1 Semiparametric Single Index Models

The Semiparametric Single Index Model is

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

where Y is the dependent variable, $X \in R^q$ is the row vector of explanatory variables, β_0 is $q \times 1$ vector of unknown parameters, and u is the error satisfying $E[u|X] = 0$. The term $X\beta_0$ is called a "single index" because it is a scalar (a single index) even though x is a vector. Here the linear index is specified while the function $g(\cdot)$ is left unspecified.

Background: Parametric Estimation

1. **Censored Regression Model:** $Y_i = \max\{0, Y_i^*\} = Y_i^* 1\{Y_i^* > 0\}$, where $Y_i^* = X_i\beta_0 + \epsilon_i$. $E[\epsilon_i|X_i] = 0$. The density function of ϵ is $f(\cdot)$ with distribution function $F(\cdot)$. Then the censored conditional expectation is

$$\begin{aligned}
E[Y_i|X_i] &= 0 \times P(Y_i = 0|X_i) + E[Y_i|Y_i > 0, X_i]P(Y_i > 0|X_i) \\
&= E[Y_i^*|Y_i^* > 0, X_i]P(Y_i^* > 0|X_i) \\
&= E[X_i\beta_0 + \epsilon_i|\epsilon_i > -X_i\beta_0, X_i]P(\epsilon_i > -X_i\beta_0|X_i) \\
&= X_i\beta_0 P(\epsilon_i > -X_i\beta_0|X_i) + \int_{-X_i\beta_0}^{\infty} t f(t) dt \\
&\equiv g(X_i\beta_0).
\end{aligned}$$

The conditional expectation is a function of x only through the index $x\beta_0$. The function $g(\cdot)$ maps the index into the response: $Y = g(x\beta_0) + u$, where $E[u|x] = 0$.

The parametric density function of ϵ is $f(\cdot;\theta_0)$ with distribution function $F(\cdot;\theta_0)$. Since the density function of Y is

$$\begin{aligned}
f_Y(y) &= \begin{cases} f(y - x\beta_0;\theta_0) & \text{if } y > 0 \\ P(Y^* \leq 0) & \text{if } y = 0 \end{cases} \\
&= \begin{cases} f(y - x\beta_0;\theta_0) & \text{if } y > 0 \\ F(-x\beta_0;\theta_0) & \text{if } y = 0 \end{cases}
\end{aligned}$$

The log likelihood function of the censored model is given by

$$\log L = \frac{1}{n}\sum_{i=1}^{n}\left[d_i \log f(Y_i - X_i\beta;\theta) + (1 - d_i)\log(F(-X_i\beta;\theta))\right],$$

where $d_i = 1\{Y_i > 0\}$.

2. **Truncated Regression Model:** $Y_i = Y_i^*$ only when $Y_i^* > 0$, where $Y_i^* = X_i\beta_0 + \epsilon_i$ and $E[\epsilon_i|X_i] = 0$, but there is no information on X_i and Y_i when $Y_i^* \leq 0$. The truncated conditional expectation given X_i is

$$\begin{aligned} E[Y_i|X_i] &= E[Y_i^*|Y_i^* > 0, X_i] \\ &= E[X_i\beta_0 + \epsilon_i|\epsilon_i > -X_i\beta_0, X_i] \\ &= X_i\beta_0 + \frac{\int_{-X_i\beta_0}^{\infty} \epsilon f(\epsilon)d\epsilon}{\int_{-X_i\beta_0}^{\infty} f(\epsilon)d\epsilon} \\ &\equiv g(X_i\beta_0). \end{aligned}$$

The parametric density of ϵ is $f(\cdot;\theta_0)$ with distribution function $F(\cdot;\theta_0)$. Since the density function of Y is

$$f_{Y|Y>0}(y) = \frac{f(y - x\beta_0)}{P(Y > 0|x)} = \frac{f(y - x\beta_0)}{1 - F(-x\beta_0;\theta_0)}$$

The log likelihood function of the trancated model is given by

$$\log L = \frac{1}{n}\sum_{i=1}^{n}\left[\log f(Y_i - X_i\beta;\theta) - \log(1 - F(-X_i\beta;\theta))\right].$$

3. **Binary Choice Parametric Model:** Consider

$$Y_i = \begin{cases} 1, \text{ if } Y_i^* \equiv X_i\beta_0 - \epsilon_i > 0 \\ 0, \text{ if } Y_i^* \equiv X_i\beta_0 - \epsilon_i \leq 0, \end{cases}$$

or $Y_i = 1\{Y_i^* > 0\} = 1\{X_i\beta_0 - \epsilon_i > 0\}$, where $E[\epsilon_i|X_i] = 0$. The parametric linear index $X_i\beta_0$ governs the choices while the distribution of the error term ϵ_i is not specified, i.e. the distribution function $F(\cdot)$ is unknown. We can observe Y (0 or 1), but cannot observe Y^*. The model $Y_i^* \equiv X_i\beta_0 - \epsilon_i$ is a **latent variable model**. We are mainly interested in estimating β_0 based on the data (Y_i, X_i). This is a semiparametric estimation problem. Note that $\epsilon = Y^* - E[Y^*|X] \neq u = Y - E[Y|X]$ since $Y^* \neq Y$.

For example, Y^* denotes the difference between an individual's market wage (observable) and reservation wage (generally unobservable). Y represents a labor force participation decision. $Y = 1$ if and only if $Y^* > 0$. X contains a set of economic factors that could influence the decision, such as age, education, marital status, work history, and number of children.

Let $E[Y_i^*|x] = H(x)$ and $Y^* = H(x) - \epsilon$.If Y^* were observable, H can be nonparametrically estimated. The population distribution of (Y^*, X) would identify H if H is a continuous function of X. The distribution function $F(\cdot)$ of ϵ is also identified because $\epsilon = H(x) - Y_i^*$ is identified.

However, Y^* is unobservable since the market wage is observable only for employed individuals and the reservation wage is never observable. But $G(x) \equiv E[Y|x]$ can be nonparametrically estimated. From $Y^* = H(x) - \epsilon$,

$$G(x) \equiv E[Y|x] = P(Y^* > 0|x) = P(\epsilon < H(x)|x) = F(H(x)). \tag{1}$$

Therefore, H and F are identified and nonparametrically estimatable only if (1) has a unique solution for H and F in terms of G. **Whether H and F are identified and estimated nonparametrically?** No unless H is restricted! For example, suppose that x is a scalar and $G(x) = e^x/(1+e^x)$. One solution to (1) is

$$\begin{cases} H(x) = x \\ F(u) = e^u/(1+e^u), \quad -\infty \leq u \leq \infty. \end{cases}$$

Another solution is

$$\begin{cases} H(x) = e^x/(1+e^x) \\ F(u) = u, \;\; 0 \leq u \leq 1. \end{cases}$$

Assume that H has the single-index structure $H(x) = x\beta_0$ and that $F(\cdot)$ is known.

Compare

$$E[Y^*|x] = E[X\beta_0 - \epsilon|X = x] = x\beta_0$$

and

$$\begin{aligned} E[Y|x] &= 1 \times P(Y=1|x) + 0 \times P(Y=0|x) \\ &= P(Y=1|x) \equiv p(x) \\ &= P(X\beta_0 - \epsilon > 0|X = x) \\ &= P(\epsilon < x\beta_0) \\ &= F(x\beta_0) \qquad (2) \\ &= g(x\beta_0) \end{aligned}$$

The probability $P(Y=1|x)$ is a function of x only through the index $x\beta_0$. The function $g(\cdot)$ maps the index into the response probability. $Y = g(x\beta_0) + u$, where $E[u|x] = 0$. The parameter β_0 reflects the impact of changes in X on the probability of participating in the labor market. Since

$$\partial p(x)/\partial x_k = g'(x\beta_0)\beta_{0k} = f(x\beta_0)\beta_{0k},$$

the partial effect of x_k on $p(x)$ depends on x through $f(x\beta_0)$.

Note: 1) If $F(\cdot)$ is strictly increasing, $f(z) > 0$ for all $z > 0$ and the sign of the effect of x_k is given by the sign of β_{0k}, i.e. the direction of the effects of x_k on $E[Y^*|x]$ and $E[Y|x]$ are identical.

2) The relative effects do not depend on x since

$$\frac{\partial p(x)/\partial x_j}{\partial p(x)/\partial x_k} = \frac{\beta_{0j}}{\beta_{0k}}$$

is a constant.

3) If ϵ has a symmetric distribution about zero, with unique mode at zero, the largest effect of x_k on the probability $p(x)$ is $f(0)\beta_{0k}$ when $x\beta_0 = 0$. For example, in the probit case it is $1/\sqrt{2\pi}\beta_{0k} \approx 0.399\beta_{0k}$; in the logit case it is $0.25\beta_{0k}$.This implies that the logit estimates can be expected to be larger by a factor of about $0.4/0.25 = 1.6$ than the probit estimates. Or, multiply the logit estimates by 0.625 to make them comparable to the probit estimates.

Special Specification Examples:

- Probit Model: $\epsilon \sim N(0,1)$ with the density $\phi(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$. The conditional expectation is

$$E[Y|x] = P(Y=1|x) = \Phi(x\beta_0),$$

  where $\Phi(t) = \int_{-\infty}^{t}\phi(v)dv$ is the CDF of a standard normal variable.

- Logistic Model: $\epsilon \sim$ *Logistic* with the density function $f(u) = \frac{\exp(-u)}{(1+\exp(-u))^2}$.The conditional expectation is

$$E[Y|x] = P(Y=1|x) = \frac{\exp(x\beta_0)}{1+\exp(x\beta_0)}.$$

Maximum Likelihood Estimation of Parametric Binary Response Index Models: Suppose that the distribution of ϵ is known. The conditional density of y given x is

$$f(y|x;\beta_0) \equiv F(x\beta_0)^y(1-F(x\beta_0))^{1-y}, \quad y = 0,1.$$

Identification can be guaranteed by the conditional Kullback-Leibler information inequality:

$$\int_y \log\left(\frac{f(y|x;\beta_0)}{f(y|x;\beta)}\right) f(y|x;\beta_0)v(dy) \geq 0$$

for all nonnegative functions $f(y|x;\beta)$ such that $\int_y f(y|x;\beta)v(dy) = 1$ for all possible values of x. The logarithm of the conditional likelihood function of the binary choice model is

$$\log L(\beta) \equiv \sum_{i=1}^{n} \log f(y_i|x_i;\beta) = \sum_{i=1}^{n} \left[y_i \log F(x_i\beta) + (1-y_i)\log(1-F(x_i\beta)) \right].$$

The MLE $\hat{\beta}$ is a solution (if it exists) of $\partial \log L(\beta)/\partial\beta = 0$, where

$$\partial \log L(\beta)/\partial\beta = \sum_{i=1}^{n} \frac{y_i - F(x_i\beta)}{F(x_i\beta)(1-F(x_i\beta))} f(x_i\beta)x_i'.$$

By a Taylor expansion,

$$0 = \frac{\partial \log L(\beta)}{\partial\beta}\Big|_{\hat{\beta}} = \frac{\partial \log L(\beta)}{\partial\beta}\Big|_{\beta_0} + \frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta'}\Big|_{\beta^*} (\hat{\beta} - \beta_0),$$

where β^* lies between $\hat{\beta}$ and β_0, and

$$\begin{aligned} \frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta'} &= -\sum_{i=1}^{n} \left(\frac{y_i - F(x_i\beta)}{F(x_i\beta)(1-F(x_i\beta))} \right)^2 f^2(x_i\beta)x_i'x_i \\ &\quad + \sum_{i=1}^{n} \frac{y_i - F(x_i\beta)}{F(x_i\beta)(1-F(x_i\beta))} f'(x_i\beta)x_i'x_i. \end{aligned}$$

Therefore,

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\left(\frac{\partial^2 \log L(\beta)}{\partial\beta\partial\beta'}\Big|_{\beta^*} \right)^{-1} \frac{\partial \log L(\beta)}{\partial\beta}\Big|_{\beta_0}.$$

For Probit model and Logistic models, we can show (see Amemiya (1984), "Advanced Econometrics", P273-274) that $\partial^2 \log L(\beta)/\partial\beta\partial\beta' < 0$ for $\beta \in B$ (an open bounded subset of $R^q, \beta_0 \in B$) which justifies the conditional MLE, and when $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_i'x_i$ is a finite nonsingular matrix, the MLE estimator is root-n consistent and asymptotically normal:

$$\sqrt{n}(\hat{\beta}_{MLE} - \beta_0) \to N(0, A_0^{-1}),$$

where $A_0 = -E\left[\partial^2 L(\beta_0)/\partial\beta\partial\beta'\right] = E\left[\frac{f^2(x_i\beta_0)}{F(x_i\beta_0)(1-F(x_i\beta_0))} x_i'x_i\right]$. The parameters in Probit and Logit models can be estimated in EViews or Stata.

The **disadvantage** of the parametric method is that different distributional assumption for ϵ lead to different functional forms for the conditional probability of $Y = 1$ (see

(2)). The consistent parametric estimation of $E[Y|x]$ above requires the correct distributional specification of ϵ. Misspecification of the distribution of ϵ will lead to inconsistent parametric estimation.

The **advantages** of a semiparametric single index model (not specify $F(\cdot)$ a prior): It can avoid the problem of error distribution misspecification. It is more general than the binary choice model since Y is not necessarily binary: Y can be continuous or discrete in semiparametric single index model. Also, it is an alternative approach designed to mitigate effects arising from the curse of dimensionality.

Why Single-Index Models $Y = g(X\beta_0) + u$ Are Useful?

1) A Single-Index Model does not assume that $g(\cdot)$ is known, and hence it is more flexible and less restrictive than are parametric models for conditional mean functions, such as linear models and binary probit models. Use of a semiparametric single-index model reduces the risk of obtaining misleading results.

2) Although nonparametric estimation of a conditional mean function maximizes flexibility and minimizes (but does not eliminate) the risk of specification error, the price of this flexibility can be high for several reasons: (i) Nonparametric estimation precision decreases rapidly as the dimension of X increases. To obtain acceptable estimation precision if X is multidimensional (as it often is in economic application), impracticably large samples may be needed. However, a single index model avoids the curse of dimensionality because the index $X\beta$ aggregates the dimensions of X. At the same time β can be estimated with the same rate of convergence, $n^{-1/2}$, that is achieved in a parametric model. (ii) Nonparametric estimation results (usually without simple analytic forms) can be difficult to display and interpret when X is multidimensional. (iii) Nonparametric estimation does not permit extrapolation: it does not provide predictions of $E[Y|x]$ at points x that are not in the support of X. This is a serious drawback in policy analysis and forecasting. A single-index model, by contrast, permits extrapolation within limits: it yields predictions of $E[Y|x]$ at values of x that are not in the support of X but are in the support of $X\beta$.

Identification Condition:

β_0 and $g(\cdot)$ must be uniquely determined by the population distribution of (Y, X).

- $g(\cdot)$ cannot be a constant function; otherwise, β_0 is not identified.
- Perfect multicollinearity is not allowed in different components of x.

- β_0 cannot contain a location parameter. It only is identifiable up to a scale. Compare

$$\begin{aligned} E[Y|x] &= g(x\beta_0), \\ E[Y|x] &= g^*(\gamma + x\beta_0\delta). \end{aligned}$$

They are observationally equivalent. They could not be distinguished empirically even if the population distribution of (Y, X) were known. β_0 and $g(\cdot)$ are not identified unless restrictions are imposed that uniquely specify γ and δ. Therefore, β_0 should be location normalized and scale normalized: x does not contain a constant and β has unit length $|\beta| = 1$ or the first component of x has a unit coefficient (and is continuous).

- x should contain at least one continuous random variable. Otherwise, there exist an infinite number of different choices of $g(\cdot)$ and β that satisfy the finite set of restrictions imposed by $E[Y|x] = g(x\beta)$. Give an example to illustrate this.... Suppose that (X_1, X_2) is two-dimensional and discrete with support: $(0,0), (0,1), (1,0), (1,1)$. The coefficient of x_1 is normalized to be 1. Then

$$E[Y|x] = g(x_1 + \beta_2 x_2).$$

The left hand is identified while the right hand is not.

The identification conditions of a single index model are summarized in the following:

(i) x should not contain a constant (intercept), and x 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 $x\beta_0$.

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

Estimation:

If $g(\cdot)$ were known, use the nonlinear LS method to estimate β_0:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - g(X_i\beta))^2 w(X_i), \tag{3}$$

where $w(\cdot)$ is an appropriate weight function for possible heteroscedasticity. Suppose that $g(\cdot)$ is unknown. The kernel method can not be used to estimate $g(X_i\beta)$ directly because $g(\cdot)$ and β_0 are both unknown. However, for a given value of β, since $Y_i = g(X_i\beta_0) + u_i$ and $E[u_i|X_i] = 0$, we can estimate

$$G(X_i\beta) \equiv E[Y_i|X_i\beta] = E[g(X_i\beta_0)|X_i\beta]$$

by the kernel method.

(Recall: $E(y|h(x)) = E[E(y|h(x))|x]$ and $E(y|h(x)) = E[E(y|x)|h(x)]$, where $h(x)$ is a random vector that is a function of x).

Note that when $\beta = \beta_0, G(X_i\beta) = g(X_i\beta_0)$,while in general, $G(X_i\beta) \neq g(X_i\beta_0)$ if $\beta \neq \beta_0$. A leave-one-out nonparametric kernel estimator of $g(X_i\beta)$ is given by

$$
\begin{aligned}
&\hat{G}_{-i}(X_i\beta) \equiv \hat{E}_{-i}[Y_i|X_i\beta] \\
=\ & \frac{(nh)^{-1}\sum_{j\neq i}^{n} Y_j k\left(\frac{X_j\beta - X_i\beta}{h}\right) w(X_j)1\{X_i \in A_n\}}{(nh)^{-1}\sum_{j\neq i}^{n} k\left(\frac{X_j\beta - X_i\beta}{h}\right) w(X_j)1\{X_i \in A_n\}} \\
\equiv\ & \frac{(nh)^{-1}\sum_{j\neq i}^{n} Y_j k\left(\frac{X_j\beta - X_i\beta}{h}\right) w(X_j)1\{X_i \in A_n\}}{\hat{p}_{-i}(X_i\beta)},
\end{aligned}
$$

where $\hat{p}_{-i}(X_i\beta) = (nh)^{-1}\sum_{j\neq i}^{n} k\left(\frac{X_j\beta - X_i\beta}{h}\right) w(X_j)1\{X_i \in A_n\}$ is the leave-one-out estimator of the PDF $p(\cdot)$ of $X\beta$ at $X_i\beta$, and $1\{X_i \in A_n\}$ is a trimming function to trim out small values of $\hat{p}_{-i}(X_i\beta)$, defined below.

1) Ichimura (1993)'s Estimator: Replace $g(X_i\beta)$ in (3) with $\hat{G}_{-i}(X_i\beta)$ and use a trimming function to trim out small values of $\hat{p}_{-i}(X_i\beta)$. Let

$$
\begin{aligned}
A_\delta &= \{x : p(x\beta) \geq \delta, \forall \beta \in B\}, \\
A_n &= \{x : |x - x^*| \leq 2h_n \text{ for some } x^* \in A_\delta\},
\end{aligned}
$$

where $\delta > 0$ is a constant, B is a compact subset in R^q, $A_\delta \subset A_n$, and as $n \to \infty$, $h_n \to 0$ and A_n shrinks to A_δ. Ichimura (1993)'s estimator is

$$\hat{\beta}_I = \arg\min_{\beta} \sum_{i=1}^{n} \left[Y_i - \hat{G}_{-i}(X_i\beta)\right]^2 w(X_i)1\{X_i \in A_\delta\},$$

where $w(X_i)$ is a positive weight function which is bounded in A_δ. The trimming function ensures that the random denominator in the kernel estimator is positive with high probability so as to simplify the asymptotic analysis. Under some regularity conditions about $g(\cdot)$, $p(\cdot)$ and the kernel $k(\cdot)$, and $E|Y|^m < \infty$ for some $m \geq 3$, $\lim_{n\to\infty} \ln(h)/[nh^{3+3/(m-1)}] = 0$ and $\lim_{n\to\infty} nh^8 = 0$, the estimator $\hat{\beta}_I$ is root-n consistent and asymptotically normal:

$$\sqrt{n}(\hat{\beta}_I - \beta_0) \to N(0, \Omega_I),$$

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

$$
\begin{aligned}
\Sigma &= E\left[w(X_i)\sigma^2(X_i)\left(g'(X_i\beta_0)\right)^2 (X_i - E_A(X_i|X_i\beta_0))'(X_i - E_A(X_i|X_i\beta_0))\right], \\
V &= E\left[w(X_i)\left(g'(X_i\beta_0)\right)^2 (X_i - E_A(X_i|X_i\beta_0))'(X_i - E_A(X_i|X_i\beta_0))\right],
\end{aligned}
$$

where $E_A(X_i|v) = E(X_i|x_A\beta_0 = v)$ with x_A having the distribution of X_i conditional on $X_i \in A_\delta$. A consistent estimator for Ω_I is $\hat{\Omega}_I = \hat{V}^{-1}\hat{\Sigma}\hat{V}^{-1}$, where

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n} w(X_i)\hat{u}_i^2 \left(\hat{g}'(X_i\hat{\beta}_I)\right)^2 (X_i - \hat{E}(X_i|X_i\beta))'(X_i - \hat{E}(X_i|X_i\beta)),\\ \hat{V} &= \frac{1}{n}\sum_{i=1}^{n} w(X_i) \left(\hat{g}'(X_i\hat{\beta}_I)\right)^2 (X_i - \hat{E}(X_i|X_i\beta))'(X_i - \hat{E}(X_i|X_i\beta))\end{aligned}$$

with $\hat{u}_i = Y_i - \hat{g}(X_i\hat{\beta}_I)$ and

$$\hat{E}(X_i|X_i\beta) = \sum_{j=1}^{n} X_j k\left((X_j - X_i)\hat{\beta}_I/h\right) / \sum_{j=1}^{n} k\left((X_j - X_i)\hat{\beta}_I/h\right).$$

It shows that $\hat{\beta}_I$ can be computationally costly in practice. For the **Bandwidth Choice**, since $\lim_{n\to\infty} \ln(h)/[nh^{3+3/(m-1)}] = 0$ for some $m \geq 3$ and $\lim_{n\to\infty} nh^8 = 0$, the range of permissible smoothing parameters allows for optimal smoothing: $h = O(n^{-1/5})$. And alternatively we can choose h and β simultaneously by minimizing

$$\sum_{i=1}^{n} \left[Y_i - \hat{G}_{-i}(X_i\beta, h)\right]^2 w(X_i)1\{X_i \in A_\delta\},$$

where $\hat{G}_{-i}(X_i\beta, h) = \hat{G}_{-i}(X_i\beta)$.

2) Direct Semiparametric Estimator:

From $E[Y|x] = g(x\beta_0)$, we get

$$E\left[\frac{\partial E[Y|x]}{\partial x}\right] = E[g'(x\beta_0)\beta_0] = E[g'(x\beta_0)]\beta_0 \equiv C\beta_0 \tag{4}$$

and

$$E\left[w(x)\frac{\partial E[Y|x]}{\partial x}\right] = E[w(x)g'(x\beta_0)\beta_0] = E[w(x)g'(x\beta_0)]\beta_0 \equiv C_2\beta_0, \tag{5}$$

both of which are proportional to β_0. Then one can estimate β_0 by estimating (4) and (5), respectively, in the following ways:

1. **The average derivative-based estimator:**

$$\hat{\beta}_{ave} \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{E}[Y_i|X_i]}{\partial X_i},$$

where $\hat{E}[Y_i|X_i] = \frac{\sum_{j=1}^{n} Y_j K((X_i - X_j)/a)}{\sum_{j=1}^{n} K((X_i - X_j)/a)}$, $K((X_i - X_j)/a)$ is a product kernel function, a is the vector of smoothing parameter. If one uses $|\beta| = 1$ as the normalization rule, the scale normalization is $\hat{\beta}_{ave}/|\hat{\beta}_{ave}|$; if one chooses to normalize the coefficient of the first regressor to be one, the scale normalization is $\hat{\beta}_{ave}/\hat{\beta}_{ave,1}$.(Use a trimming function to avoid the "small denominator problem").

2. **The weighted average derivative estimator** (see Powell, Stock and Stoker (1989), "Semiparametric Estimation of Index Coefficients", Econometrica Vol 57, No 6, P1403-1430): If $f(x) = 0$ at the boundary of the support of X (e.g. X has unbounded support), choose the weight $w(x) = f(x)$ in (5). Then

$$
\begin{aligned}
E\left[f(X)\frac{\partial E[Y|X]}{\partial X}\right] &= \int \frac{\partial E[Y|X]}{\partial X} f^2(X)dX \\
&= 0 - 2\int E[Y|X]f(X)\frac{\partial f(X)}{\partial X}dX \\
&= -2E\left[g(X\beta_0)\frac{\partial f(X)}{\partial X}\right] \\
&= -2E\left[Y\frac{\partial f(X)}{\partial X}\right] \\
&\equiv \delta,
\end{aligned}
$$

which can be estimated by

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

where $\hat{f}^{(1)}(X_i)$ (a $q \times 1$ vector) is the first-order partial derivative of the kernel estimator

$$\hat{f}(X_i) = \frac{1}{na_1 \cdots a_q}\sum_{j=1}^{n} k\left(\frac{X_{1i} - X_{1j}}{a_1}\right)\cdots k\left(\frac{X_{qi} - X_{qj}}{a_q}\right).$$

The sth component in $\hat{f}^{(1)}(X_i)$ is

$$\frac{\partial \hat{f}(X_i)}{\partial X_{si}} = \frac{1}{n}\sum_{j=1}^{n} a_s^{-2} k^{(1)}\left(\frac{X_{si} - X_{sj}}{a_s}\right)\prod_{t\neq s} a_t^{-1} k\left(\frac{X_{ti} - X_{tj}}{a_t}\right).$$

The PSS's estimator $\hat{\delta}$ does not have a random denominator, and therefore, one does not need to introduce a trimming nuisance parameter. Under some smoothness and moments conditions, PSS prove that

$$\sqrt{n}(\hat{\delta} - \delta) \to N(0, \Omega_{PSS}),$$

where $\Omega_{PSS} = 4E[\sigma^2(X)f^{(1)}(X)f^{(1)}(X)'] + 4var(f(X)g^{(1)}(X\beta_0))$. A normalized vector β can be obtained via $\hat{\delta}/|\hat{\delta}|$. For the **Bandwidth Choice**, choose h to minimize $E[|\hat{\delta} - \delta|^2]$: the optimal bandwidth is of the form $h_s = c_s n^{-2/(2q+v+2)}$, where q is the dimension of x and v is the order of the kernel, and c_s is a constant. When $v = 2$, the optimal bandwidth is $h_s = c_s n^{-1/(q+2)}$, $s = 1, 2, \cdots, q$.

3. Choose $w(x) = 1$ in (5),

$$
\begin{aligned}
E\left[\frac{\partial E[Y|X]}{\partial X}\right] &= \int \frac{\partial E[Y|X]}{\partial X} f(X)dX \\
&= 0 - 2\int E[Y|X]\frac{\partial f(X)}{\partial X}dX \\
&= -2E\left[g(X\beta_0)\frac{\partial f(X)}{\partial X}/f(X)\right] \\
&= -2E\left[Y\frac{\partial f(X)}{\partial X}/f(X)\right] \\
&\equiv \sigma,
\end{aligned}
$$

which can be estimated by

$$
\hat{\sigma} = -\frac{2}{n}\sum_{i=1}^{n} Y_i \frac{\hat{f}^{(1)}(X_i)}{\hat{f}(X_i)} 1\{\hat{f}(X_i) \geq b_n\},
$$

where $b_n > 0$ satisfies $\lim_{n\to\infty} b_n = 0$. A normalized vector β can be obtained via $\hat{\sigma}/|\hat{\sigma}|$.

The disadvantage of the direct average derivative estimation method is that it is applicable only when x is a q-vector of continuous variables since the derivative with respect to discrete variables is not defined. Also, the first-stage nonparametric estimation suffers from the curse of dimensionality which gives rise to a potential finite-sample problem.

The advantage of the direct average derivative estimation method is the computational simplicity in that β_0 and $g(x\beta_0)$ can be directly estimated without using nonlinear iteration procedures. In large sample setting, symptotically, the curse of dimensionality problem disappears because the second stage estimate has a parametric root-n-rate of convergence and the dimension of x does not affect the rate of convergence of the average derivative estimator obtained at the second stage.

However, in small-sample application, the iterative method of Ichimura (1993) is more appealing as it avoids having to conduct high-dimensional nonparametric estimation.

Estimation of Nonparametric Function $g(\cdot)$: Suppose that β_n is one of the estimators, e.g. $\hat{\beta}_I, \hat{\beta}_{ave}, \hat{\delta}$ or $\hat{\sigma}$. With β_n, we can estimate $E[Y|x] = g(x\beta_0)$ by

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

Since $\beta_n - \beta_0 = O_p(n^{-1/2})$ converges to zero faster than standard nonparametric estimators, the asymptotic distribution of $\hat{g}(x\beta_n)$ is the same as the case with β_n being replaced by β_0. Hence, from the asymptotic normality result in Chapter 2 (the case $q = 1$, since $x\beta_0$ is a scalar), we have

$$\sqrt{nh}(\hat{g}(x\beta_n) - g(x\beta_0) - h^2 B(x\beta_0)) \to N(0, \kappa\sigma^2(x\beta_0)/f(x\beta_0)),$$

where $B(x\beta_0) = \frac{\kappa_2}{2}\{2f'(x\beta_0)g'(x\beta_0) + f(x\beta_0)g''(x\beta_0)\}/f(x\beta_0)$, and the other notations are defined in the same way as before.

Testing the Single Index Model:

$$\begin{aligned} H_0 &: E[Y|X = x] = G(x\beta_0) \\ H_1 &: E[Y|X = x] = g(x\beta_0) \end{aligned}$$

where $G(\cdot)$ is a known function while $g(\cdot)$ is an unspecified function. The test statistic is

$$T = \sqrt{h}\sum_{i=1}^n w(X_i\hat{\beta})\left[Y_i - G(X_i\hat{\beta})\right]\left[\hat{G}_{-i}(X_i\hat{\beta}) - G(X_i\hat{\beta})\right] \to N(0, \sigma_T^2),$$

where $w(\cdot)$ is a weight function that downweights extreme observations, often defined in practice as 90% or 95% of the central range of the index values of $X_i\hat{\beta}$ with $\hat{\beta}$ being the estimate under H_0, and $\hat{G}_{-i}(X_i\hat{\beta})$ is the leave-one-out nonparametric estimate.

Other Estimators of β_0 in the Binary Choice Model:

Klein and Spady (1993)'s Estimator:

$$\hat{\beta}_{KS} = \arg\max_{\beta}\sum_{i=1}^n \left[(1 - Y_i)\ln(1 - \hat{g}(X_i\beta)) + Y_i \ln(\hat{g}(X_i\beta))\right],$$

where

$$\hat{g}(X_i\beta) = \frac{\sum_{j\neq i}^n Y_j k\left(\frac{X_j\beta - X_i\beta}{h}\right)}{\sum_{j\neq i}^n k\left(\frac{X_j\beta - X_i\beta}{h}\right)}.$$

Lewbel (2000)'s Estimator: The model is of the form:

$$Y_i = 1\{v_i + X_i\beta_0 + \epsilon_i > 0\},$$

where v_i is a special continuous regressor whose coefficient is normalized to be one and X_i is of dimension q.Let $f(v|x)$ denote the conditional density of v_i given X_i, and let $F_\epsilon(\epsilon|v,x)$ be the conditional CDF of ϵ_i given (v_i, X). Suppose that $F_\epsilon(\epsilon|v,x) = F_\epsilon(\epsilon|x)$ and that $E[X_i\epsilon_i] = 0$. Let $s = -X\beta_0 - \epsilon$. Denote $\tilde{Y}_i = [Y_i - 1\{v_i > 0\}]/f(v_i|X_i)$. L_2 and $-L_1$ are positive and sufficiently large. Supp$(v) = (L_1, L_2)$. Simple calculation shows that

$$\begin{aligned}
E[\tilde{Y}|X] &= E\left[\frac{Y - 1\{v>0\}}{f(v|X)}|X\right] \\
&= E\left[\frac{E[Y - 1\{v>0\}|v,X]}{f(v|X)}|X\right] \\
&= \int_{L_1}^{L_2} \frac{E[Y - 1\{v>0\}|v,X]}{f(v|X)} f(v|X)dv \\
&= \int_{L_1}^{L_2} E[1\{v + X\beta_0 + \epsilon > 0\} - 1\{v>0\}|v,X]dv \\
&= \int_{L_1}^{L_2}\int_{\Omega_{\epsilon|X}} [1\{v + X\beta_0 + \epsilon > 0\} - 1\{v>0\} f_\epsilon(\epsilon|X)d\epsilon dv \\
&= \int_{L_1}^{L_2}\int_{\Omega_{\epsilon|X}} [1\{v - s > 0\} - 1\{v>0\} f_\epsilon(\epsilon|X)d\epsilon dv \\
&= \int_{\Omega_{\epsilon|X}} \left(\int_{L_1}^{L_2} [1\{v > s\} - 1\{v>0\}]dv\right) f_\epsilon(\epsilon|X)d\epsilon \\
&= \int_{\Omega_{\epsilon|X}} \left(-1\{s>0\}\int_0^s 1dv + 1\{s<0\}\int_s^0 1dv]dv\right) f_\epsilon(\epsilon|X)d\epsilon \\
&= \int_{\Omega_{\epsilon|X}} (-s)\, f_\epsilon(\epsilon|X)d\epsilon = E[X\beta_0 + \epsilon|X] \\
&= X\beta_0 + E[\epsilon|X]
\end{aligned}$$

and

$$X'E[\tilde{Y}|X] = X'(X\beta_0 + E[\epsilon|X]) = X'X\beta_0 + X'E[\epsilon|X].$$

Hence

$$\begin{aligned}
E[X'\tilde{Y}] &= E[X'E[\tilde{Y}|X]] = E[X'X]\beta_0 + E[X'E[\epsilon|X]] \\
&= E[X'X]\beta_0 + E[X'\epsilon] \\
&= E[X'X]\beta_0.
\end{aligned}$$

That is,

$$\beta_0 = (E[X'X])^{-1} E[X'\tilde{Y}]. \tag{6}$$

Denote $\hat{Y}_i = [Y_i - 1\{v_i > 0\}]/\hat{f}(v_i|X_i)$, where $\hat{f}(v_i|X_i)$ is the nonparametric kernel conditional density estimator of $f(v_i|X_i)$. The sample analog of (6) gives a feasible estimator of β_0 :

$$\hat{\beta}_L = \left(\sum_{i=1}^{n} X_i' X_i \right)^{-1} \sum_{i=1}^{n} X_i' \hat{Y}_i,$$

which is obtained by regressing $\hat{Y}_i$ on X_i. Lewbel (2000) proved that this estimator is $\sqrt{n}$-consistent and asymptotically normal.

Han (1987)'s Maximum Rank Correlation (MRC) Estimator: For binary choice model $y = 1\{x\beta_0 - \epsilon > 0\}$ with the independence of x and ϵ,

$$E[Y|x] = P(Y = 1|x) = P(X\beta_0 - \epsilon > 0|X = x) = F(x\beta_0),$$

where $F(\cdot)$ is the distribution function of ϵ. The monotonicity of $F(\cdot)$ ensures that

$$E[Y_i - Y_j|X_i, X_j] = E[Y_i|X_i] - E[Y_j|X_j] = F(X_i\beta_0) - F(X_j\beta_0) \geq 0$$

whenever $X_i\beta_0 > X_j\beta_0$. Note that $Y_i - Y_j$ can be valued $1, 0, -1$. Hence,

$$E[Y_i - Y_j|X_i, X_j] = 1 \times P(Y_i - Y_j > 0|X_i, X_j) - 1 \times P(Y_i - Y_j < 0|X_i, X_j) \geq 0,$$

i.e.

$$P(Y_i > Y_j|X_i, X_j) \geq P(Y_i < Y_j|X_i, X_j) \text{ whenever } X_i\beta_0 > X_j\beta_0$$

or

$$\text{when } X_i\beta_0 > X_j\beta_0, \text{ more likely than not } Y_i > Y_j.$$

The intuition is that given an inequality $X_i\beta_0 > X_j\beta_0$ for a pair of samples, it is more likely that $Y_i > Y_j$, i.e. the rankings of the Y_i and the rankings of the $X_i\beta_0$ would be positively correlated. **The idea of the MRC estimator is to maximize with respect to β the rank correlation between the Y_i and the $X_i\beta_0$.** The MRC estimator $\hat{\beta}_H = \arg\max_\beta S_H(\beta)$, where

$$S_H(\beta) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} [1\{Y_i > Y_j\} 1\{X_i\beta > X_j\beta\}]$$

or

$$S_H(\beta) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j > i}^{n} [1\{Y_i > Y_j\} 1\{X_i\beta > X_j\beta\} + 1\{Y_i < Y_j\} 1\{X_i\beta < X_j\beta\}].$$

Han proves the strong consistency of his MRC estimator. Sherman (1993) shows that the MRC estimator is $\sqrt{n}$-consistent and has an asymptotic normal distribution by the U-statistic decomposition theory.

Example 6 (Semiparametric Single Index Model, see ex6) The data generating process is

$$Y_i = 1 + (2X_i + 5Z_i + 1)^2 + u_i, \; i = 1, 2, \cdots, n,$$

where $X_i \sim U[0,1]$ and $Z_i \sim N(0,1)$, $u_i \sim N(0, X_i)$;

In the design, the dependent variable Y is a continuous random variable, $g(v) = 1 + (2v+1)^2$ and the parameter $\beta_0 = 2.5$ (after scale normalization). The sample size is $n = 400$. The sample are independent. In the nonparametric estimation, the bandwidth is chosen as $h = an^{-1/5}$, where $a = 0.4$. Use Ichimura Method.

Example 7 (Binary Choice Model, see ex7) The data generating process is $Y_i = 1\{Y_i^* > 0\}$, and the latent variable

$$Y_i^* = 1 + 2X_i + 5Z_i + \epsilon_i, \; i = 1, 2, \cdots, n,$$

where $X_i \sim U[-1,1]$ and $Z_i \sim N(0,1)$, $\epsilon_i \sim N(0,1)$.

In the design, the dependent variable Y is a binary choice variable, and its conditional expectation given X and Z is

$$\begin{aligned} E[Y|X,Z] &= P(\epsilon > -1 - 2X - 5Z) = \Phi(1 + 2X + 5Z) \\ &= \Phi(1 + 2(X + 2.5Z)). \end{aligned}$$

The nonparametric function $g(v) = \Phi(1 + 2v)$, where $\Phi(\cdot)$ is the distribution function of $N(0,1)$, and the parameter $\beta_0 = 2.5$ (after scale normalization). The sample size is $n = 400$. The sample are independent. In the nonparametric estimation, the bandwidth is chosen as $h = an^{-1/5}$, where $a = 0.4$. Use Ichimura Method.

Exercises

1. Consider the following model

$$Y = \begin{cases} 1, \text{ if } X\beta_0 - \epsilon > 0 \\ 0, \text{ if } X\beta_0 - \epsilon \le 0, \end{cases}$$

where $E[\epsilon|X] = 0$. Show that $P(Y = 1|X) = E[Y|X] = F(X\beta_0)$, where $F(\cdot)$ is the cdf of ϵ. Explain that, if ϵ and X are not independent (for instance, let $\epsilon =$

$\omega(X\beta_0)\varepsilon$, where $\omega(\cdot)$ is an unknown function, $\varepsilon \sim \textit{Logistic}$, and ε is independent of X), $E[Y|X]$ also has a single-index form, that is, $E[Y|X] = g(X\beta_0)$, where $g(\cdot)$ is some link function.

2. Repeat the work in Example 6 by using the weighted average derivative estimation (GAUSS program is required).

3. Repeat the work in Example 7 by using the weighted average derivative estimation and the Lewbel's approach (GAUSS program is required).

4. Consider the following binary choice model

$$Y = 1\{v + X\beta_0 - \epsilon > 0\},$$

where v is a continuous regressor, X is a random row vector of regressors with dimension q, $E[X\epsilon] = 0$, EXX' exists and is nonsingular. Let $g(v|x)$ be the conditional density of v given $X = x$, $f(\epsilon|\cdot)$ the conditional density function of the error term ϵ with $f(\epsilon|v, x) = f(\epsilon|x)$, and the conditional distribution of v given X has support $(-L, L)$, where L is some positive number. Denote $\tilde{Y} = [Y - 1\{v > 0\}]/f(v|X)$. Prove that $\beta_0 = (E[X'X])^{-1} E[X'\tilde{Y}]$ and provide a feasible estimator of β_0.

5. (1) Suppose that Y is a $\{0, 1\}$ binary variable. Show that $P(Y = 1|x) = E(Y|x)$; (2) If Y is a binary variable taking values in $\{1, 2\}$, is it true that $P(Y = 1|x) = E(Y|x)$?