sqrt{\frac{\log n}{nh}}\right),$$

which is $O_p\left(n^{-1/3}\sqrt{\log n}\right)$ when $h \asymp n^{-1/3}$ and $O_p\left(n^{-3/10}\sqrt{\log n}\right)$ when $h \asymp n^{-2/5}$; in either case this magnitude is $o_p(n^{-1/4})$ as required.

PROOF OF THEOREM 1. By a geometric series expansion

$$\begin{aligned}
\sqrt{n}(\widehat{\tau} - \tau_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi(Z_i; \tau_0, p(X_i)) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi_p(Z_i; \tau_0, p(X_i))(\widehat{p}(X_i) - p(X_i)) \\
&+ \frac{1}{2\sqrt{n}}\sum_{i=1}^n \Psi_{pp}(Z_i; \tau_0, p(X_i))(\widehat{p}(X_i) - p(X_i))^2 \\
&+ \frac{1}{6\sqrt{n}}\sum_{i=1}^n \Psi_{ppp}(Z_i; \tau_0, p(X_i))(\widehat{p}(X_i) - p(X_i))^3 \\
&+ \frac{1}{24\sqrt{n}}\sum_{i=1}^n \Psi_{pppp}(Z_i; \tau_0, p(X_i))(\widehat{p}(X_i) - p(X_i))^4 + o_p(n^{-3/4}).
\end{aligned}$$

Because the derivatives of Ψ with respect to p are dominated by a function with finite moment, because p is bounded away from zero.

When $h \asymp n^{-1/3}$, we can further drop the cubic and quartic terms to obtain

$$\begin{aligned}
\sqrt{n}(\widehat{\tau} - \tau_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi(Z_i; \tau_0, p(X_i)) + \frac{1}{\sqrt{n}}\sum_{i=1}^n s_p(X_i)(\widehat{p}(X_i) - p(X_i)) \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^n \zeta_i \cdot (\widehat{p}(X_i) - p(X_i)) + \frac{1}{2\sqrt{n}}\sum_{i=1}^n s_{pp}(X_i)(\widehat{p}(X_i) - p(X_i))^2 \\
&+ \frac{1}{2\sqrt{n}}\sum_{i=1}^n \xi_i \cdot (\widehat{p}(X_i) - p(X_i))^2 + o_p(n^{-1/3}),
\end{aligned}$$

where

$$\xi_i = \Psi_{pp}(Z_i; \tau_0, p(X_i)) - E[\Psi_{pp}(Z_i; \tau_0, p(X_i))|X_i],$$

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

$$\begin{aligned}
\Psi_p(Z_i; \tau_0, p(X_i)) &= -\frac{y_i \cdot t_i}{p(X_i)^2} - \frac{y_i \cdot (1 - t_i)}{[1 - p(X_i)]^2} \\
\Psi_{pp}(Z_i; \tau_0, p(X_i)) &= 2\frac{y_i \cdot t_i}{p(X_i)^3} - 2\frac{y_i \cdot (1 - t_i)}{[1 - p(X_i)]^3}.
\end{aligned}$$

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 know that $\beta_n(X_i) = O_p(h^2)$. The particular form of this bias function can be found in Fan and Gijbels (1996), for example.

We then write

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n s_p(X_i)(\widehat{p}(X_i) - p(X_i)) \\ = & \frac{1}{\sqrt{n}} \sum_{j=1}^n s_p(X_j)\varepsilon_j + \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j \left[\sum_{i \neq j} w_{ij} s_p(X_i) - s_p(X_j) \right] + \frac{1}{\sqrt{n}} \sum_{i=1}^n s_p(X_i)\beta_n(X_i), \end{aligned} \tag{11}$$

where the first term is $O_p(1)$ and jointly asymptotically normal with the leading term $\frac{1}{\sqrt{n}} \sum_{i=1}^n \Psi(Z_i; \tau_0, p(X_i))$, 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 $O(h^4)$. The reason is that w_{ij} are approximately symmetric [see Linton (2001)] and so $\sum_{i \neq j} w_{ij} s_p(X_i) - s_p(X_j)$ is rather like $\sum_{i \neq j} w_{ji} s_p(X_i) - s_p(X_j)$ in terms of its magnitude, and this latter quantity is just the bias function from smoothing $s_p(X_i)$ against X_i. Replacing the local linear weights by an approximation, we have

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

The third term in (11) 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

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

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

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

See Linton (1995) for a similar calculation.

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

$$\begin{aligned}\sqrt{n}(\widehat{\tau}-\tau_0) \;\simeq\; & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Z_i;\tau_0,p(X_i))+s_p(X_i)\varepsilon_i \quad [=O_p(1)] \\ & +\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j\left[\sum_{i\neq j}w_{ij}s_p(X_i)-s_p(X_j)\right] \quad \left[=O_p(h^2)\right] \\ & +\sum_{i\neq j}\sum\varphi_n(Z_i,Z_j) \quad \left[=O_p(n^{-1/2}h^{-1/2})\right] \\ & +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_p(X_i)\beta_n(X_i) \quad \left[=O_p(h^2\sqrt{n})\right] \\ & +\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}s_{pp}(X_i)M_n(X_i) \quad \left[=O_p(h^4\sqrt{n})+O_p(n^{-1/2}h^{-1})\right],\end{aligned}$$

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. ■

PROOF OF THEOREM 2. Define the quantity

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

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

$$\sqrt{n}(\widehat{\tau}^{bc}-\tau_0)=\sqrt{n}(\widehat{\tau}-\tau_0)-\frac{b_2}{\sqrt{nh}}(1+O_p(\sqrt{\frac{\log n}{nh}}+h^2)).$$

■

REFERENCES

Andrews, D.W.K., (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. Econometrica 62, 43-72.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and J. A. Wellner (1993). Efficient and adaptive estimation for semiparametric models. The John Hopkins University Press, Baltimore and London.

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.

Hall, P., and J.S. Marron (1987). Estimation of Integrated Squared Density Derivatives. Statistics and Probability Letters 6, 109-115.

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.

Jones, M.C., and S.J. Sheather (1991). Using non-Stochastic terms to Advantage in Kernel-based Estimation of Integrated Squared Density Derivatives. Statistics and Probability Letters 11, 511-514.

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.

Linton, O.B. (2001). Symmetrizing and unitizing transformations for linear smoothing weights. Computational Statistics 16, 153-164.

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.

Newey, W.K. and D.F. McFadden (1994). Large sample estimation and hypothesis testing. in *Handbook of Econometrics, vol. IV*, eds. D.F. McFadden and R.F. Engle III. North Holland.

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

Powell, J.L., and T.M. Stoker (1996). Optimal bandwidth choice for density-weighted averages. Journal of Econometrics 75, 291-316.

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.

Srinavasan, T.N. (1970). Approximations to Finite Sample Moments of Estimators whose Exact Sampling Distributions are unknown. Econometrica, 38, 533-541.

Xiao, Z. and O.B. Linton (2001). Second order approximation for an adaptive estimator in a linear regression. Econometric Theory 17, 984-1024.

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

Figure 1: Figure 1. n=100

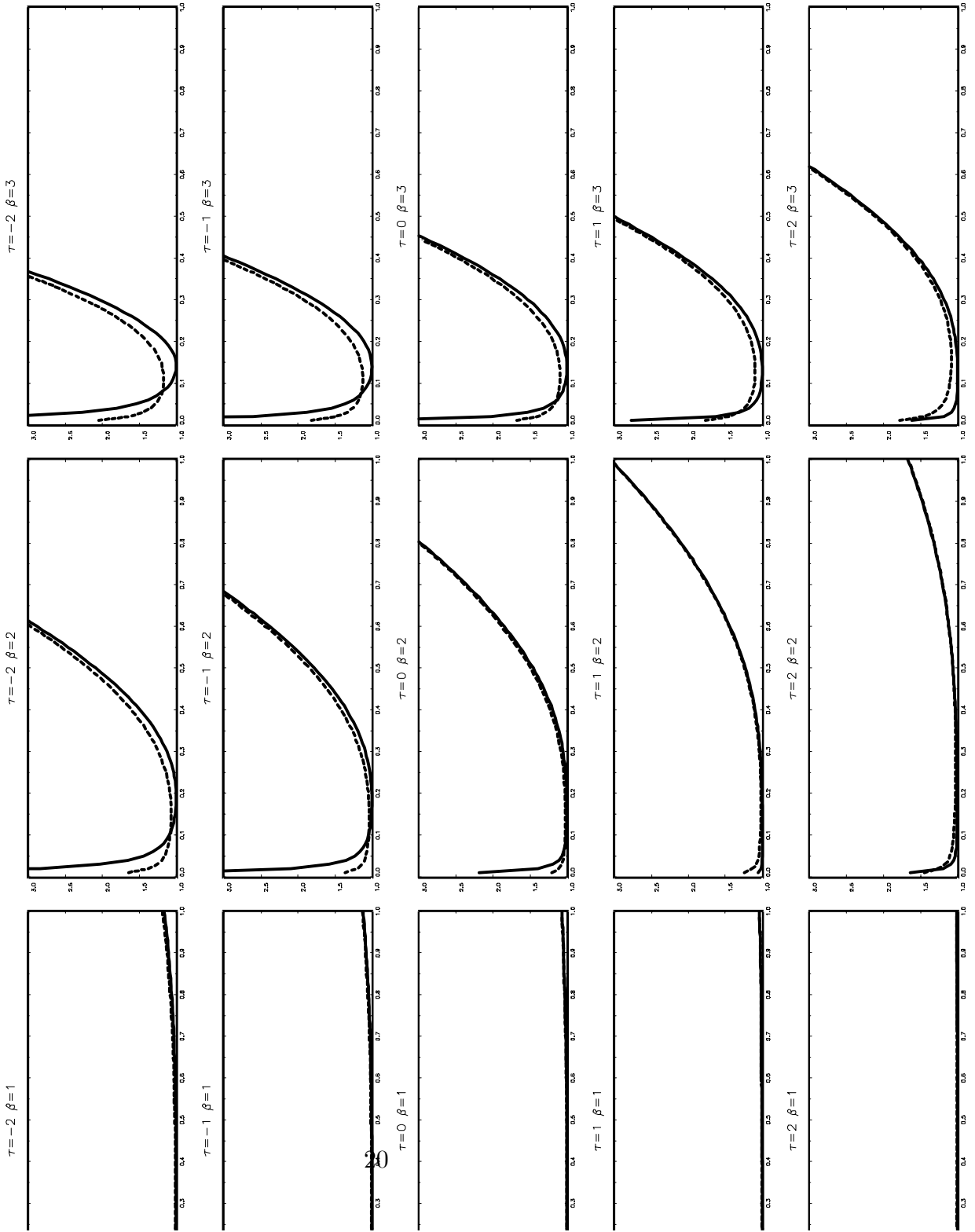

Figure 2: Figure 2. n=1000

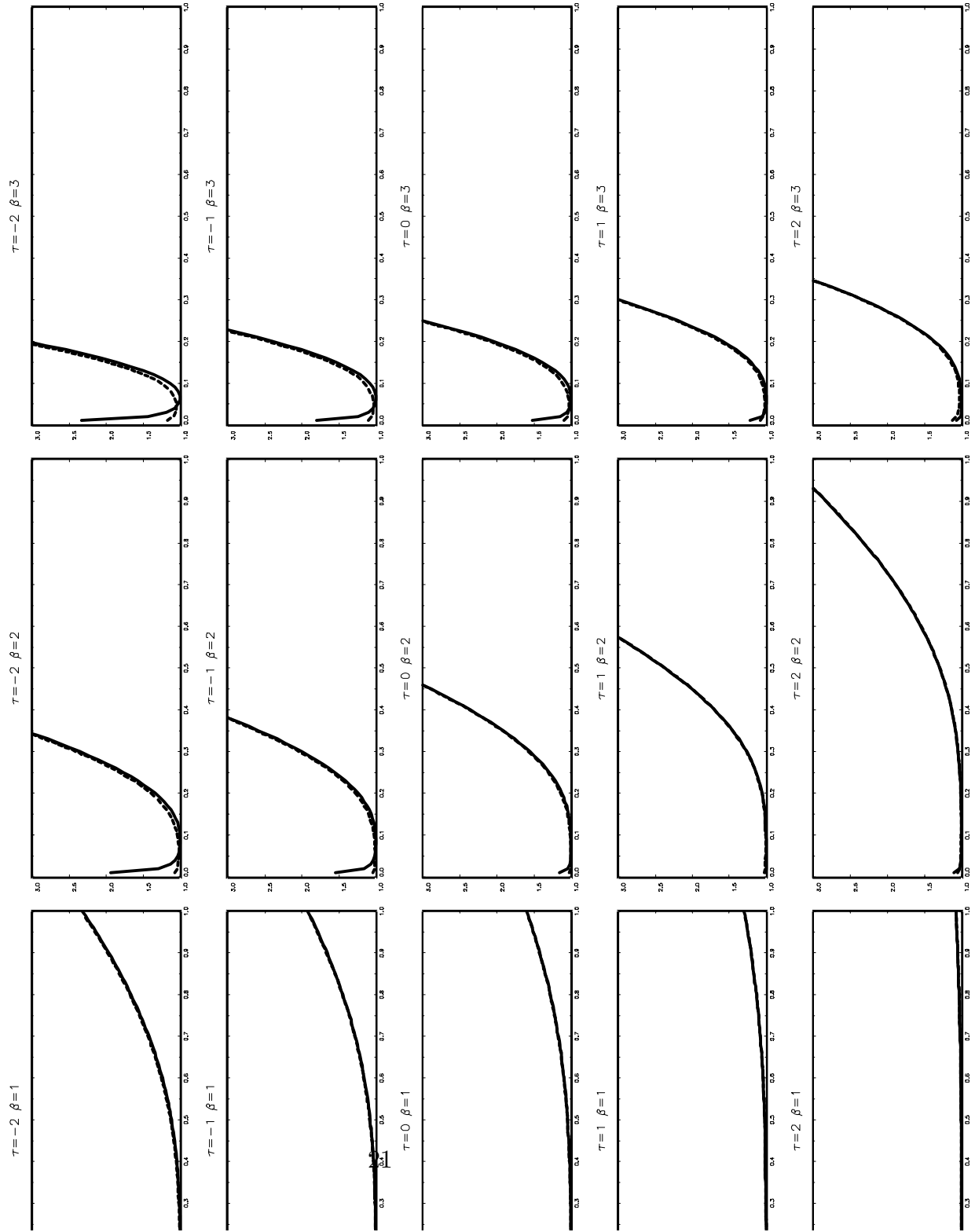

Figure 3: Figure 3. n=10,000