Supplementary Material 1
Review on Semiparametric Single Index Model

Hansen (2022, Sections 26.7-26.12)

Yasuyuki Matsumura (Kyoto University)

Last Updated: October 7, 2025

https://yasu0704xx.github.io

Introduction

- A semiparametric single index model is given by

$$Y = g(X^T \beta_0) + u,$$

where

$$\begin{aligned}
Y \in \mathbb{R} &: \text{a dependent variable,}\\
X \in \mathbb{R}^q &: \text{a } q \times 1 \text{ explanatory vector,}\\
\beta_0 \in \mathbb{R}^q &: \text{a } q \times 1 \text{ vector of unknown parameters,}\\
u \in \mathbb{R} &: \text{an error term which satisfies } \mathbb{E}(u \mid X) = 0,\\
g(\cdot) &: \text{an unknown distribution function.}
\end{aligned}$$

- Even though x is a $q \times 1$ vector, $x^T \beta_0$ is a scalar of a single linear combination, which is called a single index.
- By the form of the single index model, we obtain

$$\mathbb{E}(Y \mid X) = g(X^T \beta_0),$$

  which means that the conditional expectation of Y only depends on the vector X through a single index $X^T \beta_0$.
- The model is semiparametric when $\beta \in \mathbb{R}^q$ is estimated with the parametric methods and $g(\cdot)$ with the nonparametric methods.
  - If $g(\cdot)$ is the identity function, then the model turns out to be a linear regression model.
  - If $g(\cdot)$ is the CDF of Normal$(0, 1)$, then the model turns out to be a probit model.
  - If $g(\cdot)$ is the CDF of logistic distribution, then the model turns out to be a logistic regression model.

Textbooks

- Pagan and Ullah (1999, Chapter 7)
- Li and Racine (2007, Chapter 8)
- Horowitz (2009, Chapters 2 and 4)
- 西山・人見（2023，第 3 章）
- 末石（2024，第 4 章）

Contents

Identification

Identification Conditions

Proposition 8.1 (Identification of a Single Index Model)

For the semiparametric single index model $Y = g(x^T\beta_0) + u$, identification of β_0 and $g(\cdot)$ requires that

(i) x should not contain a constant/an intercept, and must contain at least one continuous variable. Moreover, $\|\beta_0\|=1$.

(ii) $g(\cdot)$ is differentiable and is not a constant function on the support of $x^T\beta_0$.

(iii) For the discrete components of x, varying the values of the discrete variables will not divide the support of $x^T\beta_0$ into disjoint subsets.

Identification Condition (i)

- Note that the location and the scale of β_0 are not identified.
- The vector x cannot include an intercept because the function $g(\cdot)$ (which is to be estimated in nonparametric manners) includes any location and level shift.
    - That is, β_0 cannot contain a location parameter.
- Some normalization criterion (scale restrictions) for β_0 are needed.
    - One approach is to set $\|\beta_0\| = 1$.
    - The second approach is to set one component of β_0 to equal one. This approach requires that the variable corresponding to the component set to equal one to be continuously distributed and has a non-zero coefficient.
    - Then, x must be dimension 2 or larger. If x is one-dimensional, then $\beta_0 \in \mathbb{R}^1$ is simply normalized to 1, and the model is the one-dimensional nonparametric regression $E(Y \mid x) = g(x)$ with no semiparametric component.

Identification Conditions (ii) and (iii)

- The function $g(\cdot)$ cannot be a constant function and must be differentiable on the support of $x^T\beta_0$.
- x must contain at least one continuously distributed variable and this continuous variable must have non-zero coefficient.
    - If not, $x^T\beta_0$ only takes a discrete set of values and it would be impossible to identify a continuous function $g(\cdot)$ on this discrete support.

Ichimura's (1993) Method

Ichimura's (1993) Method

- Suppose that the functional form of $g(\cdot)$ *were* known.
- Then we could estimate β_0 by minimizing the least-squares criterion:

$$\sum_{i=1} \left[Y_i - g(X_i^T \beta)\right]^2$$

  with respect to β.
- We could think about replacing $g(\cdot)$ with a nonparametric estimator $\widehat{g}(\cdot)$.
- However, since $g(z)$ is the conditional mean of Y_i given $X_i^T \beta_0 = z$, $g(\cdot)$ depends on unknown β_0. thus we cannot estimate $g(\cdot)$ here.

- Nevertheless, for a fixed value of β, we can estimate

$$G(X_i^T\beta) := \mathbb{E}(Y_i \mid X_i^T\beta) = \mathbb{E}(g(X_i^T\beta_0) \mid X_i^T\beta).$$

- In general $G(X_i^T\beta) \neq g(X_i^T\beta)$.
- When $\beta = \beta_0$, it holds that $G(X_i^T\beta_0) = g(X_i^T\beta_0)$.
- That is, conditioning on a general $X_i^T\beta$, G and g do not, in general, coincide, while they coincide when $X_i^T\beta = X_i^T\beta_0$.

Ichimura's (1993) Weighted Semiparametric Least Squares Estimation

- First, we estimate $G(X_i^T\beta)$ with the leave-one-out NW estimator:

$$\widehat{G}_{-i}(X_i^T\beta) \equiv \widehat{\mathbb{E}}_{-i}(Y_i \mid X_i^T\beta) = \frac{\sum_{j\neq i} Y_j K\left(\frac{X_j^T\beta - X_i^T\beta}{h}\right)}{\sum_{j\neq i} K\left(\frac{X_j^T\beta - X_i^T\beta}{h}\right)}.$$

- Second, using the leave-one-out NW estimator $\widehat{G}_{-i}(X_i^T\beta)$, we estimate β with

$$\widehat{\beta} \equiv \arg\min_{\beta} \sum_{i=1}^{n} \left[Y_i - \widehat{G}_{-i}(X_i^T\beta)\right]^2 w(X_i)\mathbf{1}(X_i \in A_n) \equiv \arg\min_{\beta} S_n(\beta), \qquad (1)$$

which is called Ichimura's estimator (the WSLS estimator).

- $w(X_i)$ is a nonnegative weight function.
- $A_\delta = \{x : p(x^T\beta) \geq \delta,$ for $^\forall \beta \in \mathcal{B}\}$.
- $A_n = \{x : ||x - x^\star|| \leq 2h,$ for $^\exists x^\star \in A_\delta\}$, which shrinks to A_δ as $n \to \infty$ and $h \to 0$.
- $\mathbf{1}(X_i \in A_n)$ is a trimming function to trim out small values of $\widehat{p}(X_i^T\beta) = \dfrac{1}{nh}\sum_{j\neq i} K\left(\frac{X_j^T\beta - X_i^T\beta}{h}\right)$, so that we do not suffer from the random denominator problem.

Asymptotic Distribution of Ichimura's (1993) Estimator

- Let $\widehat{\beta}$ denote the semiparametric estimator of β_0 obtained from minimizing $S_n(\beta)$.
- To derive the asymptotic distribution of $\widehat{\beta}$, we introduce the following conditions:

Assumption 8.1

The set A_δ is compact, and the weight function $w(\cdot)$ is bounded and posotive on A_δ. Define the set

$$D_z = \{z : z = x^T\beta, \beta \in \mathcal{B}, x \in A_\delta\}.$$

Letting $p(\cdot)$ denote the PDF of $z \in D_z$, $p(\cdot)$ is bounded below by a positive constant for ${}^\forall z \in D_z$

Assumption 8.2

$g(\cdot)$ and $p(\cdot)$ are 3 times differentiable w.r.t. $z = x^\beta$. The third derivatives are Lipschitz continuous uniformly over $\mathcal{B}$ for $^\forall z \in D_z$.

Assumption 8.3

The kernel function is a bounded second order kernel, which has bounded support; is twice differentiable; and its second derivative is Lipschitz continuous.

Assumption 8.4

$\mathbb{E}(|Y^m|) < \infty$ for $^\exists m \geq 3$. var$(Y \mid x)$ is bounded and bounded away from zero for $^\forall x \in A_\delta$. $\frac{q \ln(h)}{nh^{3+\frac{3}{m-1}}} \to 0$ and $nh^8 \to 0$ as $n \to \infty$.

Theorem 8.1 (Asymptotic Distribution of Ichimura's (1993) Estimator)

Under assumptions 8.1 through 8.4,

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} \text{Normal}(0, \Omega_I),$$

with

$$\Omega_I = V^{-1}\Sigma V^{-1},$$

$$V = \mathbb{E}\{w(X_i)(g_i^{(1)})^2 \times (X_i - E_A(X_i \mid X_i^T\beta_0))(X_i - E_A(X_i \mid X_i^T\beta_0))^T\},$$

$$\Sigma = \mathbb{E}\{w(X_i)\sigma^2(X_i)(g_i^{(1)})^2 \times (X_i - E_A(X_i \mid X_i^T\beta_0))(X_i - E_A(X_i \mid X_i^T\beta_0))^T\},$$

$$(g_i^{(1)}) = \frac{\partial g(v)}{\partial v} \Big|_{v=X_i^T\beta_0},$$

$$\mathbb{E}_A(X_i \mid v) = \mathbb{E}(X_i \mid x_A^T\beta_0 = v),$$

x_A has the distribution of X_i conditional on $X_i \in A_\delta$.

Proof (Theorem 8.1)

- See Ichimura (1993); and Hardle, Hall and Ichimura (1993) for the proof of Theorem 8.1.
- Horowitz (2009) provides an excellent heuristic outline for the proof, using only familiar Taylor series methods, the standard LLN, and the Lindeberg-Levy CLT.

Optimal Weight under Homoscedasticity

- We introduce the following homoscedasticity assumption:

$$\mathbb{E}(u_i^2 \mid X_i) = \sigma^2.$$

- Under this assumption, the optimal choice of $w(\cdot)$ is $w(X_i) = 1$.
- In this case,

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \widehat{G}_{-i}(X_i^T \beta)^2) \mathbf{1}(X_i \in A_n)$$

is semiparametrically efficient in the sense that Ω_I is the semiparametric variance lower bound (conditional on $X \in A_\delta$).

Optimal Weight under Heteroscedasticity

- In general, $\mathbb{E}(u_i^2 \mid X_i) = \sigma^2(X_i)$.
- An infeasible case: If one assues that $\mathbb{E}(u_i^2 \mid X_i) = \sigma^2(X_i^T\beta_0)$, that is, the conditional variance depends only on the single index $X_i^T\beta_0$, the choice of $w(X_i) = \frac{1}{\sigma^2(X_i^T\beta_0)}$ can lead to a semiparametrically efficient estimation.
- We could employ a two-step procedure as follows.
- Suppose that the conditional variance is a function of $X_i^T\beta_0$ (Let $\sigma^2(X_i^T\beta_0)$ denote it).

- 1st step: Use $w(X_i) = 1$ to obtain a $\sqrt{n}$-consistent estimator of β_0.
- Let $\tilde{\beta}_0$ denote the estimator of β_0, and $\tilde{u}_i = Y_i - \widehat{g}(X_i^T \tilde{\beta}_0)$ denote the residual obtained from $\tilde{\beta}_0$.
- We can obtain a consistent nonparametric estimator of the conditional variance: $\widehat{\sigma}^2(X_i^T \tilde{\beta}_0)$.
- 2nd step: Estimate β_0 again using $w(X_i) = \frac{1}{\widehat{\sigma}^2(X_i^T \tilde{\beta}_0)}$:

$$\widehat{\beta}_0 = \arg\min_\beta \sum_{i=1}^{n} \left[Y_i - \widehat{G}_{-i}(X_i^T \beta) \right]^2 \frac{1}{\widehat{\sigma}^2(X_i^T \tilde{\beta}_0)} \mathbf{1}(X_i \in A_n).$$

- The estimator $\widehat{\beta}_0$ is semiparametrically efficient because $\widehat{\sigma}^2(v) - \sigma^2(v)$ converges to zero at a particular rate uniformly over $v \in D_v$ (D_v is the support of $X_i^T \beta_0$). [1]

[1] $\widehat{\sigma}^2(X_i^T \beta)$ can be used instead.

Bandwidth Selection for Ichimura's (1993) Estimator

- Recall that we assume in Assumption 8.4 that $\frac{q\ln(h)}{nh^{3+\frac{3}{m-1}}} \to 0$ and $nh^8 \to 0$ as $n \to \infty$, where $m \geq 3$ is a positive integer whose specific value depends on the existence of the number of finite moments of Y along with the smoothness of the unknown function $g(\cdot)$. [2]
- The range of permissive smoothing parameters allows for optimal smoothing, i.e., $h = O(n^{-\frac{1}{5}})$, which satisfies Assumption 8.4.

[2] Assumption 8.4 is a sufficient condition ensuring that nonparametric estimation of g does not affect the convergence rate of the parametric part.

- Our aim is to choose $\widehat{\beta}$ close to β_0, and h close to the value h_0, which minimize the average of

$$\mathbb{E}\{\widehat{g}(X_i^T\beta_0 \mid X_i^T\beta_0) - g(X_i^T\beta_0)\}^2.$$

- Hardle, Hall and Ichimura (1993) suggest picking β and the bandwidth h jointly by minimization of $S_n(\beta)$.

- In the proof of Theorem 8.1, we can establish the following decomposition of the least squares criterion:

$$
\begin{aligned}
S_n(\beta, h) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i \widehat{G}_{-i}(X_i^T\beta))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - G(X_i^T\beta))^2 + \frac{1}{n}\sum_{i=1}^{n}(G_{-i}(X_i^T\beta_0) - g(X_i\beta_0))^2 + op(1) \\
&\equiv S(\beta) + T(h) + op(1).
\end{aligned}
$$

- Minimize $S_n(\beta, h)$ simultaneously over both $(\beta, h) \in \mathcal{B}_n \times \mathcal{H}_n$ is equivalent to
    - first minimizing $S(\beta)$ over $\beta \in \mathcal{B}_n$; and
    - second minimizing $T(h)$ over $h \in \mathcal{H}_n$.
- Let $(\widehat{\beta}, \widehat{h})$ be the minimizers of $S_n(\beta, h)$.
- Suppose that we use the second order kernel. Hardle, Hall and Ichimura (1993) show that the MSE optimal bandwidth satisfies

$$
\widehat{h} = O(n^{-\frac{1}{5}}), \quad \frac{\widehat{h}}{h_0} \xrightarrow{p} 1.
$$

Regularity Conditions in Ichimura (1993)

For identification:

- A second order kernel
- h satisfies Assumption 8.4.
- $\mathbb{E}[|Y^m|] < \infty$ for $^{\exists}m \geq 3$.

Regularity Conditions in HHI (1993)

For asymptotic properties:

- A higher order kernel (at least 4)
- $h = O(n^{-\frac{1}{5}})$
- Y has moments of any order.

Direct Semiparametric Estimators

Direct Semiparametric Estimators for β

- Here we review:
  - Hardle and Stoker's (1989) Average Derivative Estimator,
  - Powell, Stock and Stoker's (1989) Density-Weighted Average Derivative Estimator,
  - Li, Lu and Ullah's (2003) Estimator, and
  - Hristache, Juditsky and Spokoiny's (2001) Improved Average Derivative Estimator.
- The advantage of the direct estimation method is that we can estimate β_0 and $g(x^T\beta_0)$ directly without running the nonlinear least squares, which leads to the computational simplicity.
- We still suffer from a finite-sample problem.

Hardle and Stoker's (1989) Average Derivative Estimator

- Suppose that x is a $q \times 1$ vector of continuous variables.
- Then we obtain the average derivative of $\mathbb{E}(Y \mid X = x)$:

$$\mathbb{E}\left[\frac{\partial \mathbb{E}(Y \mid X = x)}{\partial x}\right] = \mathbb{E}\left[g^{(1)}(x^T \beta_0)\right] \beta_0$$

- Recall that the scale of β_0 is not identified, which means that the constant $\mathbb{E}\left[g^{(1)}(x^T \beta_0)\right]$ does not matter. That is, a normalized estimation of $\mathbb{E}\left[\frac{\partial \mathbb{E}(Y|X=x)}{\partial x}\right]$ is an estimation of normalized β_0.

- Let $\widehat{\mathbb{E}}(Y_i \mid X_i)$ denote the NW estimator of $\mathbb{E}(Y_i \mid X_i)$:

$$\widehat{\mathbb{E}}(Y_i \mid X_i) = \frac{\sum_{j=1}^{n} Y_j K\left(\frac{X_i - X_j}{a}\right)}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{a}\right)}.$$

- Assuming that the kernel function is differentiable, we can estimate β_0, estimating $\mathbb{E}\left[\frac{\partial \mathbb{E}(Y|X=x)}{\partial x}\right]$ with its sample analogue:

$$\tilde{\beta}_{ave} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \widehat{\mathbb{E}}(Y_i \mid X_i)}{\partial X_i}.$$

- The scale normalization can also be implemented by $\frac{\tilde{\beta}_{ave}}{|\tilde{\beta}_{ave}|}$.

- An issue raised with this estimator is the random denominator problem, which leads to a difficulty in the derivation of the asymptotic properties.
- Rilstone (1991) establishes the $\sqrt{n}$-normality using a trimming function.

Bandwidth Selection for Hardle and Stoker's (1989) Average Derivative Estimator

- The estimation of β_0 involves the q-dimensional multivariate nonparametric estimation of the first order derivatives.
- Smoothing Parameters for $\widehat{f}^{(1)}_{-i}(X_i)$: Hardle and Tsybakov (1993) suggest to choose the smoothing parameters $h_1, \cdots, h_q$ to minimize MSE of $\widehat{\delta}$.
- They show that the asymptotically optimal bandwidth is given by $h_s = c_s n^{-\frac{2}{2q+v+2}}$, for all $s = 1, \ldots, q$, where c_s is the constant, and v is the order of kernel.
- Powell and Stoker (1996) provide a method for estimating c_s.
- Horowitz (2009) suggests to select h_s based on bootstrap resampling.

- Smoothing Parameters for $\widehat{g}({X_i}^T\beta_n)$: Once we select the optimal h_s's, we can obtain an estimator of β. Let β_n denote a generic estimator.
- We estimate $\mathbb{E}[y|x] = g(x^T\beta_0)$ by $\widehat{g}(x^T\beta_n, h) = \widehat{g}(x^T\beta_n)$. The smoothing parameter associated with the scalar index $x^T\beta_n$ can be selected by least squares cross-validation:

$$\widehat{h} = \arg\min_h \sum_{i=1}^{n} [Y_i - \widehat{g}_{-i}(X_i^T\beta_n, h)]^2.$$

- Under some regularity conditions, the selection of h is of order $Op(n^{-\frac{1}{5}})$.

- As we obtain the average derivative above, we also obtain the weighted average derivative of $\mathbb{E}(Y \mid X = x)$:

$$\mathbb{E}\left[w(x)\frac{\partial \mathbb{E}(Y \mid X = x)}{\partial x}\right] = \mathbb{E}\left[w(x)g^{(1)}(x^T\beta_0)\right]\beta_0.$$

- Let $w(x)$ be the density function $f(x)$, and δ denote the density-weighted average derivative of $\mathbb{E}(Y \mid X = x)$.
- Then we obtain

$$\begin{aligned}\delta &= \mathbb{E}\left[f(X)\frac{\partial \mathbb{E}(Y \mid X = x)}{\partial x}\right] \\ &= \mathbb{E}\left[f(X)g^{(1)}(X^T\beta_0)\right] \\ &= \int g^{(1)}(x^T\beta_0)f^2(x)dx \\ &= g(x^T\beta_0)f^2(x) - 2\int g(x^T\beta_0)f^{(1)}(x)f(x)dx.\end{aligned}$$

- Assume that $f(x) = 0$ at the boundary of the support of X. Then we observe that $g(x^T\beta_0)f^2(x) = 0$, that is,

$$\begin{aligned}\delta &= -2\int g(x^T\beta_0)f^{(1)}(x)f(x)dx \\ &= -2\mathbb{E}[g(X^T\beta_0)f^{(1)}(X)] \\ &= -2\mathbb{E}[Yf^{(1)}(X)].\end{aligned}$$

- We can estimate δ by its sample analogue:

$$\widehat{\delta} = -\frac{2}{n}\sum_{i=1}^{n} Y_i \widehat{f}_{-i}^{(1)}(X_i), \tag{2}$$

where $\widehat{f}_{-i}(X_i)$ is the leave-one-out NW estimator of $f(X)$:

$$\widehat{f}_{-i}(X_i) = \frac{1}{n-1}\sum_{j\neq i}\left(\frac{1}{h}\right)^q K\left(\frac{X_i - X_j}{h}\right).$$

- There is no denominator messing with uniform convergence. There is only a density estimator, no conditional mean needed.
- The textbook uses the NW estimator $\widehat{f}^{(1)}(X_i)$ in (2), while Powell, Stock and Stoker (1989) define their estimator using the leave-one-out NW estimator $\widehat{f}_{-i}^{(1)}(X_i)$.
- Here we proceed with Powell, Stock and Stoker (1989).

- A useful representation of $\widehat{\delta}$ is given by

$$\widehat{\delta} = \frac{-2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \left(\frac{1}{h}\right)^{q+1} Y_i K^{(1)} \left(\frac{X_i - X_j}{h}\right).$$

- Under some assumptions, if $h \to 0$ and $nh^{q+2} \to \infty$ hold, then the density-weighted average derivative estimator $\widehat{\delta}$ satisfies that

$$\sqrt{n}(\widehat{\delta} - \mathbb{E}[\widehat{\delta}]) \xrightarrow{d} \text{Normal}(0, \Sigma_\delta),$$

where $\Sigma_\delta = 4\mathbb{E}[\sigma^2(X) f^{(1)}(X) f^{(1)}(X)^T] + 4\text{Var}(f(X) g^{(1)}(X))$.

- Recall that $K(\cdot)$ is differentiable and symmetric, that is, $K^{(1)}(u) = -K^{(1)}(-u)$. Then, we obtain the standard U-statistics form of $\widehat{\delta}$:

$$\widehat{\delta} = -\binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{1}{h}\right)^{q+1} K^{(1)}\left(\frac{X_i - X_j}{h}\right)(Y_i - Y_j).$$

- Letting Z_i denote $(Y_i, X_i^T)^T$ and $p_n(Z_i, Z_j)$ denote $-\frac{1}{h^{q+1}} K^{(1)}\left(\frac{X_i - X_j}{h}\right)(Y_i - Y_j)$, $\widehat{\delta}$ can be rewritten as

$$\widehat{\delta} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_n(Z_i, Z_j).$$

- This representation of $\widehat{\delta}$ permits a direct analysis of its asymptotic properties, based on the asymptotic theory of U-statistics. Further discussions can be seen in Serfling (1980); van der Vaart (1998, Chapter 12).

- The asymptotic bias is a bit complicated.
- Let q be the dimension of X, and set

$$p = \begin{cases} \dfrac{q+4}{2} & \text{if } q \text{ is even,} \\[2ex] \dfrac{q+3}{2} & \text{if } q \text{ is odd.} \end{cases}$$

- The kernel function $K(\cdot)$ for the estimation of $f(\cdot)$ is required to be of order at least p.
- The asymptotic bias is $\sqrt{n}(\mathbb{E}(\widehat{\delta}) - \delta) = O(n^{\frac{1}{2}} h^p)$, which is $o(1)$ if $nh^{2p} \to 0$.

- $nh^{2p} \to 0$ is violated if h is selected to be optimal for the estimation of $f(\cdot)$ or $f^{(1)}(\cdot)$. That is, this requirement needs the bandwidth h to undersmooth to reduce the bias. Further discussions on the bandwidth selection follow in Section 8.4.
- Cattaneo, Crump and Jansson (2010, 2011) introduce another asymptotic theory to relax strong assumptions .
- Nishiyama and Robinson (2005): Density-weighted average derivative estimators can be refined by bootstrapping methods.

Li, Lu and Ullah's (2003) Estimator

- We consider the estimation of the average derivative $\mathbb{E}[g^{(1)}(X^T\beta_0)]$ again.
- We can also use the local polynomial method for the estimation of $g^{(1)}(X^T\beta_0)$.
- Let $\widehat{g}^{(1)}(X_i^T\beta_0)$ denote the kernel estimator of $g^{(1)}(X_i\beta_0)$, which is obtained from an m-th order local polynomial regression.
- Li, Lu and Ullah (2003) suggest to use $\tilde{\beta}_{ave} = \frac{1}{n}\widehat{g}^{(1)}(X_i^T\beta_0)$ to estimate $\beta = \mathbb{E}[g^{(1)}(X^T\beta_0)]$.

- Their approach does not require the condition $f(x) = 0$ at the boundary of the support of X. However, they require to assume that
  - the support of X is a compact set, and that
  - the density $f(x)$ is bounded below by a positive constant at the support of X,

  which avoids the use of a trimming function.

- Under the assumptions so far and some additional conditions, if we use a second order kernel, where $n\sum_{s=1}^{q} a_s^{2m} \to 0$ and $\dfrac{na_1 \cdots a_q \sum_{s=1}^{q}}{\ln(n)} \to \infty$ with m denoting the order of local polynomial estimation, then,

$$\sqrt{n}(\tilde{\beta}_{ave} - \beta) \xrightarrow{d} \text{Normal}\left(0, \Phi + \text{var}(g^{(1)}(X^T\beta_0))\right),$$

where $\Phi = \mathbb{E}\left[\dfrac{\sigma^2(X)f^{(1)}(X)f^{(1)}(X)^T}{f^{(2)}(X)}\right]$.

- The proof of the asymptotic normality can be derived from U-statistics theory.
- Newey (1994) shows that the asymptotic variance does not depend on the specific nonparametric estimation method.

Hristache, Juditsky and Spokoiny's (2001) Improved Average Derivative Estimator

- Powell, Stock and Stoker's (1989) density-weighted average derivative estimator requires the density of X to be increasingly smooth as the dimension of X increases.
- This is necessary to make $\sqrt{n}(\widehat{\delta} - \delta)$ asymptotically normal with a mean of 0.
- Practical Consequence: The finite-sample performance of the density-weighted average derivative estimator is likely to be deteriorated as the dimension of X increases, especially if the density of X is not very smooth.
- Specifically, the estimator's bias and MSE are likely to increase as the the dimension of X increases.

- Hristache, Juditsky and Spokoiny (2001) introduce an iterated average derivative estimator that overcomes this problem.
- Their estimator is based on the observation that $g(x^T\beta_0)$ does not vary when x varies in a direction that is orthogonal to β_0.
- Therefore, only the directional derivative of $\mathbb{E}(Y \mid X = x)$ along the direction of β is needed for estimation.
- Suppose that this direction were known. Then estimating the directional derivative would be a one-dimensional nonparametric estimation problem, and there would be no curse of dimensionality.

- In practice, the direction of β is unknown.
- Hristache, Juditsky and Spokoiny (2001) show that this can be estimated with sufficient accuracy through an iterative procedure.
- Their idea is to use prior information about the vector β for improving the quality of the gradient estimate by extending a weighting kernel in the direction of small directional derivatives, and they demonstrate that the whole procedure requires at most $2\log(n)$ iterations.
- Under relatively mild assumptions, their estimator is $\sqrt{n}$-consistent.
- See Horowitz (2009, Section 2.6) for further discussions.

Estimation of $g(\cdot)$

- Let β_n denote a $\sqrt{n}$-consistent estimator of β, or δ.
- Once we obtain β_n, we can estimate $g(x^T\beta_0)$ by

$$\widehat{g}(x^T\beta_n) = \frac{\sum_{j=1}^{n} Y_j K\left(\frac{(X_j - x)^T\beta_n}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{(X_j - x)^T\beta_n}{h}\right)}.$$

- Recall that β_n is a $\sqrt{n}$-consistent estimator of β, that is, $\beta_n - \beta_0 = Op(n^{-\frac{1}{2}})$,
- This converges to zero faster than standard nonparametric estimators.
- Then, the asymptotic distribution of $\widehat{g}(x^T\beta_n)$ is the same as that of $\widehat{g}(x^T\beta_0)$.

Corollary 8.1

$$\sqrt{nh}[\widehat{g}(x^T\beta_n) - g(x^T\beta_0) - h^2 B(x_0^\beta)] \xrightarrow{d} \text{Normal}\left(0, \frac{\kappa\sigma^2(x^T\beta_0)}{f(x^T\beta_0)}\right).$$

Generalized Cases?

- The direct average derivative estimation method discussed previously is applicable only when x is a $q \times 1$ vector of continuous variables because the derivative w.r.t. discrete variables is not defined.
- Horowitz and Hardle (1996) discuss how direct (noniterative) estimation can be generalized to cases for which some components of x are discrete. Horowitz (2009) provides an excellent overview of this method.

Finite-Sample Problem

- Nonparametric estimation in the 1st stage may suffer from the curse of dimensionality.
- In small-sample settings, the iterative method of Ichimura (1993) may be more appealing as it avoids having to conduct high-dimensional nonparametric estimation.

Carroll, Fan, Gijbels and Wand (1997)

- They consider the problem of estimating a general partially linear single index model which contains both a partially linear model and a single index model as special cases.