Lec 19: Single index models

Weiping Zhang

December 31, 2020

Single Index Models

Single Index Regression and Ichimura's Estimator

Asymptotic Distribution of Ichimura's Estimator

Klein and Spady's Binary Choice Estimator

Average Derivative Estimator

Testing the SIM

Index Models

- A object of interest such as the conditional density $f(y|x)$ or conditional mean $E(y|x)$ is a single index model when it only depends on the vector x through a single linear combination $x'\beta$, called single index.
- Most parametric models are single index, including Normal regression, Logit, Probit, Tobit, and Poisson regression.
- In a semiparametric single index model, the object of interest depends on x through the function $g(x'\beta)$ where $\beta \in \mathbb{R}^p$ and $g : \mathbb{R} \to \mathbb{R}$ are unknown. g is sometimes called a link function. In single index models, there is only one nonparametric dimension. These methods fall in the class of dimension reduction techniques.

- The semiparametric single index regression model is

$$E(y|x) = g(x'\beta) \tag{1}$$

  where g is an unknown link function.
- The semiparametric single index binary choice model is

$$P(y = 1|x) = E(y|x) = g(x'\beta) \tag{2}$$

  where g is an unkown distribution function. We use g(rather than, say, F) to emphasize the connection with the regression model.
- In both contexts, the function g includes any location and level shift, so the vector X_i cannot include an intercept. The level of β is not identified, so some normalization criterion for β is needed.

Identification

- It is typically easier to impose this on β than on g. One approach is to set $\|\beta\| = 1$. A second approach is to set one component of β to equal one. (This second approach requires that this variable correctly has a non-zero coefficient.)
- The vector X_i must be dimension 2 or larger. If X_i is one-dimensional, then β is simply normalized to one, and the model is the one-dimensional nonparametric regression $E(y|x) = g(x)$ with no semiparametric component.
- Identification of β and g also requires that X_i contains at least one continuously distributed variable, and that this variable has a non-zero coefficient. If not, $X'\beta$ only takes a discrete set of values, and it would be impossible to identify a continuous function g on this discrete support.

- Specifically, let γ be any constant and δ be any nonzero constant. Define the function g^* by the relation $g^*(\gamma + \delta v) = g(v)$ for all v in the support of $X'\beta$. Then $E(y|x) = g(X'\beta)$ and $E(y|x) = g^*(\gamma + x'\beta\delta)$ are observationally equivalent.
- Therefore, β and g are not identified unless restrictions are imposed that uniquely specify γ and δ. The restriction on γ is called a *location normalization*, and the restriction on δ is called a *scale normalization*. Location normalization can be achieved by requiring X to contain no constant (intercept) component. Scale normalization can be achieved by setting the β coefficient of one component of X equal to one.

Example: SIM with only discrete covariates

- Suppose $X = (X_1, X_2)$ is two-dimensional and discrete with support consisting of the corners of the unit square: (0,0), (1,0), (0,1), and (1,1). Set the coefficient X_1 equal to one to achieve scale normalization. Suppose that the values of $E(Y|X = x)$ at the points of support of X are as shown in the following table:

(x_1, x_2)	$E(y\|x)$	$g(x_1 + \beta_2 x_2)$
(0,0)	0	$g(0)$
(1,0)	0.1	$g(1)$
(0,1)	0.3	$g(\beta_2)$
(1,1)	0.4	$g(1 + \beta_2)$

- There are infinitely many such models, so β_2 and g are not identified.

- Another requirement for identification is that g must be differentiable. To understand why, observe that the distinguishing characteristic of a single-index model that makes identification possible is that $E(y|x)$ is constant if x changes in such a way that $x'\beta$ stays constant.
- However, if $X'\beta$ is a continuously distributed random variable, as it is if X has at least one continuous component with a nonzero coefficient, the set of X values on which $X'\beta = c$ has probability zero for any c. Events of probability zero happen too infrequently to permit identification.
- If g is differentiable, then $g(X'\beta)$ is close to $g(c)$ whenever $X'\beta$ is close to c. The set of X values on which $X'\beta$ is within any specified nonzero distance of c has nonzero probability for any c in the interior of the support of $X'\beta$. This permits identification of β through "approximate" constancy of $X'\beta$.

Identification in Single-Index Models

- **Theorem** Suppose that $E(y|X = x)$ satisfies model (1) and X is a p-dimensional random variable. Then β and g are identified if the following conditions hold:
  1. g is differentiable and not constant on the support of $X'\beta$.
  2. The components of X are continuously distributed random variables that have a joint probability density function.
  3. The support of X is not contained in any proper linear subspace of $\mathbb{R}^p$.
  4. $\beta_1 = 1$.
- Ichimura (1993) and Manski (1988) provide proofs of several versions of this theorem. It is also possible to prove a version that permits some components of X to be discrete. Two additional conditions are needed. These are as follows: (1) varying the values of the discrete components must not divide the support of $X'\beta$ into disjoint subsets and (2) g must satisfy a nonperiodicity condition.

Example

Identification of a SIM with Continuous and Discrete Covariates

- Suppose that X has one continuous component, X_1, whose support is [0,1], and one discrete component, X_2, whose support is the two-point set 0,1. Assume that X_1 and X_2 are independent and that g is strictly increasing on [0,1]. Set $\beta_1 = 1$ to achieve scale normalization. Then $X'\beta = X_1 + \beta_2 X_2$.
- Observe that $E[y|X = (x_1, 0)] = g(x_1)$ and $E[y|X = (x_1, 1)] = g(x_1 + \beta_2)$. Observations of X for which $X_2 = 0$ identify g on [0,1]. However, if $\beta_2 > 1$, the support of $X_1 + \beta_2$ is disjoint from [0,1], and β_2 is, in effect, an intercept term in the model for $E[y|X = (x_1, 1)]$. So β_2 is not identified in this model.

- The situation is different if $\beta_2 < 1$, because the supports of X_1 and $X_1 + \beta_2$ then overlap. The interval of overlap is $[\beta_2, 1]$. Because of this overlap, there is a subset of the support of X on which $X_2 = 1$ and $g(X_1 + \beta_2) = g(v)$ for some $v \in [0, 1]$. The subset is $\{X : X_1 \in [\beta_2, 1], X_2 = 1\}$. Since $g(v)$ is identified for $v \in [\beta_2, 1]$ by observations of X_1 for which $X_2 = 0$, β_2 can be identified by solving

$$E[y|X = (x_1, 1)] = g(x_1 + \beta_2) \tag{3}$$

on the set of x_1 values where the ranges of $E[y|X = (x_1, 1)]$ and $g(x_1 + \beta_2)$ overlap.

- To see why g must satisfy a nonperiodicity condition, suppose g were periodic on $[\beta_2, 1]$ instead of strictly increasing. Then (3) would have at least two solutions, so β_2 would not be identified.
- The assumption that g is strictly increasing on [0,1] prevents this kind of periodicity, but many other shapes of g also satisfy the nonperiodicity requirement. See Ichimura (1993) for details.

semiparametric single index regression model

- The semiparametric single index regression (SIR) model is

$$y_i = g(X_i'\beta) + e_i, \qquad E(e_i|X_i) = 0 \tag{4}$$

- This model generalizes the linear regression model (which sets $g(z)$ to be linear), and is a restriction of the nonparametric regression model.
- The gain over full nonparametrics is that there is only one nonparametric dimension, so the curse of dimensionality is avoided.

- Suppose g were known. Then you could estimate β by (nonlinear) least-squares. The LS criterion would be

$$S_n(\beta, g) = \sum_{i=1}^{n} (y_i - g(X_i'\beta))^2$$

  We could think about replacing g with an estimate $\hat{g}$, but since $g(z)$ is the conditional mean of y_i given $X_i'\beta = z$, g depends on β, so a two-step estimator is likely to be inefficient.
- In his PhD thesis, Ichimura proposed a semiparametric estimator, published later in the Journal of Econometrics (1993).

- Ichimura suggested replacing g with the leave-one-out NW estimator

$$\hat{g}_{-i}(X_i'\beta) = \frac{\sum_{j\neq i} \mathcal{K}_h((X_j - X_i)'\beta)y_i}{\sum_{j\neq i} \mathcal{K}((X_j - X_i)'\beta)}$$

  The leave-one-out version is used since we are estimating the regression at the i-th observation.

- Since the NW estimator only converges uniformly over compact sets, Ichimura introduces trimming for the sum-of-squared errors. The criterion is then

$$S_n(\beta) = \sum_{i=1}^{n} (y_i - \hat{g}_{-i}(X_i'\beta))^2 1_i(b)$$

  He is not too specific about how to pick the trimming function, and it is likely that it is not important in applications.

- The estimator of β is then

$$\hat{\beta} = \arg\min_{\beta} S_n(\beta)$$

- The criterion is somewhat similar to cross-validation. Indeed, Hardle, Hall, and Ichimura (Annals of Statistics, 1993) suggest picking β and the bandwidth h jointly by minimization of $S_n(\beta)$.
- In his paper, Ichimura claims that the $\hat{g}_{-i}(X_i'\beta)$ could be replaced by any other uniformly consistent estimator and the consistency of $\hat{\beta}$ would be maintained, but his asymptotic normality result would be lost. In particular, his proof rests on the asymptotic orthogonality of the derivative of $\hat{g}_{-i}(X_i'\beta)$ with e_i, which holds since the former is a leave-one-out estimator, and fails if it is a conventional NW estimator.

Asymptotic Distribution of Ichimura's Estimator

- Let β_0 denote the true value of β. The tricky thing is that $\hat{g}_{-i}(X_i'\beta)$ is not estimating $g(X_i'\beta_0)$, rather it is estimating

$$G(X_i'\beta) = E(y_i|X_i'\beta) = E(g(X_i'\beta_0)|X_i'\beta)$$

  The second equality since $y_i = g(X_i'\beta_0) + e_i$.

- That is

$$G(z) = E(y_i|X_i'\beta = z)$$

  and $G(X_i'\beta)$ is then evaluate at $X_i'\beta$.

- Note that

$$G(X_i'\beta_0) = g(X_i'\beta_0)$$

  but for other values of β,

$$G(X_i'\beta) \neq g(X_i'\beta)$$

- Hardle, Hall, and Ichimura (1993) show that the LS criterion is asymptotically equivalent to replacing $\hat{g}_{-i}(X_i'\beta)$ with $G(X_i'\beta)$, so

$$S_n(\beta) \simeq S_n^*(\beta) = \sum_{i=1}^{n}(y_i - G(X_i'\beta))^2.$$

  This approximation is essentially the same as Andrews' MINPIN argument, and relies on the estimator $\hat{g}_{-i}(X_i'\beta)$ being a leave-one-out estimator, so that it is orthogonal with the error e_i.
- This means that $\hat{\beta}$ is asymptotically equivalent to the minimizer of $S_n^*(\beta)$, a Nonlinear Least Squares (NLLS) estimator problem.

- The asymptotic distribution of the NLLS estimator is identical to least-squares on

$$X_i^* = \frac{\partial}{\partial \beta} G(X_i'\beta).$$

This implies

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightsquigarrow N(0, V)$$

where $V = Q^{-1}\Omega Q^{-1}$, $Q = E(X_i^* X_i^{*\prime})$ and $\Omega = E(X_i^* X_i^{*\prime} e_i^2)$.

- To complete the derivation, we now find this X_i^*. As $\hat{\beta}$ is $n^{-1/2}$ consistent, we can use a Taylor expansion of $g(X_i'\beta_0)$ to find

$$g(X_i'\beta_0) \simeq g(X_i'\beta) + g^{(1)}(X_i'\beta)X_i'(\beta_0 - \beta)$$

where $g^{(1)}(z) = \frac{d}{dz} g(z)$.

- Then

$$\begin{aligned} G(X_i'\beta) &= E(g(X_i'\beta_0)|X_i'\beta) \\ &\simeq E[g(X_i'\beta) + g^{(1)}(X_i'\beta)X_i'(\beta_0 - \beta)|X_i'\beta] \\ &= g(X_i'\beta) - g^{(1)}(X_i'\beta)E(X_i|X_i'\beta)'(\beta - \beta_0) \end{aligned}$$

Since $g(X_i'\beta)$ and $g^{(1)}(X_i'\beta)$ are measurable with respect to $X_i'\beta$. Another Taylor expansion for $g(X_i'\beta)$ yields that this is approximately

$$\begin{aligned} G(X_i'\beta) &\simeq g(X_i'\beta_0) + g^{(1)}(X_i'\beta)(X_i - E(X_i|X_i'\beta))'(\beta - \beta_0) \\ &\simeq g(X_i'\beta_0) + g^{(1)}(X_i'\beta_0)(X_i - E(X_i|X_i'\beta_0))'(\beta - \beta_0) \end{aligned}$$

the final approximation for β in a $n^{-1/2}$ neighborhood of β_0. (The error is of smaller stochastic order)

- We see that

$$X_i^* = \frac{\partial}{\partial \beta} G(X_i'\beta) \simeq g^{(1)}(X_i'\beta_0)(X_i - E(X_i|X_i'\beta_0)).$$

  Ichimura rigorously establishes this result.
- This asymptotic distribution is slightly different than that which would be obtained if the function g were known a priori. In this case, the asymptotic design depends on X_i, not $E(X_i|X_i'\beta_0)$.
- The quantity

$$Q = E(g^{(1)}(X_i'\beta_0)^2(X_i - E(X_i|X_i'\beta_0))(X_i - E(X_i|X_i'\beta_0))')$$

  denotes the cost of the semiparametric estimation.

- Recall when we described identification that we required the dimension of X_i to be 2 or larger. Suppose that X_i is one-dimensional. Then $X_i - E(X_i|X_i'\beta_0) = 0$ so $Q = 0$ and the above theory is vacuous (as it should be).
- The Ichimura estimator achieves the semiparametric efficiency bound for estimation of β when the error is conditionally homoskedastic. Ichimura also considers a weighted least-squares estimator setting the weight to be the inverse of an estimate of the conditional variance function. This weighted LS estimator is then semiparametrically efficient.

Klein and Spady's Binary Choice Estimator

- Klein and Spady (Econometrica, 1993) proposed an estimator of the semiparametric single index binary choice model which has strong similarities with Ichimura's estimator.
- The model is

$$y_i = 1(X_i'\beta \geq e_i)$$

  where e_i is an error.
- If e_i is independent of X_i and has distribution function g, then the data satisfy the single index regression

$$E(y|x) = g(x'\beta)$$

  It follows that Ichimura's estimator can be directly applied to this model.

- Klein and Spady suggest a semiparametric likelihood approach. Given g, the log-likelihood is

$$L_n(\beta, g) = \sum_{i=1}^{n} (y_i \log g(X_i'\beta) + (1 - y_i) \log(1 - g(X_i'\beta))).$$

  This is analogous to the sum-of-squared errors function $S_n(\beta, g)$ for the semi-parametric regression model.

- Similarly with Ichimura, Klein and Spady suggest replacing g with the leave-one-out NW estimator

$$\hat{g}_{-i}(X_i'\beta) = \frac{\sum_{j \neq i} \mathcal{K}_h((X_j - X_i)'\beta) y_j}{\sum_{j \neq i} \mathcal{K}_h((X_j - X_i)'\beta)}$$

- Making this substitution, and adding trimming function, this leads to the feasible likelihood criterion

$$L_n(\beta) = \sum_{i=1}^{n} (y_i \log \hat{g}_{-i}(X_i'\beta) + (1-y_i) \log(1-\hat{g}_{-i}(X_i'\beta))) 1_i(b).$$

- Klein and Spady emphasize that the trimming indicator should not be a function of β, but instead of a preliminary estimator. They suggest

$$1_i(b) = 1(\hat{f}_{X'\tilde{\beta}}(X_i'\tilde{\beta}) \geq b)$$

  where $\tilde{\beta}$ is a preliminary estimator of β and $\hat{f}$ is an estimate of the density of $X_i'\tilde{\beta}$. Klein and Spady observe that trimming does not seem to matter in their simulations.

- The Klein-Spady estimator for β is the value $\hat{\beta}$ which maximizes $L_n(\beta)$.

- In many respects the Ichimura and Klein-Spady estimators are quite similar.
- Unlike Ichimura, Klein-Spady impose the assumption that the kernel $\mathcal{K}$ must be fourth-order (e.g. bias reducing). They also impose that the bandwidth h satisfy the rate $n^{-1/6} < h < n^{-1/8}$, which is smaller than the optimal $n^{-1/9}$ rate for a 4th order kernel. It is unclear to me if these are merely technical sufficient conditions, or if there a substantive difference with the semiparametric regression case.
- Klein and Spady also have no discussion about how to select the bandwidth. Following the ideas of Hardle, Hall and Ichimura, it seems sensible that it could be selected jointly with β by maximization of $L_n(\beta)$, but this is just a conjecture.

- They establish the asymptotic distribution for their estimator. Similarly as in Ichimura, letting g denote the distribution of e_i, define the function

$$G(X_i'\beta) = E(g(X_i'\beta_0)|X_i'\beta).$$

Then

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow N(0, H^{-1})$$

$$H = E\Big(\frac{\partial}{\partial\beta}G(X_i'\beta)\frac{\partial}{\partial\beta}G(X_i'\beta)'\frac{1}{g(X_i'\beta_0)(1 - g(X_i'\beta_0))}\Big)$$

- They are not specific about the derivative component, but if I understand it correctly it is the same as in Ichimura, so

$$\frac{\partial}{\partial \beta} G(X_i'\beta) \simeq g^{(1)}(X_i'\beta_0)(X_i - E(X_i|X_i'\beta_0)).$$

  The Klein-Spady estimator achieves the semiparametric efficiency bound for the single index binary choice model.

- Thus in the context of binary choice, it is preferable to use Klein-Spady over Ichimura. Ichimura's LS estimator is inefficient (as the regression model is heteroskedastic), and it is much easier and cleaner to use the Klein-Spady estimator rather than a two-step weighted LS estimator.

Average Derivative Estimator

- If X is continuous, Powell, Stock and Stoker (Econometrica, 1989) proposed a simple approach to estimate β.
- Let the conditional mean be

$$E(y|x) = \mu(x)$$

Then the derivative is

$$\mu^{(1)}(x) = \frac{\partial}{\partial x}\mu(x)$$

and a weighted average is

$$E(\mu^{(1)}(X)w(X))$$

where $w(x)$ is a weight function. It is particularly convenient to set $w(x) = f(x)$, the marginal density of X.

- Thus PSS define this as the average derivative

$$\delta = E(\mu^{(1)}(X)f(X))$$

  This is a measure of the average effect of X on y. It is a simple vector, and therefore easier to report than a full nonparametric estimator.
- There is a connection with the single index model, where

$$\mu(x) = g(x'\beta)$$

- For then

$$\mu^{(1)}(x) = \beta g^{(1)}(x'\beta), \quad \delta = c\beta$$

  where $c = E(g^{(1)}(x'\beta)f(X))$.
- Since β is identified only up to scale, the constant c doesn't matter. That is, a (normalized) estimate of δ is an estimate of normalized β.

- PSS observe that by integration by parts

$$\begin{aligned}
\delta &= E(\mu^{(1)}(X)f(X)) \\
&= \int \mu^{(1)}(x)f^2(x)dx \\
&= -2\int \mu(x)f(x)f^{(1)}(x)dx \\
&= -2E(\mu(X)f^{(1)}(X)) \\
&= -2E(yf^{(1)}(X)).
\end{aligned}$$

- By the reasoning in CV, an estimator of this is

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

  where $\hat{f}_{-i}(X_i)$ is the leave-one-out density estimator, and $\hat{f}_{-i}^{(1)}(X_i)$ is its first derivative.
- This is a convenient estimator. There is no denominator messing with uniform convergence. There is only a density estimator, no conditional mean needed.
- PSS show that $\hat{\delta}$ is $\sqrt{n}$ consistent and asymptotic normal, with a convenient covariance matrix.

Testing the SIM

- Horowitz & Härdle (1994) designed a test that considers the following hypotheses:

$$H_0 : E(Y|X = x) = U(x'\beta) \leftrightarrow H_1 : E(Y|X = x) = g(x'\beta) \tag{5}$$

  Here U (the link under H_0) is a known function and g (the link under H_1) an unspecified function. For example, the null hypothesis could be a logit model and the alternative a semiparametric model of SIM type.

- The main idea that inspires the test relies on the fact that if the model under the null is true then a nonparametric estimation of $E(Y|X'\hat{\beta} = v)$ gives a correct estimate of $F(v)$. Thus, the specification of the parametric model can be tested by comparing the nonparametric estimate of $E(Y|X'\hat{\beta} = v)$ with the parametric fit using the known link U.
- The test statistic is defined as

$$T = \sqrt{h} \sum_{i=1}^{n} w(X_i'\hat{\beta})[Y_i - U(X_i'\hat{\beta})][\hat{g}_{-i}(X_i'\hat{\beta}) - U(X_i'\hat{\beta})] \quad (6)$$

where $\hat{g}_{-i}(\cdot)$ is a leave-one-out NW estimate for the regression of Y on the estimated index values, h is the bandwidth used in the kernel regression. $w(\cdot)$ is a weight function that downweights extreme observations.

- In practice the weight function is defined as such that it considers only 90% or 95% of the central range of the index values of $X_i'\hat{\beta}$. Horowitz & Härdle (1994) propose to take $\hat{\beta}$, the estimate under H_0. That is, the same index values $X_i'\hat{\beta}$ are used to compute both the parametric and the semiparametric regression values.
- Let us take a closer look at the intuition behind this test statistic. The first difference term in the sum measures the deviation of the estimated regression from the true realization, that is it measures $Y_i - E(Y|X_i)$. If H_0 holds, then this measure ought to be very small on average. If, however, the parametric model under the null fails to replicate the observed values Y_i well, then T will increase. Obviously, we reject the hypothesis that the data were generated by a parametric model if T becomes unplausibly large.

- The second difference term measures the distance between the regression values obtained under the null and under the semiparametric alternative. Suppose the parametric model captures the characteristics of the data well so that $Y_i - U(X_i'\hat{\beta})$ is small. Then even if the semiparametric link deviates considerably from the parametric alternative on average, these deviations will be downweighted by the first difference term. Seen differently, the small residuals of the parametric fit are blown up by large differences in the parametric and semiparametric fits, $\hat{g}_{-i}(X_i'\hat{\beta}) - U(X_i'\beta)$. Thus, if H_0 is true, the residuals should be small enough to accommodate possible strong differences in the alternative fits. Again, a small statistic will lead to maintaining the null hypothesis.
- It can be shown that under H_0 and under some suitable regularity conditions T is asymptotically distributed as a $N(0, \sigma_T^2)$ where σ_T^2 denotes the asymptotic sampling variance of the statistic.

A Goodness-of-fit for SIM

- To test the significance of SIM, i.e.,

$$H_0 : P(E(Y|X = \cdot) = g(\beta'\cdot)) = 1, \quad for\ some\ \beta\ and\ g. \tag{7}$$

Let $g_\beta(v) = E(Y|X'\beta = v)$, and

$$\beta_0 = \arg \min_{\beta:\|\beta\|=1} E[Y - g_\beta(\beta' X)]^2.$$

- It easily can be seen that (7) is equivalent to

$$H_0 : P(E[(Y - g_{\beta_0}(\beta_0' X))|X] = 0) = 1. \tag{8}$$

or,

$$H_0 : E[(Y - g_{\beta_0}(\beta_0' X))I(X < x)] \equiv 0. \tag{9}$$

where $X < x$ means that every component of X is less than the corresponding component of x

- Suppose that $\{(X_i, Y_i) : i = 1, \ldots, n\}$ is a random sample. Let $\hat{Y}_i$'s be the fitted values of $E(Y|\beta' X)$ using some nonparametric method. Corresponding to (9), we construct the following residual marked empirical process

$$S_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) I(X_i \leq x). \tag{10}$$

- To calculate the fitted value $\hat{Y}_i$, we need to estimate $g(\cdot)$ and β. For fixed β, we estimate $g_\beta(v)$ using local linear kernel smoothing

$$\hat{g}_\beta(v) = \frac{\sum_{i=1}^{n} W_{n,h}(X_i'\beta - v) Y_i}{\sum_{i=1}^{n} W_{n,h}(X_i'\beta - v)} \tag{11}$$

where $W_{n,h}(X_i'\beta - v) = s_{n,\beta,2}(v) K_h(X_i'\beta - v)$- $s_{n,\beta,1} n^{-1} K_h(X_i'\beta - v)\{(X_i'\beta - v)/h\}$ with $s_{n,\beta,k}(v) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i'\beta - v)\{(X_i'\beta - v)/h\}^k, k = 0, 1, 2.$

- Here and later, $K(\cdot)$ is a kernel function, $K_h(\cdot) = K(\cdot/h)/h$ and h is a bandwidth.
- There are many methods to estimate the parameter β. See for example Härdle and Stoker (1989), Ichimura and Lee (1991), Härdle, Hall and Ichimura (1993) and Weisberg and Welsh (1994). Having obtained an estimate of β we estimate $g(v)$ by $\hat{g}_{\hat{\beta}}(v)$ as in (11), and obtain the fitted value of Y_i as $\hat{Y}_i = \hat{g}_{\hat{\beta}}\left(X_i^T\hat{\beta}\right)$ and hence the process $S_n(x)$.
- Denote

$$
\begin{aligned}
&l(x, g, \beta_0) \\
&= \Big[\int w(z)\{z - \mu(z,\beta_0)\}\{z - \mu(z,\beta_0)\}' g'(z'\beta_0)^2 f(z)dz\Big]^- \\
&\qquad \cdot (x - \mu(x,\beta_0)),
\end{aligned}
$$

where $\mu(x, \beta_0) = E(X|X'\beta_0 = x'\beta_0)$.

- Let

$$H(x) = \{I_D(X < x) - E\Big([g'(X'\beta_0)I_D(X < x)\{X - E(X|X'\beta_0)\}]'\Big)l(X, g, \beta_0) - E[I_D(X < x)|X'\beta_0]\}\epsilon,$$

where $\epsilon = Y - g(X'\beta)$ and
$I_D(X < x) = I(X'\beta_0 \in D)I(X < x)$ with D being a compact region on which $X'\beta_0$ has positive density.

- Denote

$$S_D(x) = \frac{1}{\sqrt{n}} \sum_{X_i'\hat{\beta} \in D} (Y_i - \hat{Y}_i)I(X_i < x).$$

and

$$B_D(x) = n^{1/10} E[g''(X'\beta_0)I_D(X < x)]/2$$

- **Theorem** Under some regularity conditions, we have under H_0,

$$S_D(x) + B_D(x) \Rightarrow Q(x)$$

where $Q(x)$ is a mean-zero Gaussian process with covairance function $EQ(x_1)Q(x_2) = E[H(x_1)H(x_2)]$. "$\Rightarrow$" denotes the weak convergence.

- There is a bias term for the residual marked empirical process $S_D(x)$, namely $B_D(x)$. We have to remove it if we want to use the process $S_D(x)$ for the purpose of testing. Therefore, we define the bias-corrected statistic as

$$CCV_D = \int [S_D(x) + B_D(x)]^2 dF_n(x),$$

where $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i < x)$. By above theorem, we have

$$CCV_D \to \int [Q(x)]^2 dF(x)$$

in distribution, where $F(x)$ is the cumulative distribution function X.

- As we have commented previously, the bias term will cause trouble in practice. In principle, it can be estimated using the usual method which, however, necessitates the selection of another bandwidth. Moreover, the limiting distribution still depends on the derivative of the unknown function $g(\cdot)$, which is difficult to estimate. It is hard to give a closed form for the distribution.
- Instead, we adopt the bootstrap approach to obtain an estimate of the bias and mimic the unknown distribution.

Boostrap method

- We now use Bootstrap method for the purpose of bias-correction and asymptotical distribution. We first estimate β_0 and $g(\cdot)$ as before, then generate independent bootstrap observations from the following model

$$Y_i^* = \hat{g}_{\hat{\beta}}(X_i'\hat{\beta}) + \epsilon_i^*, \tag{12}$$

  with $\epsilon_i^* = (Y_i - \hat{Y}_i)\eta_i^*$, where η_i^*'s are i.i.d random variables, each with zero mean, unit variance, finite moments of all orders and independent of $\{(X_i, Y_i), i = 1, \ldots, n\}$.

- Based on the Bootstrap samples, we can re-estimate β and $g(\cdot)$, and denote the estimators by $\hat{\beta}^*$ and $\hat{g}^*_{\hat{\beta}^*}(v)$, respectively. The bootstrap counterpart of $S_{\mathcal{D}}(x)$ is

$$S^*_{\mathcal{D}}(x) = n^{-1/2} \sum_{X_i^T\hat{\beta}\in\mathcal{D}} \left(Y_i^* - \hat{Y}_i^*\right) I\left(X_i < x\right)$$

Explanation of basic idea

- We temporarily assume that β_0 is known. Let $z_i = X_i^T \beta_0$. Then $\hat{g}_{\hat{\beta}}(X_i^T \hat{\beta})$ in (12) changes to $\hat{g}_{\beta_0}(z_i)$. Under some assumptions, we have

$$\hat{g}_{\beta_0}(v) = g_{\beta_0}(v) + \frac{1}{2} g''_{\beta_0}(v) h^2 + \frac{1}{n f_{\beta_0}(v)} \sum_{i=1}^{n} K_h (z_i - v) \varepsilon_i + o_P\left(h^2\right)$$

where $f_{\beta_0}(v)$ is the density function of $X^T \beta_0$. For each bootstrap sample, we have

$$\begin{aligned} \hat{g}^*_{\beta_0}(v) =& g_{\beta_0}(v) + g''_{\beta_0}(v) h^2 + \frac{1}{n f_{\beta_0}(v)} \sum_{i=1}^{n} K * K_h (z_i - v) \varepsilon_i \\ &+ \frac{1}{n f_{\beta_0}(v)} \sum_{i=1}^{n} K_h (z_i - v) \varepsilon_i^* + o_P\left(h^2\right) \end{aligned} \tag{13}$$

where $K * K$ denotes the convolution of K.

- The bias for $\hat{g}^*_{\beta_0}(x)$ is

$$E\left[\hat{g}^*_{\beta_0}(v) - \hat{g}_{\beta_0}(v) \middle| (z_i, Y_i), i = 1, \ldots, n\right] = \frac{1}{2} g''(v) h^2$$
$$+ \frac{1}{n f_{\beta_0}(v)} \sum_{i=1}^{n} \left\{K * K_h(z_i - v) - K_h(z_i - v)\right\} \varepsilon_i + o_P\left(h^2\right) \tag{14}$$

- Note that the second term on the right hand side above is $O_P\left(h^2\right)$ (as h is proportional to $n^{-1/5}$) . Equation (14) implies that $\hat{g}_{\beta_0}(\cdot)$ and $\hat{g}^*_{\beta_0}(\cdot)$ have different bias terms. Therefore if we try to make a pointwise inference about the regression function g, we have to use another bandwidth and oversmooth the regression function such that the second term in (14) is $o_P(h^2)$.

- However, the difference in (14) can be reduced by the summation of the residual marked empirical process in our problem, namely

$$n^{-1/2}\sum_{z_i\in\mathcal{D}}\left\{nf_{\beta_0}(z_i)\right\}^{-1}\sum_{j=1}^{n}\{K*K_h(z_i-v)-K_h(z_i-v)\}\varepsilon_i=o_P(1) \tag{15}$$

- because $\int\{K*K(v)-K(v)\}dv=0$. By (15), we can show that $S^*_{\mathcal{D}}(x)$ and $S_{\mathcal{D}}(x)$ have the same bias term asymptotically. Note that by (13), the bias $E\left[\hat{g}^*_{\beta_0}(v)-\hat{g}_{\beta_0}(v)|\left(z_i,Y_i\right),i=1,\ldots,n\right]$ can be obtained by the average of the resample. Therefore the bias terms in the $S^*_{\mathcal{D}}(x)$ and $S_{\mathcal{D}}(x)$ can be easily calculated and removed.

- Let $\tilde{Y}_i = \hat{Y}_i - B_i$, $\tilde{Y}_i^* = \hat{Y}_i^* - B_i$, where $\hat{Y}_i^* = \hat{g}^*_{\hat{\beta}^*}(X_i'\hat{\beta}^*)$, $B_i = E[\hat{g}^*_{\hat{\beta}^*}(X_i'\hat{\beta}) - \hat{g}_{\hat{\beta}}(X_i'\hat{\beta})|(X_i, Y_i), i = 1, \ldots, n]$. Thus, $\tilde{Y}_i$ and $\tilde{Y}_i^*$ can be seen as the bias-corrected version of Y_i and Y_i^*.
- Let

$$\tilde{S}_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - \tilde{Y}_i) I(X_i < x),$$

and

$$\tilde{S}_n^*(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i^* - \tilde{Y}_i^*) I(X_i < x)$$

- **Theorem** Under some regularity conditions, we have under H_0,

$$\tilde{S}_n(x) \Rightarrow Q(x), \qquad \tilde{S}_n^*(x) \Rightarrow Q(x).$$

- Let
$$T_n = \int \tilde{S}_n(x) dF_n(x)$$
- By the above Bootstrap procedure and theorem, we can use the Bootstrap version
$$T_n^* = \int \tilde{S}_n^*(x) dF_n(x)$$
to approximate the distribution of T_n. Therefore, the p-value is
$$p - value = P(T_n^* \geq T_n) \approx \frac{1}{M} \sum_k I(T_{n,k}^* \geq T_n).$$
- For details, please refer to Xia et al. (2004). (Statistica Sinica, 14:1-39)