Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 136 (2007) 65–88
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

A method of estimating the average derivative

Anurag N. Banerjee

Department of Economics, University of Southampton, Southampton, UK

Available online 24 August 2005

Abstract

We derive a simple semi-parametric estimator of the "direct" Average Derivative, $\boldsymbol{\delta} = E(D[m(\mathbf{x})])$, where $m(\mathbf{x})$ is the regression function and S, the support of the density of $\mathbf{x}$ is compact. We partition S into disjoint bins and the local slope $D[m(\mathbf{x})]$ within these bins is estimated by using ordinary least squares. Our average derivative estimate $\widehat{\boldsymbol{\delta}}_a$, is then obtained by taking the weighted average of these least squares slopes. We show that this estimator is asymptotically normally distributed. We also propose a consistent estimator of the variance of $\widehat{\boldsymbol{\delta}}_a$. Using Monte-Carlo simulation experiments based on a censored regression model (with Tobit Model as a special case) we produce small sample results comparing our estimator with the Härdle–Stoker [1989. Investigating smooth multiple regression by the method of average derivatives. Journal of American Statistical Association 84, 408, 986–995] method. We conclude that $\widehat{\boldsymbol{\delta}}_a$ performs better that the Härdle–Stoker estimator for bounded and discontinuous covariates.

JEL classification: C13; C14; C15

Keywords: Semi-parametric estimation; Average derivative estimator; Linear regression

1. Introduction

Social scientists and applied economists are often more interested in the derivative of a conditional expectation function than the conditional expectation function itself.

E-mail address: a.n.banerjee@soton.ac.uk.

0304-4076/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2005.07.010

To estimate derivatives applied researchers typically work with parametric specifications of the conditional expectation function, $m(\mathbf{x})$. The derivatives, $D[m(\mathbf{x})]$ are easily computed after the conditional expectation function has been estimated. The popularity of parametric estimators in empirical analysis is primarily due to the ease with which these estimators can be implemented. However, the validity of this approach ultimately rests on a number of functional form assumptions which may be hard to justify.

Econometricians have, therefore, developed alternative procedures to estimate derivatives of conditional expectations which are based on nonparametric methods. But, nonparametric methods can have relatively slow rates of convergence. Furthermore convergence rates depend on the dimensionality and smoothness of the underlying conditional expectation function, $m(\mathbf{x})$ (Härdle, 1992, pp. 91). However, in many applications, the objective of the analysis is not to estimate the entire derivative curve of a conditional expectation function at each data point. Instead, it is often sufficient to construct statistics to estimate the average derivative (AD) as

$$\boldsymbol{\delta} = \int D[m(\mathbf{x})]f(\mathbf{x})\,\mathrm{d}\mathbf{x} = E(D[m(\mathbf{x})]), \tag{1}$$

where $f(\mathbf{x})$ is the marginal density of the $\mathbf{x}$'s and $D[m(\mathbf{x})]$ is the first derivative of $m(\mathbf{x})$.

The primary interest for average derivative estimate (ADE) came from the index models. Also further motivation of ADEs can be found in specific measurement problems in economics, such as measuring the positive definiteness of the aggregate income effects matrix for assessing the "Law of Demand" (Härdle et al. (1991)).[1]

From a statistical standpoint, the benefit of the additional averaging in (1) is that $\sqrt{T}$ asymptotics are obtainable, where T is sample size. The most frequently used nonparametric techniques are based on kernel estimators. Average derivatives can be estimated at a rate of $\sqrt{T}$ by using kernel based ADEs. These estimators have been proposed by Härdle and Stoker (1989) and Powell et al. (1993). They avoid the slow rates of convergence encountered in more traditional nonparametric estimation. At the same time, the asymptotic properties of these estimators only depend on the joint distribution of the data, and do not rely on functional form assumptions of the conditional expectation function.

While the theory of these estimators is elegant, applications of these estimators often require modifications of the kernel based ADE framework (Zhang and Karuhamuni, 2000). The main purpose of this paper is to estimate the average derivative under conditions which make a straight-forward application of the standard ADE problematic. In particular, the support of the main covariates in many applications is often bounded (for example; the number of hours worked with a discontinuity at the legal limit for regular hours), whereas the ADE literature essentially assumes that the support of the covariates is unbounded and the density is

[1]See Stoker (1991a) and Härdle (1992) for more discussion on applications of ADE.

smooth. We therefore present an ADE framework for bounded covariates and possible non-smooth covariate densities.

In this article we shall present a simple estimator when the covariates of the model has bounded support. Unlike the kernel based non-parametric estimators described above, it is based on simple weighted average of the slope coefficients which are obtained by piecewise disjoint linear regression. We shall show that this ADE is "better" than the traditional ADE when the support base is bounded. Since our assumptions are not similar our method is complementary to the kernel based method.

Section 2 describes our method along with the kernel based methods proposed by Härdle and Stoker (1989) and Stoker (1991b). Section 3 gives the asymptotic properties of our estimator and its variance. We also discuss how to estimate the AD when some of the covariates are discrete in Section 3.1. In Section 4, a short comparison with Härdle–Stoker (H–S) estimator has been made using Monte-Carlo simulation method. Concluding remarks are presented in Section 5. The proofs of the theorems are given in the appendix.

2. Method of estimation

Suppose the y_t's are independent random variables generated by data generating process:

$$y_t = m(\mathbf{x}_t) + u_t, \quad t = 1, \ldots, T, \tag{2}$$

where $E(y|\mathbf{x}) = m(\mathbf{x})$ is the unknown regression function. The covariates $\mathbf{x}$ is a d-dimensional[2] random vector with density $f(\mathbf{x})$ and have a positive definite variance–covariance matrix, $\mathbf{\Sigma}$. The error terms u_t's are distributed with $E(u) = 0$ and $Var(u) = \sigma_u^2$. We are interested in estimating the functional $\boldsymbol{\delta} = E(D[m(\mathbf{x})])$. Integrating by parts we get the "indirect" representation of the AD as

$$\boldsymbol{\delta} = E(\mathscr{L}(\mathbf{x})y),$$

where $\mathscr{L}(\mathbf{x}) = -D[f(\mathbf{x})]/f(\mathbf{x})$.

2.1. Estimation by Kernel methods

The Härdle–Stoker (1989) method proposes the following the "indirect" estimate:

$$\widehat{\boldsymbol{\delta}}_{\mathrm{hs}} = -\frac{1}{T}\sum_{t=1}^{T} y_t \frac{\widehat{Df}_h(\mathbf{x}_t)}{\widehat{f}_h(\mathbf{x}_t)} I\{\widehat{f}_h(\mathbf{x}_t) > b_{\mathrm{T}}\}, \tag{3}$$

where $\widehat{f}_h(\mathbf{x})$ and $\widehat{Df}_h(\mathbf{x})$ are the kernel density estimates of $f(\mathbf{x})$ and $D[f(\mathbf{x})]$, respectively with a bandwidth h. In addition, $I\{\cdot\}$ denotes the indicator function, and b_{T} is a sequence of truncation values which converges to zero. Theorem (3.1) in

[2]Banerjee (1994) analyses the case of $d = 1$.

Härdle and Stoker (1989) establishes that, under appropriate assumptions,

$$\sqrt{T}(\widehat{\boldsymbol{\delta}}_{\mathrm{hs}} - \boldsymbol{\delta}) \xrightarrow{\mathrm{d}} N(0, Var(D[m(\mathbf{x})]) - u\mathscr{L}(\mathbf{x})). \tag{4}$$

Notice that when the error term u_t is independent of $\mathbf{x}_t$,

$$Var(\widehat{\boldsymbol{\delta}}_{\mathrm{hs}}) = \mathrm{Var}(D[m(\mathbf{x})]) + \sigma_u^2 E(\mathscr{L}(\mathbf{x}))^2.$$

A nice property of this estimator is that it converges at a rate of $\sqrt{T}$ despite the fact that the nonparametric estimator of the individual derivatives converges at a much slower rate (Ullah and Vinod, 1988).

Stoker (1991b), defined the "direct" estimator of δ as

$$\widehat{\boldsymbol{\delta}}_{\mathrm{s}} = \frac{1}{T}\sum_{t=1}^{T} D[\widehat{m}(\mathbf{x}_t)]I\{\widehat{f}_h(\mathbf{x}_t) > b_n\}, \tag{5}$$

where $\widehat{m}(\mathbf{x}_t)$ is the kernel regression estimator of $m(\mathbf{x})$. Stoker also found that under appropriate conditions, $\sqrt{T}(\widehat{\boldsymbol{\delta}}_{\mathrm{s}} - \boldsymbol{\delta})$ is also asymptotically distributed as normal. Li (1996) showed the asymptotic equivalence between $\widehat{\boldsymbol{\delta}}_{\mathrm{s}}$ and $\widehat{\boldsymbol{\delta}}_{\mathrm{hs}}$.

2.2. Estimation by piecewise local linear regression

The estimator we are going to define, uses piecewise local linear regression (PLLR) as a method.

We shall only assume some smoothness properties of the regression function and moment restrictions on the random variables which we shall state in the next section. One important difference of this method from the other methods is that there are no smoothness assumptions on the marginal density of $\mathbf{x}$, i.e. $f(\mathbf{x})$. We assume that, the support of $\mathbf{x}$ is the compact set S. Without loss of generality it is assumed to be a subset of $[0,1]^{\mathrm{d}}$. So, in contrast with the Härdle and Stoker method, in our method $\mathscr{L}(\mathbf{x})$ does not exist at the boundaries: there can be even other finite number of discontinuities of $f(\mathbf{x})$. Thus we cannot use the method proposed by Härdle and Stoker. On the other hand if $\mathbf{x}$ has unbounded support we cannot use our method since we assume the support of $f(\mathbf{x})$ to be compact. Though in this case we may be able to use the Härdle and Stoker method with some smoothness conditions on $f(\mathbf{x})$. So comparisons of our two methods in terms of the asymptotics cannot be made and our method complement the Härdle and Stoker method.

Let us motivate our method when x is univariate. When $d = 1$, S can be clearly taken to be the interval $[0,1]$. The interval $S = [0,1]$ is then partitioned in equal intervals. We denote such a partition as $P = \{(t_r, t_{r+1}] : 0 < t_1 < t_2 < \cdots < t_{k-1} < 1\}$. Denote $(t_r, t_{r+1}]$ as H_r (H_r is called a bin). These bins are of equal size ($|H_r| = h$). For any bin H_r, which has at least 3 observations we can linearly regress y_t on x_t, for all $x_t \in H_r$. We denote the estimated coefficient of the slope of the regression as $\widehat{\beta}_r$. This is a least squares estimate of the tangent of the regression curve $m(\mathbf{x})$, in the interval H_r. We then take a weighted average of the slopes in each of the bin H_r. The weights are taken to be the proportion of the total number of observations in the bin H_r.

Let us now generalise the idea when the dimension of $\mathbf{x}$ is d. Let the support of $\mathbf{x}$ be the compact set $S \subseteq [0,1]^{\mathrm{d}}$. Assume without loss of generality, the interval $[0,1]$ is the support of the marginal density of x_i. As before we partition the support in equal intervals and denote the partition as P_i. Let the partition P_i be $\{(t_{ir}, t_{i(r+1)}] : 0 < t_{i1} < t_{i2} < \cdots < t_{i(k_i-1)} < 1\}$. We denote $(t_{ir}, t_{i(r+1)}]$ as H_{r_i} (H_{r_i} is called a bin in the ith dimension). These bins are of equal size ($|H_{r_i}| = h^{1/d}$). The partition for the whole support of $f(\mathbf{x})$ is then $\mathbf{P} = \times_{i=1}^{d} P_i$, and $\mathbf{H}_r = \times_{i=1}^{d} H_{r_i}$ is the bin to be considered in this d-dimensional space. We shall only consider those bins such that $\mathbf{H}_r \subseteq S$ therefore the number of bins are at most $\mathbf{\Pi}_{i=i}^{d} k_i$. Note that when x is univariate then $H_r \subseteq S$ for all r. The same is true if the covariates are independent. Since there is no a priori reason to have different number of partitions for each dimension, we take the same number of partitions (k) for each dimension. Then the size of each bin is

$$|H_r| = h = k^{-d}. \tag{6}$$

The rest of the method is similar to the univariate case, which we now explain. Supposing we have at least $l \geqslant d+2$ points in $\mathbf{H}_r$ we linearly regress y_t on $\mathbf{x}_t$ as

$$y_t = \alpha_r + \boldsymbol{\beta}_r' \mathbf{x}_t, \quad \text{where } \mathbf{x}_t \in \mathbf{H}_r. \tag{7}$$

Define the Bernoulli random variable,

$$I\{\mathbf{x}_t \in \mathbf{H}_r\} = I_{t,r} \sim \mathrm{Bern}(p_r), \tag{8}$$

where $p_r = \int_{\mathbf{H}_r} f(\mathbf{x})\,\mathrm{d}\mathbf{x}$. We denote the estimated coefficient of the slope of the regression, $\boldsymbol{\beta}_r$ as

$$\widehat{\boldsymbol{\beta}}_r = [\mathbf{S}_x^r]^{-1} \mathbf{S}_{xy}^r,$$

where

$$\mathbf{S}_{xy}^r = \frac{1}{T_r} \sum_{t=1}^{T} (\mathbf{x}_t - \overline{\mathbf{x}}_r) I_{t,r} y_t,$$

$$\mathbf{S}_x^r = \frac{1}{T_r} \sum_{t=1}^{T} (\mathbf{x}_t - \overline{\mathbf{x}}_r)(\mathbf{x}_t - \overline{\mathbf{x}}_r)' I_{t,r},$$

$$\overline{\mathbf{x}}_r = \frac{1}{T_r} \sum_{t=1}^{T} \mathbf{x}_t I_{t,r}$$

and

$$T_r = \sum_{t=1}^{T} I_{t,r},$$

the number of observation in the bin $\mathbf{H}_r$. This is a least squares estimate of the tangent of the regression curve $m(\mathbf{x})$, within the interval $\mathbf{H}_r$. To estimate the AD we then take an weighted average of the slopes in each of the bin $\mathbf{H}_r$. The weights are

taken to be the proportion of the total number of observations in the bin $\mathbf{H}_r$,

$$w_r = \frac{T_r}{T}.$$

Definition 1. We define our ADE as

$$\widehat{\boldsymbol{\delta}}_{\mathrm{a}} = \sum_{r=1}^{k} w_r \widehat{\boldsymbol{\beta}}_r I\{T_r \geqslant l\}. \tag{9}$$

Note that in Definition (9), we assume that if there are insufficient number of observations (i.e. $T_r < l$) in the bin $\mathbf{H}_r$ to regress, those observations contribute nothing to the ADE. This throwing away of observations does not matter in large samples.

We will show later that under some natural assumptions, asymptotically

$$\sqrt{T}(\widehat{\boldsymbol{\delta}}_{\mathrm{a}} - \boldsymbol{\delta}) \simeq N(0, Var(D[m(\mathbf{x})]) + \sigma_u^2 \boldsymbol{\Sigma}^{-1}).$$

We can also estimate the error variance σ_u^2, by defining a weighted average of the variance estimates in each bin $\mathbf{H}_r$. We denote the estimate variance s_r^2 in each bin as the usual least squares estimator

$$s_r^2 = \frac{1}{T_r} \sum_{t=1}^{T} \widehat{u}_{r,t}^2 I_{t,r},$$

where

$$\widehat{u}_{r,t} = y_t - \overline{y}_r - (\mathbf{x}_t - \overline{\mathbf{x}}_r)' \widehat{\boldsymbol{\beta}}_r \tag{10}$$

are the residuals of the linear regression (7) and

$$\overline{y}_r = \frac{1}{T_r} \sum_{t=1}^{T} y_t I_{t,r}.$$

Definition 2. We define our average error variance estimator as

$$s_{\mathrm{a}}^2 = \sum_{r=1}^{k} w_r s_r^2 I\{T_r \geqslant l\}. \tag{11}$$

Under some assumptions we will show later that,

$$s_{\mathrm{a}}^2 \xrightarrow{P} \sigma_u^2.$$

Consequently, we also show that the large sample variance of our ADE, $Var(D[m(\mathbf{x})]) + \sigma_u^2 \boldsymbol{\Sigma}^{-1}$ can be consistently estimated by $\widehat{\mathbf{V}}_{\mathrm{a}}$, as defined below.

Definition 3. We define the estimated variance of $\widehat{\boldsymbol{\delta}}_{\mathrm{a}}$ as

$$\widehat{\mathbf{V}}_{\mathrm{a}} = \left(\sum_{r=1}^{k} w_r \widehat{\boldsymbol{\beta}}_r \widehat{\boldsymbol{\beta}}_r' I\{T_r \geqslant l\} - \widehat{\boldsymbol{\delta}}_{\mathrm{a}} \widehat{\boldsymbol{\delta}}_{\mathrm{a}}' \right) + s_{\mathrm{a}}^2 \sum_{r=1}^{k} w_r [\mathbf{S}_x^r]^{-1} I\{T_r \geqslant l\}. \tag{12}$$

Furthermore we need to select our bin size depending on the number of observations T. We propose a rule of thumb to choose the bin size as.

Rule of Thumb. We choose the bin size as

$$h = \frac{(\ln(T))^{1/2}}{T^{3/4}}. \tag{13}$$

It is clear from (6) that we can calculate the number of partitions as

$$k = \left\lfloor \frac{T^{3/4d}}{(\ln(T))^{1/2d}} \right\rfloor,$$

where $\lfloor v \rfloor$ is the integer part of the number v. The reason to choose such a bin size will be clear when we discuss our large sample results.

3. Large sample results

We shall now prove some large sample results under the following assumptions:

A(1) The support of $f(\mathbf{x})$ is the compact set $S \subset [0,1]^{\mathrm{d}}$ and $0 < \underline{c} \leqslant f(\mathbf{x}) < \overline{C}$.
A(2) The second derivative of $m(\mathbf{x})$, $D^2[m(\mathbf{x})]$, exists and is bounded by M_2.
A(3) The covariates $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ are independent of the errors $(u_1, \ldots, u_T)$, and the variance of u_t, $E(u_t^2) = \sigma_u^2 < \infty$.
A(4) As $T \longrightarrow \infty$,

$$\sqrt{T}h \longrightarrow 0 \quad \text{and} \quad \frac{\ln(T)}{Th} \longrightarrow 0.$$

We will make some brief comments on the assumptions. The first Assumption A(1) is not popular in the non-parametric econometrics literature. The assumption that the density $f(\mathbf{x})$ is bounded below is necessary to ensure that there are at least l-observations in each bin to perform the required regression (in large sample). However we also want the density to be bounded above since we do not want to put too much weight on any particular $\widehat{\boldsymbol{\beta}}_r$. The smoothness assumption of the regression function A(2) is also necessary for the same reason, and is similar to assumptions made in Fan and Gijbels (1992). Assumption A(3) is a standard assumption for linear models required for consistency of the estimates of slope parameters. Finally the last assumption A(4) ensures that the size of the bins shrinks at the rate of $\sqrt{T}$, but the size should not get too small too quickly ($\ln(T)/Th \longrightarrow 0$) otherwise there will be insufficient number of observations in the bin to perform a regression. Note that the bin-size proposed in (13) satisfies A(4).

Theorem 1. *Under Assumptions* A(1)–A(4),

$$\sqrt{T}(\widehat{\boldsymbol{\delta}}_{\mathrm{a}} - \boldsymbol{\delta}) \xrightarrow{D} N(0, Var(D[m(\mathbf{x})]) + \sigma_u^2 \boldsymbol{\Sigma}^{-1}),$$

where $\widehat{\boldsymbol{\delta}}_{\mathrm{a}}$ *is the ADE defined in* (9).

Observe that if $m(\mathbf{x})$ is linear (i.e. $m(\mathbf{x}) = \alpha + \boldsymbol{\beta}'\mathbf{x}$) we have $\boldsymbol{\delta} = \boldsymbol{\beta}$. Then the asymptotic variance coincides with the asymptotic variance of the classical least squares estimator of $\boldsymbol{\beta}$. So in the case of linearity of the regression function, we do not lose efficiency when compared to the simple OLS method.

The following theorems help us to estimate the variance of $\widehat{\boldsymbol{\delta}}_{\mathrm{a}}$ consistently.

Theorem 2. *Under Assumptions* A(1)–A(4) *and* $E(u^4) < \infty$,

$$s_{\mathrm{a}}^2 \xrightarrow{P} \sigma_u^2,$$

where s_{a}^2 *is the estimate of the variance of the errors as defined in* (11).

Theorem 3. *Under Assumptions* A(1)–A(4) *and* $E(u^4) < \infty$,

$$\widehat{\mathbf{V}}_{\mathrm{a}} \xrightarrow{P} Var(D[m(\mathbf{x})]) + \sigma_u^2 \boldsymbol{\Sigma}^{-1},$$

where $\widehat{\mathbf{V}}_{\mathrm{a}}$ *is the estimated variance of the ADE defined in* (12).

Theorem (3) can be used for testing linear restrictions such as the null hypothesis $\mathrm{H}_0 : \mathbf{Q}\boldsymbol{\delta} = \mathbf{q}_0$, where $\mathbf{Q}$ is a full rank matrix. The test of this hypothesis can be based on the Wald Statistic $W = (\mathbf{Q}\widehat{\boldsymbol{\delta}}_{\mathrm{a}} - \mathbf{q}_0)'[\mathbf{Q}\widehat{\mathbf{V}}_{\mathrm{a}}\mathbf{Q}]^{-1}(\mathbf{Q}\widehat{\boldsymbol{\delta}}_{\mathrm{a}} - \mathbf{q}_0)$ which will have a limiting $\chi^2(rank(\mathbf{Q}))$ distribution under the null.

3.1. Models with discrete covariates

When there are discrete covariates present the model (2) can be written as

$$y_t = m(\mathbf{z}_t, \mathbf{x}_t) + u_t, \quad t = 1, \ldots, T, \tag{14}$$

where $\mathbf{z}_t$ is a $(d_1 \times 1)$ vector of discrete covariates taking finite number of values. For our purposes we are interested in estimating

$$\boldsymbol{\delta} = E(D_x[m(\mathbf{z}, \mathbf{x})])$$

which can be written as

$$\boldsymbol{\delta} = \sum_j \Pr(\mathbf{z} = \mathbf{j})\boldsymbol{\delta}(\mathbf{j}), \tag{15}$$

where $\boldsymbol{\delta}(\mathbf{j}) = \int D_x[m(\mathbf{j}, \mathbf{x})] f(\mathbf{j}, \mathbf{x})\, \mathrm{d}x$, the average derivative conditional on $\mathbf{j}$. In the case of single index models: $m(\mathbf{z}, \mathbf{x}) = G(\boldsymbol{\alpha}'\mathbf{z} + \boldsymbol{\beta}'\mathbf{x})$, Horowitz and Härdle (1996) proposed a kernel based method to estimate $\boldsymbol{\delta}$.

Using our method, the estimation of $\delta_a(\mathbf{j})$ is easily done in the way described in Section 2 conditional on $\mathbf{j}$. We then form a weighted average of these $\hat{\boldsymbol{\delta}}(\mathbf{j})$ as

$$\widehat{\boldsymbol{\delta}}_a = \sum_j w(\mathbf{j})\widehat{\boldsymbol{\delta}}_a(\mathbf{j}), \tag{16}$$

where $w(\mathbf{j}) = \#(\mathbf{Z} = \mathbf{j})/T$.

By law of large numbers $w(\mathbf{j})$ is a consistent estimator for $\Pr(\mathbf{Z} = \mathbf{j})$ and by Theorem (1),

$$\sqrt{T}(\widehat{\boldsymbol{\delta}}_a(\mathbf{j}) - \boldsymbol{\delta}(\mathbf{j})) \xrightarrow{D} N(0, Var(m(\mathbf{j}, \mathbf{x})) + \sigma_u^2 \boldsymbol{\Sigma}(\mathbf{j})^{-1}),$$

where $\boldsymbol{\Sigma}(\mathbf{j}) = Var(\mathbf{x}|\mathbf{j})$. Therefore $\sqrt{T}(\widehat{\boldsymbol{\delta}}_a - \boldsymbol{\delta})$ is also distributed asymptotically normal.

3.2. Efficiency comparison with Härdle–Stoker ADE

Under the assumption of independence between the covariates and the errors, the asymptotic variances of the $\widehat{\boldsymbol{\delta}}_{hs}$ (or $\widehat{\boldsymbol{\delta}}_s$)[3] (4) and the asymptotic variance of $\widehat{\boldsymbol{\delta}}_a$ (Theorem 1) can be compared as[4]

$$Var(\hat{\boldsymbol{\delta}}_{hs}) = Var(D[m(\mathbf{x})]) + \sigma_u^2 E(\mathscr{L}(\mathbf{x}))^2 > Var(D[m(\mathbf{x})]) + \sigma_u^2 \boldsymbol{\Sigma}^{-1} = Var(\hat{\boldsymbol{\delta}}_a).$$

Does this imply that $\widehat{\boldsymbol{\delta}}_a$ is asymptotically more efficient than $\widehat{\boldsymbol{\delta}}_{hs}$ or $\widehat{\boldsymbol{\delta}}_d$? The answer to that question is not necessarily so, since a crucial condition used in Härdle and Stoker's (1989) proof of $\sqrt{T}$ asymptotic normality is that the bias is $o(\sqrt{T})$. Typically the bias is bounded by a Taylor expansion, assuming that $f(\mathbf{x})$ is $d+2$ continuously differentiable. However, in the setup here, the support of the covariates $\mathbf{x}$ is bounded. This, in general, makes the above derivation invalid particularly at the points near the boundary and also at the points of discontinuities of the density of the covariates, since we need the assumption of compact support of f for asymptotic normality of $\widehat{\boldsymbol{\delta}}_a$. Hence we compare them through simulation methods.

4. Small sample results

We will now study the small sample properties of our estimator and compare it with the H–S estimator using Monte-Carlo simulations. Since the primary interest for ADE came from the index models we shall study the following censored regression model, with five dummy variables $\mathbf{z} = (z_1, \ldots, z_5)$ and five "continuous" covariates $\mathbf{x} = (x_1, \ldots, x_5)$, as

$$y_t = \begin{cases} \mathbf{1}'\mathbf{z}_t + \mathbf{1}'\mathbf{x}_t + \sigma u_t, & \text{if } y_t > 0, \\ 0 & \text{otherwise,} \end{cases} \quad t = 1, \ldots, T, \tag{17}$$

[3]Newey and Stoker (1993) studies the efficiency of kernel based ADEs.

[4]This follows from the fact that $Cov(\mathbf{X}, \mathscr{L}(\mathbf{X})) = \mathbf{I}_d$.

where

$u_t \sim iid\ G(u)$,
$z_{it} \sim iid\ \text{Bern}\left(\frac{1}{2}\right)$,
$\mathbf{x}_t \sim iid$ with density $f(\mathbf{x})$
and $\mathbf{1}'$ is a vector of ones.

For the simulation we shall specify three different error distributions ($G(u)$) as follows:

1. the standard normal distribution (Φ) (This gives us the standard Tobit model),
2. the t-distribution with 3 degrees of freedom (t_3) and
3. the logistic distribution (Λ).

We also choose to simulate $\mathbf{x}$ from two different densities,

(a) $f_1(\mathbf{x}) = \Pi_{i=1}^{5} f_1^{(i)}(x_i)$, where

$$f_1^{(i)}(x_i) = \begin{cases} 1 & \text{if } x_i \in [0,1], \\ 0 & \text{otherwise,} \end{cases} \quad (i = 1, \ldots, 5),$$

being uniformly distributed over a unit 5-dimensional cube and
(b) $f_2(\mathbf{x}) = \Pi_{i=1}^{5} f_2^{(i)}(x_i)$, where

$$f_2^{(i)}(x_i) = \begin{cases} \frac{1}{2} & \text{if } x_i \in [0, \frac{1}{2}), \\ \frac{3}{2} & \text{if } x_i \in [\frac{1}{2}, 1], \\ 0 & \text{otherwise,} \end{cases} \quad (i = 1, \ldots, 5),$$

a 5-dimensional step function density. We also assume $\mathbf{z}$ and $\mathbf{x}$ are independent.

Using Greene (1999) we have for given $\mathbf{z} = \mathbf{j}$,

$$\frac{\partial E(y|\mathbf{x}, \mathbf{j})}{\partial x_i} = G\left(\frac{\mathbf{1}'\mathbf{j} + \mathbf{1}'\mathbf{x}}{\sigma}\right), \quad i = 1, \ldots, 5. \tag{18}$$

Then the average derivative at $\mathbf{z} = \mathbf{j}$ of this model takes the form of

$$\boldsymbol{\delta}(\mathbf{j}) = E_f G\left[\left(\frac{\mathbf{1}'\mathbf{j} + \mathbf{1}'\mathbf{x}}{\sigma}\right)\right]\mathbf{1}'.$$

Using (15) and the fact $\mathbf{1}'\mathbf{z} \sim Bin(n; 5, \frac{1}{2})$ we have

$$\boldsymbol{\delta}(f, G, \sigma) = \delta_0(f, G, \sigma)\mathbf{1}',$$

where

$$\delta_0(f, G, \sigma) = \sum_{n=0}^{5} \binom{5}{n} \frac{1}{2^5} E_f\left[G\left(\frac{n + \mathbf{1}'\mathbf{x}}{\sigma}\right)\right].$$

The integration of $E_f[G((n + \mathbf{1}'\mathbf{x})/\sigma)]$ was done using Monte-Carlo simulation.

We generate 10 000 data sets of size 5000(T) from the model (17) with different distribution specifications. We estimate the average derivative using the piecewise local linear regression (PLLR) method described in Section 2. The bin size used is

Table 1
Covariates distributed as f_1 density, errors normally distributed and $\sigma = 3$

$\delta_0(f_1, \Phi, 3) = 0.91827$			
Average $\widehat{\delta}$		MSE	
PLLR	H–S	PLLR	H–S
1.4402	0.14318	0.40000	0.60124
1.4201	0.14166	0.38006	0.60363
1.4591	0.14251	0.41769	0.60225
1.4364	0.14207	0.39510	0.60296
1.4737	0.14381	0.43150	0.60035

Table 2
Covariates distributed as f_2 density, errors normally distributed and $\sigma = 3$

$\delta_0(f_2, \Phi, 3) = 0.95107$			
Average $\widehat{\delta}$		MSE	
PLLR	H–S	PLLR	H–S
1.3661	−0.0526	0.28808	1.0080
1.3610	−0.0530	0.28178	1.0087
1.3585	−0.0529	0.27598	1.0084
1.3684	−0.0535	0.29196	1.0097
1.3642	−0.0526	0.28547	1.0079

Table 3
Covariates distributed as f_1 density, errors distributed as $t(3)$ and $\sigma = \sqrt{3}$

$\delta_0(f_1, t_3, \sqrt{3}) = 0.94950$			
Average $\widehat{\delta}$		MSE	
PLLR	H–S	PLLR	H–S
1.4499	0.14760	0.36794	0.64352
1.4672	0.14827	0.38225	0.64245
1.4538	0.14724	0.36251	0.64408
1.4484	0.14719	0.36774	0.64417
1.4591	0.14767	0.36501	0.64340

described in the rule of thumb (13). For comparison we also estimate the weighted average estimator (proposed by Horowitz and Härdle (1996)) $\widehat{\boldsymbol{\delta}}_{\mathrm{hs}} = \sum_{n=0}^{5} w(n)\widehat{\boldsymbol{\delta}}_{\mathrm{hs}}(n)$, where $w(n) = \#(\mathbf{1}'\mathbf{z} = n)/T$ and $\widehat{\boldsymbol{\delta}}_{\mathrm{hs}}(n)$ the estimator conditional on n. The optimal

Table 4
Covariates distributed as f_2 density, errors distributed as $t(3)$ and $\sigma = \sqrt{3}$

$\delta_0(f_2, t_3, \sqrt{3}) = 0.96721$			
Average $\widehat{\delta}$		MSE	
PLLR	H–S	PLLR	H–S
1.3659	−0.0509	0.26232	1.0370
1.3861	−0.0522	0.27962	1.0396
1.3905	−0.0510	0.27709	1.0371
1.3617	−0.0524	0.25702	1.0402
1.3558	−0.0520	0.24932	1.0393

Table 5
Covariates distributed as f_1 density, errors distributed as *Logistic* and $\sigma = \sqrt{3}$

$\delta_0(f_1, \Lambda, \sqrt{3}) = 0.91744$			
Average $\widehat{\delta}$		MSE	
PLLR	H–S	PLLR	H–S
1.4397	0.14323	0.41608	0.59992
1.4830	0.14391	0.45888	0.59891
1.4475	0.14354	0.40994	0.59949
1.4646	0.14216	0.42456	0.60164
1.4534	0.14267	0.41794	0.60082

Table 6
Covariates distributed as f_2 density, errors distributed as *Logistic* and $\sigma = \sqrt{3}$

$\delta_0(f_2, \Lambda, \sqrt{3}) = 0.94680$			
Average $\widehat{\delta}$		MSE	
PLLR	H–S	PLLR	H–S
1.3805	−0.0558	0.31468	1.0059
1.3607	−0.0549	0.30241	1.0039
1.3674	−0.0559	0.30036	1.0060
1.3561	−0.0541	0.29173	1.0023
1.3932	−0.0548	0.32254	1.0039

bandwidth of the estimator is obtained by minimising the MSE (Härdle et al., 1992). As an initial point of this optimisation we take $h = T^{-2/7}$, which is the optimal bandwidth for $d = 1$. We use the Gaussian kernel to estimate $\widehat{f}_h(\mathbf{x})$ and $\widehat{Df}_h(\mathbf{x})$. The results are given below in Tables 1–6.

As expected our PLLR estimator $\widehat{\boldsymbol{\delta}}_{\mathrm{a}}$ performs better than $\widehat{\boldsymbol{\delta}}_{\mathrm{hs}}$ in this simulation, the reason being that model violates the conditions for asymptotic normality of $\widehat{\boldsymbol{\delta}}_{\mathrm{hs}}$, specifically the smoothness criteria of $f_1(\mathbf{x})$. Comparing the results for $f_2(\mathbf{x})$ (Tables 1, 3 and 5) against $f_1(\mathbf{x})$ (Tables 2, 4 and 6) we notice our estimator performs the better when there are more discontinuities (vertices) in the covariate density. The choice of error distribution does not seem to matter that much.

We have also done a simulation study for a version of the "Sine" model (Härdle, 1992) comparing H–S estimator with the PLLR. We obtain similar results as above.[5] We conclude that our estimator complements the kernel based H–S estimator.

5. Conclusion

The paper proposes an simple method of estimating the ADE when the density of the covariates are bounded and discontinuous. We propose to estimate the AD by a weighted average of piecewise local OLS slopes denoted by $\widehat{\boldsymbol{\delta}}_{\mathrm{a}}$.

We establish the asymptotic normality of our estimator under regularity conditions similar (but not same) to those of Härdle–Stoker. These assumptions are similar but not the same as the assumptions under which proved the (asymptotic) normality of their ADE. The H–S estimator requires some smoothness conditions on the density of explanatory variable(s) $f(\mathbf{x})$ but our method does not require such assumptions though we need $f(\mathbf{x})$ to have compact support. It might be worthwhile to point out that by not requiring the density $f(\mathbf{x})$ to vanish, our method can be used to test for linearity or stability by dividing the data into different regions and calculating the ADE of each region and testing for equality like a Chow test. We also provide a consistent estimator of the asymptotic variance of $\widehat{\boldsymbol{\delta}}_{\mathrm{a}}$, which can be used for a Wald like test statistic.

In the special case when the regression function is linear, the asymptotic variance of $\widehat{\boldsymbol{\delta}}_{\mathrm{a}}$ coincides with the asymptotic variance of the classical least squares estimator of the slope. So in the case of linearity of the regression function we get the standard result. This implies that we will not lose efficiency when compared to the simple OLS method where the regression function is rightly specified as linear.

The method described, is applied to a censored regression model with bounded but discontinuous covariates, in presence of dummy variables. We simulate and compare the small sample results of H–S estimator with our estimator under various specifications covariate densities and error distributions. The results indicate that our estimator performs better than the H–S estimator when the discontinuities in the covariate density increase. Our method thus complements the H–S method for estimating the average derivative.

[5]The results are available from author's website or on request by e-mail.

Acknowledgements

I am grateful to John Aldrich, Federico Martellosio, Ashoke Kr. Sinha and two referees for their constructive comments. I thank Prof Wolfgang Härdle for comments on a previous version of this paper (Banerjee, 1994).

Appendix A. Useful lemmas and proof of theorems

Lemma 1. *If g is bounded, under the Assumptions* A(1)–A(4), *as $T \to \infty$, then we have*

1. $$\sup_r \left\| \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{x}_t) I_{t,r} - p_r \boldsymbol{\mu}_g \right\| \xrightarrow{P} 0.$$

2. $$\sup_r \left\| \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{x}_t) I_{t,r} u_t \right\| \xrightarrow{P} 0,$$

where $\boldsymbol{\mu}_g = E(g(\mathbf{x}))$.

Proof of Lemma 1. Observe that, if $\mathbf{M}_r, (1 \leqslant r \leqslant k)$ are a collection of independent random variables then,

$$\Pr\left\{ \sup_r \|\mathbf{M}_r\| > \varepsilon \right\} = 1 - \prod_{r=1}^{k} (1 - \Pr(\|\mathbf{M}_r\| > \varepsilon))$$

so

$$\Pr\left\{ \sup_r \|\mathbf{M}_r\| > \varepsilon \right\} \to 0 \quad \text{iff} \quad \prod_{r=1}^{k} (1 - \Pr(\|\mathbf{M}_r\| > \varepsilon)) \to 1,$$

$$\text{iff} \quad \sum_{r=1}^{k} \Pr(\|\mathbf{M}_r\| > \varepsilon) \to 0,$$

$$\text{iff} \quad \sum_{r=1}^{k} E\|\mathbf{M}_r\|^2 \to 0, \quad \text{(using Chebyshev's inequality).} \tag{A.1}$$

Thus, if $\sum_{r=1}^{k} E\|\mathbf{M}_r\|^2 \to 0$, then $\sup_r \|\mathbf{M}_r\| \xrightarrow{P} 0$.

(1) Let $\mathbf{M}_r = \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{x}_t) I_{t,r} - p_r \mu_g$, so

$$E\|\mathbf{M}_r\|^2 = E \left\| \frac{1}{T} \sum_{t=1}^{T} (g(\mathbf{x}_t) I_{t,r} - p_r \boldsymbol{\mu}_g) \right\|^2$$

$$\leqslant \frac{1}{T^2} E\left\| \sum_{t=1}^{T} g(\mathbf{x}_t)(I_{t,r} - p_r) \right\|^2 + p_r^2 E\left\| \frac{1}{T} \sum_{t=1}^{T} (g(\mathbf{x}_t) - \boldsymbol{\mu}_g) \right\|^2$$

$$\leqslant \frac{1}{T^2} E\left(\sum_{t=1}^{T} \|g(\mathbf{x}_t)\| |I_{t,r} - p_r| \right)^2 + p_r^2 E\left\| \frac{1}{T} \sum_{t=1}^{T} (g(\mathbf{x}_t) - \boldsymbol{\mu}_g) \right\|^2.$$

Since $\mathbf{x}$ is bounded and $E(g(\mathbf{x}_t)) = \boldsymbol{\mu}_g$, we have

$$E\|\mathbf{M}_r\|^2 \leqslant \frac{1}{T^2} C_1 E\left(\sum_{t=1}^{T} |I_{t,r} - p_r| \right)^2 + C_2 \frac{p_r^2}{T}, \tag{A.2}$$

where C_1 and C_2 are constants. As $\mathbf{x}_t$'s are independent and $I_{t,r}$ are Bernoulli random variables with parameter p_r, the first term in (A.2) can be written as

$$\begin{aligned} E\left(\sum_{t=1}^{T} |I_{t,r} - p_r| \right)^2 &= \frac{1}{T^2} \left[\sum_{t=1}^{T} E(I_{t,r} - p_r)^2 + 2 \sum_{t<s=1}^{T} E|I_{t,r} - p_r| E|I_{s,r} - p_r| \right] \\ &= \frac{1}{T^2} \left[T p_r (1 - p_r) + 2 \binom{T}{2} (2 p_r (1 - p_r))^2 \right] \\ &\leqslant Const \left(\frac{1}{T} p_r + p_r h \right). \end{aligned} \tag{A.3}$$

Therefore combining (A.2) and (A.3) we get,

$$\begin{aligned} \sum_{r=1}^{k} E\|\mathbf{M}_r\|^2 &\leqslant Const \sum_{r=1}^{k} \left(\frac{1}{T} p_r + p_r h + \frac{p_r}{T} h \right) \\ &\leqslant Const \left(\frac{1}{T} + h + \frac{1}{T} h \right). \end{aligned}$$

Hence as $T \to \infty$, the expression above goes to zero since $\sqrt{T} h \to 0$.

(2) Let $\mathbf{M}_r = \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{x}_t) I_{t,r} u_t$, so

$$\begin{aligned} E\|\mathbf{M}_r\|^2 &= E\left\| \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{x}_t) I_{t,r} u_t \right\|^2 \\ &= \frac{1}{T^2} E\left[E\left(\left\| \sum_{t=1}^{T} g(\mathbf{x}_t) I_{t,r} u_t \right\|^2 \middle| \mathbf{x} \right) \right] \\ &\leqslant \frac{1}{T^2} E\left[\sum_{t=1}^{T} \|g(\mathbf{x}_t)\|^2 I_{t,r} \sigma_u^2 \right] (\text{since } u_t\text{'s are independent}) \\ &\leqslant \frac{1}{T^2} \, Const \sum_{t=1}^{T} E I_{t,r} \\ &\leqslant \frac{1}{T} \, Const \, p_r. \end{aligned}$$

Therefore we get

$$\sum_{r=1}^{k} E\|\mathbf{M}_r\|^2 \leqslant Const \sum_{r=1}^{k} \frac{1}{T} p_r = Const \frac{1}{T},$$

hence the proof. □

Corollary 1. *It follows from Lemma* 1,

1. $\sup_r \|w_r - p_r\| \xrightarrow{P} 0$,
2. $\sup_r \|\overline{\mathbf{x}}_r - \boldsymbol{\mu}_1\| \xrightarrow{P} 0$,
3. $\sup_r \|\mathbf{S}_x^r - \boldsymbol{\Sigma}\| \to 0$,
4. $\sup_r \|[\mathbf{S}_x^r]^{-1} - \boldsymbol{\Sigma}^{-1}\| \xrightarrow{P} 0$,

where $p_r = E(I_{t,r})$, $\boldsymbol{\Sigma} = Var(\mathbf{x})$ *and* $\boldsymbol{\mu}_1 = E(\mathbf{x})$.

Proof of Corollary 1.

1. Since we can write $w_r = \frac{1}{T}\sum_{t=1}^{T} I_{t,r}$, the proof follows from Lemma 1, part (1).
2. We can write $\overline{\mathbf{x}}_r = \frac{1}{w_r}\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t I_{t,r}$ then the proof follows from the Lemma 1 part (1), that $\sup_r \|\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t I_{t,r} - p_r \boldsymbol{\mu}_1\| \xrightarrow{P} 0$ and $\sup_r \|w_r - p_r\| \xrightarrow{P} 0$.
3. Note the we can write,

$$\mathbf{S}_x^r = \frac{1}{w_r T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' I_{t,r} - \left(\frac{1}{w_r T}\sum_{t=1}^{T} \mathbf{x}_t I_{t,r}\right)\left(\frac{1}{w_r T}\sum_{t=1}^{T} \mathbf{x}_t' I_{t,r}\right). \tag{A.4}$$

   Then using Lemma 1 part (1), we have

$$\sup_r \left\| \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' I_{t,r} - p_r \boldsymbol{\mu}_2 \right\| \xrightarrow{P} 0, \quad \sup_r \left\| \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t I_{t,r} - p_r \boldsymbol{\mu}_1 \right\| \xrightarrow{P} 0$$

   and

$$\sup_r \left\| w_r - p_r \right\| \xrightarrow{P} 0,$$

   where $\boldsymbol{\mu}_2 = E(\mathbf{x}\mathbf{x}')$ and $\boldsymbol{\mu}_1 = E(\mathbf{x})$. This implies

$$\sup_r \|\mathbf{S}_x^r - (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \boldsymbol{\mu}_1')\| \xrightarrow{P} 0. \tag{A.5}$$

4. Follows from the fact that $\Sigma = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \boldsymbol{\mu}_1')$ is a positive definite matrix. □

Lemma 2. *Under our Assumptions* A(1)–A(4) *as* $T \to \infty$,

$$\sqrt{T}\left\| \sum_{r=1}^{k} w_r \mathbf{W}_r I\{T_r \leqslant l\} \right\| \xrightarrow{P} 0,$$

where $\mathbf{W}_r, 1 \leqslant r \leqslant k$ *are uniformly bounded random variables.*

Proof of Lemma 2. Notice

$$\left\|\sum_{r=1}^{k} w_r \mathbf{W}_r I\{T_r \leqslant l\}\right\| \leqslant \sup_r \|\mathbf{W}_r\| \left\|\sum_{r=1}^{k} w_r I\{T_r \leqslant l\}\right\| \leqslant \overline{W}\left\|\sum_{r=1}^{k} w_r I\{T_r \leqslant l\}\right\|,$$

where $\sup_r \|\mathbf{W}_r\| \leqslant \overline{W}$. Therefore it is enough to prove $\|\sum_{r=1}^{k} w_r I\{T_r \leqslant l\}\| \xrightarrow{P} 0$. Since the random variable $\sum_{r=1}^{k} w_r I\{T_r \leqslant l\}$ is positive it is enough to show by Markov inequality that

$$\sqrt{T} E\left(\sum_{r=1}^{k} w_r I\{T_r \leqslant l\}\right) \to 0. \tag{A.6}$$

Since $T_r \sim \text{BinomialDist}(l; T, p_r)$,

$$\begin{aligned}\sqrt{T} E\left(\sum_{r=1}^{k} w_r I\{T_r \leqslant l\}\right) &= \sqrt{T}\sum_{r=1}^{k}\sum_{j=0}^{l} \frac{j}{T}\Pr(T_r = j) \\ &= \sqrt{T}\sum_{r=1}^{k}\sum_{j=0}^{l} \frac{j}{T}\binom{T}{j} p_r^j (1-p_r)^{T-j} \\ &= \sum_{r=1}^{k} p_r \sum_{j=1}^{l} \sqrt{T}\binom{T-1}{j-1} p_r^{j-1}(1-p_r)^{T-j}.\end{aligned}$$

Notice that using the inequalities $\binom{T}{j} \leqslant T^l$ for all j and $(1-v)^{\alpha} \leqslant \exp(-\alpha v)$, we have

$$\sqrt{T}\binom{T-1}{j-1} p_r^{j-1}(1-p_r)^{T-j} \leqslant Const\ T^{l-\frac{1}{2}} \exp(-T p_r).$$

Since $0 < \underline{c} < f(x)$, we have $\exp(-T p_r) \leqslant \exp(-\underline{c}\, Th)$. Hence

$$\begin{aligned}\sqrt{T} E\left(\sum_{r=1}^{k} w_r I\{T_r \leqslant l\}\right) &\leqslant Const \sum_{r=1}^{k} p_r T^{l-\frac{1}{2}} \exp(-\underline{c}\, Th) \\ &= Const \exp\left[\left((l-0.5) - \underline{c}\frac{Th}{\ln T}\right)\ln T\right] \to 0\end{aligned}$$

as $Th/\ln T \to \infty$, (Assumption A(4)). Then $\sqrt{T} E(\sum_{r=1}^{k} w_r I\{T_r \leqslant l\}) \to 0$. □

Lemma 3. *Under our Assumptions* A(1)–A(4) *if g is a differentiable function such that* $\|D[g(\cdot)]\| \leqslant C_g$, *then*

$$g(\overline{\mathbf{x}}_r) = \frac{1}{T_r}\sum_{t=1}^{T} g(\mathbf{x}_t) I_{t,r} + E_r^{(g)},$$

where $\sup_r \sqrt{T}\|E_r^{(g)}\| \to 0$.

Proof of Lemma 3. Note that,

$$\|E^{(g)}\| = \left\| \frac{1}{T_r} \sum_{t=1}^{T} (g(\mathbf{x}_t) - g(\overline{\mathbf{x}}_r)) I_{t,r} \right\|$$
$$\leqslant \frac{1}{T_r} \sum_{t=1}^{T} I_{t,r} \|D[g(\xi_{t,r})]\| \|\mathbf{x}_t - \overline{\mathbf{x}}_r\|,$$

where $\xi_{t,r}$ is within $\mathbf{x}_t$ and $\overline{\mathbf{x}}_r$. As $\|D[g(\cdot)]\|$ is bounded and $\|\mathbf{x}_t - \overline{\mathbf{x}}_r\| I_{t,r} \leqslant h$, we have

$$\|E^{(g)}\| \leqslant C_g h.$$

Hence $\sup_r \sqrt{T} \|R_r^{(g)}\| \to 0$ as $\sqrt{T}h \to 0$ (by Assumption A(4)). □

Lemma 4. *Under our Assumptions* A(1)–A(4),

$$\widehat{\boldsymbol{\beta}}_r = D[m(\overline{\mathbf{x}}_r)] + \boldsymbol{\theta}_r + R_r^{(1)},$$

where

$$\boldsymbol{\theta}_r = [\mathbf{S}_x^r]^{-1} \mathbf{S}_{xu}^r \quad \textit{and} \quad \sup_r \sqrt{T} R_r^{(1)} \xrightarrow{P} 0$$

such that $\mathbf{S}_{xu}^r = \frac{1}{T_r} \sum_{t-1}^{T} (\mathbf{x}_t - \overline{\mathbf{x}}_r) I_{t,r} u_t$.

Proof of Lemma 4. Taking a Taylor series expansion of $m(\mathbf{x}_t)$ around $\overline{\mathbf{x}}_r$ for those $\mathbf{x}_t$'s which are in $\mathbf{H}_r$, we have,

$$m(\mathbf{x}_t) = m(\overline{\mathbf{x}}_r) + \widetilde{\mathbf{x}}_{t,r}' D[m(\overline{\mathbf{x}}_r)] + \varepsilon_{t,r}, \tag{A.7}$$

where $\varepsilon_{t,r} = \frac{1}{2} \widetilde{\mathbf{x}}_{t,r} D^2[m(\xi_{t,r})] \widetilde{\mathbf{x}}_{t,r} I_{t,r}$ for some $\boldsymbol{\xi}_{t,r}$ between $\mathbf{x}_t$ and $\overline{\mathbf{x}}_r$ and $\widetilde{\mathbf{x}}_{t,r} = (\mathbf{x}_t - \overline{\mathbf{x}}_r) I_{t,r}$. Note that, since $\|\widetilde{\mathbf{x}}_{t,r}\| = \|(\mathbf{x}_t - \overline{\mathbf{x}}_r) I_{t,r}\| \leqslant h$, and $D^2[m(\boldsymbol{\xi}_{tr})]$ is bounded we have

$$\sup_r \|\varepsilon_{t,r}\| \leqslant Const\, h^2.$$

Then for all $\mathbf{x}_t$'s which are in $\mathbf{H}_r$ we can write

$$y_t = m(\overline{\mathbf{x}}_r) + \widetilde{\mathbf{x}}_{t,r}' D[m(\overline{\mathbf{x}}_r)] + \varepsilon_{t,r} + u_t,$$

using (A.7). This implies we can write the estimate of slope in the bin $\mathbf{H}_r$ as,

$$\widehat{\boldsymbol{\beta}}_r = [\mathbf{S}_x^r]^{-1} \frac{1}{T_r} \sum_{t=1}^{T} \widetilde{\mathbf{x}}_{t,r} y_t,$$
$$= [\mathbf{S}_x^r]^{-1} \frac{1}{T_r} \sum_{t=1}^{T} \widetilde{\mathbf{x}}_{t,r} m(\mathbf{x}_t) + [\mathbf{S}_x^r]^{-1} \mathbf{S}_{xu}^r$$
$$= [\mathbf{S}_x^r]^{-1} \left(\frac{1}{T_r} \sum_{t=1}^{T} \widetilde{\mathbf{x}}_{t,r} m(\overline{\mathbf{x}}_r) + \frac{1}{T_r} \sum_{t=1}^{T} \widetilde{\mathbf{x}}_{t,r} \widetilde{\mathbf{x}}_{t,r}' D[m(\overline{\mathbf{x}}_r)] + \frac{1}{T_r} \sum_{t=1}^{T} \widetilde{\mathbf{x}}_{t,r} \varepsilon_{t,r} \right) + \boldsymbol{\theta}_r$$
$$= D[m(\overline{\mathbf{x}}_r)] + R_r^{(1)} + \boldsymbol{\theta}_r,$$

where $R_r^{(1)} = [\mathbf{S}_x^r]^{-1} \frac{1}{T_r} \sum_{t=1}^{T} \widetilde{\mathbf{x}}_{t,r} \varepsilon_{t,r}$.

Now note that by Corollary 1 part (4), $[\mathbf{S}_x^r]^{-1}$ is uniformly bounded in r and

$$\sqrt{T}\left\|\frac{1}{T_r}\sum_{t=1}^{T}\widetilde{\mathbf{x}}_{t,r}\varepsilon_{t,r}\right\| \leqslant \sqrt{T}\,Const\,h^2\,\frac{1}{T_r}\sum_{t=1}^{T}\|(\mathbf{x}_t-\overline{\mathbf{x}}_r)\|I_{t,r} \leqslant Const\,\sqrt{T}h^3.$$

As $\sqrt{T}h \to 0, \sup_r\sqrt{T}\|\frac{1}{T_r}\sum_{t=1}^{T}\widetilde{\mathbf{x}}_{t,r}\varepsilon_{t,r}\| \to 0$. Hence $\sup_r\sqrt{T}\|R_r^{(1)}\| \xrightarrow{P} 0$. □

Lemma 5. *Under our Assumptions* A(1)–A(4),

1. $\|\sqrt{T}\sum_{r=1_P}^{k} w_r\boldsymbol{\theta}_r - \mathbf{Z}\| \xrightarrow{D} 0$, *where* $\mathbf{Z}\sim N(0,\sigma_u^2\Sigma^{-1})$.
2. $\sup_r\|\boldsymbol{\theta}_r\| \to 0$.

Proof of Lemma 5. (1) Notice that

$$\begin{aligned}\sqrt{T}\sum_{r=1}^{k} w_r\mathbf{S}_{xu}^r &= \sqrt{T}\sum_{r=1}^{k} w_r\frac{1}{T_r}\sum_{t=1}^{T}(\mathbf{x}_t-\overline{\mathbf{x}}_r)I_{t,r}u_t\\ &= \frac{1}{\sqrt{T}}\sum_{r=1}^{k}\sum_{t=1}^{T}(\mathbf{x}_t-\overline{\mathbf{x}}_r)I_{t,r}u_t\\ &= \frac{1}{\sqrt{T}}\sum_{r=1}^{k}\sum_{t=1}^{T}I_{t,r}u_t - \frac{1}{\sqrt{T}}\sum_{r=1}^{k}\sum_{t=1}^{T}(\overline{\mathbf{x}}_r-\boldsymbol{\mu}_1)I_{t,r}u_t\\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\mathbf{x}_t-\boldsymbol{\mu}_1)I_{t,r}u_t + R\left(\text{since }\sum_{r=1}^{k}I_{t,r}=1\right),\end{aligned}$$

where $R = \frac{1}{\sqrt{T}}\sum_{r=1}^{k}\sum_{t=1}^{T}(\overline{\mathbf{x}}_r-\boldsymbol{\mu}_1)I_{t,r}u_t$. The first term

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\mathbf{x}_t-\boldsymbol{\mu}_1)I_{t,r}u_t \xrightarrow{D} \mathbf{Z}_1\sim N(0,\sigma_u^2\Sigma)$$

by the central limit theorem. Consider the residual term

$$\begin{aligned}\|R\| &\leqslant \sup_r\|\overline{\mathbf{x}}_r-\boldsymbol{\mu}_1\|\frac{1}{\sqrt{T}}\sum_{r=1}^{k}\sum_{t=1}^{T}I_{t,r}u_t\\ &= \sup_r\|\overline{\mathbf{x}}_r-\boldsymbol{\mu}_1\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}u_t.\end{aligned}$$

Since $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}u_t$ is bounded and by Corollary 1 part (2) $\sup_r\|\overline{\mathbf{x}}_r-\boldsymbol{\mu}_1\| \xrightarrow{P} 0$, we have $\|R\| \xrightarrow{P} 0$. Hence

$$\sqrt{T}\sum_{r=1}^{k} w_r\mathbf{S}_{xu}^r \xrightarrow{D} N(0,\sigma_u^2\Sigma). \tag{A.8}$$

Since

$$\sqrt{T}\sum_{r=1}^{k} w_r\boldsymbol{\theta}_r = \boldsymbol{\Sigma}^{-1}\sqrt{T}\sum_{r=1}^{k} w_r\mathbf{S}_{xu}^r + \sqrt{T}\sum_{r=1}^{k} w_r([\mathbf{S}_x^r]^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{S}_{xu}^r \tag{A.9}$$

from Corollary 1 part (4) we have $\sup_r\|[\mathbf{S}_x^r]^{-1} - \boldsymbol{\Sigma}^{-1}\| \xrightarrow{P} 0$. Combining with (A.8) gives us,

$$\left\|\sqrt{T}\sum_{r=1}^{k} w_r([\mathbf{S}_x^r]^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{S}_{xu}^r\right\| \leqslant \sup_r \|[\mathbf{S}_x^r]^{-1} - \boldsymbol{\Sigma}^{-1}\|\sqrt{T}\sum_{r=1}^{k} w_r\mathbf{S}_{xu}^r \xrightarrow{P} 0.$$

Therefore $\sup_r\|\sqrt{T}\sum_{r=1}^{k} w_r\boldsymbol{\theta}_r - N(0, \sigma_u^2\boldsymbol{\Sigma}^{-1})\| \xrightarrow{D} 0$.

(2) Notice that

$$\boldsymbol{\theta}_r = [\mathbf{S}_x^r]^{-1}\frac{1}{Tw_r}\sum_{t=1}^{T}(\mathbf{x}_t - \overline{\mathbf{x}}_r)I_{t,r}u_t.$$

Using Lemma 1 part (2) $\sup_r\|\frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_t - \overline{\mathbf{x}}_r)I_{t,r}u_t\| \xrightarrow{P} 0$, from Corollary 1 part (1) $\sup_r\|w_r - p_r\| \xrightarrow{P} 0$ and Corollary 1 part (4) we have $\sup_r\|[\mathbf{S}_x^r]^{-1} - \boldsymbol{\Sigma}^{-1}\| \xrightarrow{P} 0$. Hence the proof. □

Proof of Theorem 1. Notice that we can write using Lemma 4

$$\sqrt{T}\sum_{r=1}^{k} w_r\widehat{\boldsymbol{\beta}}_r = \sqrt{T}\sum_{r=1}^{k} w_r(D[m(\overline{\mathbf{x}}_r)] + \boldsymbol{\theta}_r + R_r^{(1)})$$

and therefore

$$\begin{aligned}\sqrt{T}\left\|\widehat{\boldsymbol{\delta}}_{\mathrm{a}} - \sum_{r=1}^{k} w_r\widehat{\boldsymbol{\beta}}_r\right\| &= \sqrt{T}\left\|\sum_{r=1}^{k} w_r\widehat{\boldsymbol{\beta}}_r I\{T_r < l\}\right\| \\ &= \sqrt{T}\left\|\sum_{r=1}^{k} w_r(D[m(\overline{\mathbf{x}}_r)] + \boldsymbol{\theta}_r + R_r^{(1)})I\{T_r < l\}\right\|\end{aligned}$$

by Lemma 5. Since $D[m(\overline{\mathbf{x}}_r)]$ is bounded by assumption and $R_r^{(1)}$ and $\boldsymbol{\theta}_r$ are bounded by Lemmas 4 and 5 we have

$$\sqrt{T}\left\|\widehat{\boldsymbol{\delta}}_{\mathrm{a}} - \sum_{r=1}^{k} w_r\widehat{\boldsymbol{\beta}}_r\right\| \xrightarrow{P} 0. \tag{A.10}$$

Since by Lemma 2, $\|D[m(\overline{\mathbf{x}}_r)] - \frac{1}{T_r}\sum_{t=1}^{T} D[m(\mathbf{x}_t)]I_{t,r}\| \xrightarrow{P} 0$, we can write,

$$\begin{aligned}\sqrt{T}\sum_{r=1}^{k} w_r\widehat{\boldsymbol{\beta}}_r &= \sqrt{T}\sum_{r=1}^{k} w_r\left(\frac{1}{T_r}\sum_{t=1}^{T} D[m(\mathbf{x}_t)]I_{t,r} + \boldsymbol{\theta}_r + R_r^{(1)} + R_r^{(2)}\right) \\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} D[m(\mathbf{x}_t)] + \sqrt{T}\sum_{r=1}^{k} w_r\boldsymbol{\theta}_r + \sqrt{T}\sum_{r=1}^{k} w_r(R_r^{(1)} + R_r^{(2)}),\end{aligned}$$

where $\sup_r \sqrt{T} R_r^{(i)} \xrightarrow{P} 0$, $i = 1, 2$. Therefore we have by (A.10),

$$\left\| \sqrt{T}(\widehat{\boldsymbol{\delta}}_{\mathrm{a}} - \boldsymbol{\delta}) - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (D[m(\mathbf{x}_t)] - \boldsymbol{\delta}) - \sqrt{T} \sum_{r=1}^{k} w_r \boldsymbol{\theta}_r \right\| \xrightarrow{P} 0.$$

By central limit theorem we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (D[m(\mathbf{x}_t)] - \boldsymbol{\delta}) \xrightarrow{D} N(0, Var(D[m(\mathbf{x})]))$$

and by Lemma 5

$$\sqrt{T} \sum_{r=1}^{k} w_r \boldsymbol{\theta}_r \xrightarrow{D} \mathbf{Z}.$$

Also since $\mathbf{x}$ and u are independent, we prove that

$$\sqrt{T}(\widehat{\boldsymbol{\delta}}_{\mathrm{a}} - \boldsymbol{\delta}) \xrightarrow{D} N(0, Var(D[m(\mathbf{x})]) + \sigma_u^2 \boldsymbol{\Sigma}^{-1}).$$

Lemma 6. *Under our Assumptions* A(1)–A(4) *and* $E(u^4) < \infty$, *we have*

$$\sup_r \left\| \frac{1}{T_r} \sum_{t=1}^{T} u_t^2 I_{r,t} - \sigma_u^2 \right\| \xrightarrow{P} 0.$$

Proof of Lemma 6. Following the proof in Lemma 1, let $\mathbf{M}_r = \frac{1}{T} \sum_{t=1}^{T} u_t^2 (I_{r,t} - p_r \sigma_u^2)$, then,

$$\begin{aligned} E\left\| \frac{1}{T} \sum_{t=1}^{T} u_t^2 (I_{r,t} - p_r \sigma_u^2) \right\|^2 &= \frac{1}{T^2} E\left[E\left(\left\| \sum_{t=1}^{T} (I_{t,r} - p_r) u_t^2 \right\|^2 \middle| \mathbf{x} \right) \right] \\ &\leqslant Const \, \frac{1}{T^2} \left[\sum_{t=1}^{T} E(I_{t,r} - p_r)^2 E(u^4) + 2 \sum_{t<s=1}^{T} E|I_{t,r} - p_r| E|I_{s,r} - p_r| [\sigma_u^2]^2 \right] \\ &\leqslant Const \left(\frac{1}{T} p_r + p_r h + \frac{p_r}{T} h \right) \end{aligned}$$

hence

$$\sum_{r=1}^{k} E\left\| \frac{1}{T} \sum_{t=1}^{T} u_t^2 (I_{r,t} - p_r \sigma_u^2) \right\|^2 \to 0$$

and therefore $\sup_r \| \frac{1}{T} \sum_{t=1}^{T} u_t^2 I_{r,t} - p_r \sigma_u^2 \| \xrightarrow{P} 0$. Also by Corollary 1 part (1) we have $\sup_r \| w_r - p_r \| \xrightarrow{P} 0$. Hence the proof. □

Proof of Theorem 2. Note that using Taylor series expansion as in (A.7) we can write,

$$y_t = m(\overline{\mathbf{x}}_r) + \widetilde{\mathbf{x}}_{t,r}' D[m(\overline{\mathbf{x}}_r)] + \varepsilon_{t,r} + u_t.$$

Therefore

$$\overline{y}_r = \frac{1}{T_r}\sum_{t=1}^{T} y_t I_{r,t} = m(\overline{\mathbf{x}}_r) + \overline{\varepsilon}_r + \overline{u}_r,$$

where $\overline{u}_r = \frac{1}{T_r}\sum_{t=1}^{T} u_t I_{r,t}$ and $\overline{\varepsilon}_r = \frac{1}{T_r}\sum_{t=1}^{T} \varepsilon_{t,r} I_{r,t}$. Also by Lemma 4

$$\widetilde{\mathbf{x}}_{t,r}'\widehat{\boldsymbol{\beta}}_r = \widetilde{\mathbf{x}}_{t,r}' D[m(\overline{\mathbf{x}}_r)] + \widetilde{\mathbf{x}}_{t,r}'\boldsymbol{\theta}_r + \widetilde{\mathbf{x}}_{t,r}' R_r^{(1)},$$

where $R_r^{(1)} = [\mathbf{S}_x^r]^{-1}\frac{1}{T_r}\sum_{t=1}^{T}\widetilde{\mathbf{x}}_{t,r}\varepsilon_{t,r}$. Hence

$$\begin{aligned}\widehat{u}_{r,t} &= y_t - \overline{y}_r - (\mathbf{x}_t - \overline{\mathbf{x}}_r)'\widehat{\boldsymbol{\beta}}_r \\ &= (u_t - \overline{u}_r - \widetilde{\mathbf{x}}_{t,r}'\boldsymbol{\theta}_r) + \varepsilon_r^{(1)},\end{aligned}$$

where

$$\varepsilon_r^{(1)} = \left(\varepsilon_{t,r} - \frac{1}{T_r}\sum_{t=1}^{T}(1 - \widetilde{\mathbf{x}}_{t,r}'[\mathbf{S}_x^r]^{-1}\widetilde{\mathbf{x}}_{t,r})\varepsilon_{t,r} I_{r,t}\right).$$

Note that

$$\begin{aligned}\sup_r\left\|\varepsilon_r^{(1)}\right\| &= \sup_r\left\|\varepsilon_{t,r} - \frac{1}{T_r}\sum_{t=1}^{T}(1 - \widetilde{\mathbf{x}}_{t,r}'[\mathbf{S}_x^r]^{-1}\widetilde{\mathbf{x}}_{t,r})\varepsilon_{t,r} I_{r,t}\right\| \\ &\leqslant Const\, h^2\left(1 + \frac{1}{T_r}\sum_{t=1}^{T}\|1 - \widetilde{\mathbf{x}}_{t,r}'[\mathbf{S}_x^r]^{-1}\widetilde{\mathbf{x}}_{t,r}\| I_{r,t}\right) \leqslant Const\, h^2.\end{aligned}$$

By Lemmas 4 and 1 we have $\sup_r\|\boldsymbol{\theta}_r\| \xrightarrow{P} 0, \sup_r\|\overline{u}_r\| \xrightarrow{P} 0$ and $\sup_r\|\mathbf{S}_x^r - \boldsymbol{\Sigma}\| \xrightarrow{P} 0$, hence

$$\sup_r\|\overline{u}_r^2 + \boldsymbol{\theta}_r'\mathbf{S}_x^r\boldsymbol{\theta}_r\| \xrightarrow{P} 0. \tag{A.11}$$

This implies $(u_t - \overline{u}_r - \widetilde{\mathbf{x}}_{t,r}'\boldsymbol{\theta}_r)$ is uniformly bounded in r and

$$\sup_r\|\widehat{u}_{r,t}^2 - (u_t - \overline{u}_r - \widetilde{\mathbf{x}}_{t,r}'\boldsymbol{\theta}_r)^2\| \leqslant Const\, h^2.$$

Therefore

$$\sup_r\left\|\frac{1}{T_r}\sum_{t=1}^{T}\widehat{u}_{r,t}^2 I_{r,t} - \frac{1}{T_r}\sum_{t=1}^{T}(u_t - \overline{u}_r - \widetilde{\mathbf{x}}_{t,r}'\boldsymbol{\theta}_r)^2 I_{r,t}\right\| \leqslant Const\, h^2. \tag{A.12}$$

Note that we can write,

$$\frac{1}{T_r}\sum_{t=1}^{T}(u_t - \overline{u}_r - \widetilde{\mathbf{x}}_{t,r}'\boldsymbol{\theta}_r)^2 I_{r,t} = \frac{1}{T_r}\sum_{t=1}^{T} u_t^2 I_{r,t} - \overline{u}_r^2 - \boldsymbol{\theta}_r'\mathbf{S}_x^r\boldsymbol{\theta}_r.$$

By Lemma 6 we get $\sup_r\|\frac{1}{T_r}\sum_{t=1}^{T} u_t^2 I_{r,t} - \sigma_u^2\| \xrightarrow{P} 0$. Thus combining with (A.11) we obtain $\sup_r\|s_r^2 - \sigma_u^2\| \xrightarrow{P} 0$, which implies

$$s_{\mathrm{a}}^2 = \sum_{r=1}^{k} w_r s_r^2 - \sum_{r=1}^{k} w_r s_r^2 I\{T_r < l\} \to \sigma_u^2$$

by Lemma 2. □

Proof of Theorem 3. Using Lemma (4) we have

$$\widehat{\boldsymbol{\beta}}_r\widehat{\boldsymbol{\beta}}_r' = D[m(\overline{\mathbf{x}}_r)]D[m(\overline{\mathbf{x}}_r)]' + 2D[m(\overline{\mathbf{x}}_r)](\boldsymbol{\theta}_r + R_r^{(1)})' + (\boldsymbol{\theta}_r + R_r^{(1)})(\boldsymbol{\theta}_r + R_r^{(1)})'.$$

By Lemmas 4 and 5 we have $\sup_r \|R_r^{(1)} + \boldsymbol{\theta}_r\| \xrightarrow{P} 0$, and by Assumption $D[m(\overline{\mathbf{x}}_r)]$ is bounded, therefore

$$\widehat{\boldsymbol{\beta}}_r\widehat{\boldsymbol{\beta}}_r' = D[m(\overline{\mathbf{x}}_r)]D[m(\overline{\mathbf{x}}_r)]' + R_r^{(4)}, \tag{A.13}$$

where $\sup_r \|R_r^{(4)}\| \xrightarrow{P} 0$. Also by Lemma 3 we can write,

$$D[m(\overline{\mathbf{x}}_r)]D[m(\overline{\mathbf{x}}_r)]' = \frac{1}{T_r}\sum_{t=1}^{T} D[m(\mathbf{x}_t)]D[m(\mathbf{x}_t)]I_{t,r} + R_r^{(5)}, \tag{A.14}$$

where $\sup_r \|R_r^{(5)}\| \xrightarrow{P} 0$. Combining (A.13) and (A.14) we get

$$\begin{aligned}\sum_{r=1}^{k} w_r\widehat{\boldsymbol{\beta}}_r\widehat{\boldsymbol{\beta}}_r' &= \sum_{r=1}^{k} w_r \frac{1}{T_r}\sum_{t=1}^{T} D[m(\mathbf{x}_t)]D[m(\mathbf{x}_t)]'I_{t,r} + \sum_{r=1}^{k} w_r(R_r^{(4)} + R_r^{(5)}) \\ &= \frac{1}{T}\sum_{t=1}^{T} D[m(\mathbf{x}_t)]D[m(\mathbf{x}_t)]' + \sum_{r=1}^{k} w_r(R_r^{(4)} + R_r^{(5)}).\end{aligned}$$

By weak law of large numbers we have,

$$\frac{1}{T}\sum_{t=1}^{T} D[m(\mathbf{x}_t)]D[m(\mathbf{x}_t)]' \xrightarrow{P} E(D[m(\mathbf{x})]D[m(\mathbf{x})]').$$

We know from Theorem 1 $\widehat{\boldsymbol{\delta}}_{\mathrm{a}} \xrightarrow{P} E(D[m(\mathbf{x})])$, hence using Lemma 2 we have

$$\sum_{r=1}^{k} w_r\widehat{\boldsymbol{\beta}}_r\widehat{\boldsymbol{\beta}}_r' I\{T_r \geqslant l\} - \widehat{\boldsymbol{\delta}}_{\mathrm{a}}\widehat{\boldsymbol{\delta}}_{\mathrm{a}}' \xrightarrow{P} Var(D[m(\mathbf{x})]).$$

Also from Theorem 2 $s_{\mathrm{a}}^2 \xrightarrow{P} \sigma_u^2$ and by Corollary 1 part (4), $\sup_r \|[\mathbf{S}_x^r]^{-1} - \boldsymbol{\Sigma}^{-1}\| \xrightarrow{P} 0$, and using Lemma 2 we have

$$s_{\mathrm{a}}^2 \sum_{r=1}^{k} w_r[\mathbf{S}_x^r]^{-1} I\{T_r \geqslant l\} \xrightarrow{P} \sigma_u^2\boldsymbol{\Sigma}^{-1}.$$

Hence the proof of

$$\widehat{\mathbf{V}}_{\mathrm{a}} \xrightarrow{P} Var(D[m(\mathbf{x})]) + \sigma_u^2\boldsymbol{\Sigma}^{-1}.$$

References

Banerjee, A.N., 1994. A method of estimating the average derivative. CORE Discussion paper: 9403, Universite Catholique de Louvain.

Fan, J., Gijbels, I., 1992. Variable bandwidth and local linear regression smoothers. The Annals of Statistics 20, 2008–2036.

Greene, W., 1999. Marginal effects in the censored regression model. Economics Letters 64 (1), 43–49.
Härdle, W., 1992. Applied of Non Parametric Regression, Econometric Society Monographs. Cambridge University Press, Cambridge.
Härdle, W., Stoker, T.M., 1989. Investigating smooth multiple regression by the method of average derivatives. Journal of American Statistical Association 84 (408), 986–995.
Härdle, W., Hilderbrand, W., Jerison, M., 1991. Empirical evidence for the law of demand. Econometrica 59, 1525–1550.
Härdle, W., Hart, J., Marron, J.S., Tsybakov, A.B., 1992. Bandwidth choice for average derivative estimation. Journal of the American Statistical Association 87 (417), 218–226.
Horowitz, J.L., Härdle, W., 1996. Direct semiparametric estimation of a single-index model with discrete covariates. Journal of the American Statistical Association 91, 1632–1640.
Li, W., 1996. Asymptotic equivalence of estimators of average derivative. Economic Letters 52, 241–245.
Newey, W.K., Stoker, T.M., 1993. Efficiency of weighted average derivative estimators and index models. Econometrica 61, 1199–1223.
Powell, J., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.
Stoker, T.M., 1991a. Lectures on Semiparametric Econometrics, CORE Lecture Series. Core Foundation, Louvain-la Neuve.
Stoker, T.M., 1991b. Equivalance of direct, indirect and slope estimators of average derivatives. In: Barnett, W.A., Powell, J., Tauchen, G. (Eds.), Nonparametric and Semiparametric Methods in Econometrics and Statistics. Cambridge University Press, Cambridge.
Ullah, A., Vinod, H., 1988. Flexible production function estimation by nonparametric kernel estimators. Advances in Econometrics, JAI Press.
Zhang, S., Karuhamuni, R.J., 2000. On nonparametric density estimation at the boundary. Nonparametric Statistics 12, 197–221.