Nonparametric Methods in Economics and Finance Lecture 2

Oliver Linton

May 10, 2006

Confidence Intervals

- We first consider confidence intervals for density estimates. Consistent estimates of $v(x)$ can easily be obtained, indeed

  $$\widehat{v}(x) = \|K\|_2^2 \, \widehat{f}(x)$$

  is one such.

- Estimation of the bias is more complicated because it depends on $f''(x)$, but it can be also be consistently estimated by

  $$\widehat{b}(x) = (nh^5)^{1/2} \frac{\mu_2(K)}{2} \widehat{f}''(x),$$

  where $\widehat{f}''(x)$ is a kernel estimate of $f''(x)$.

- It is necessary to use a different bandwidth in $\widehat{f}''(x)$ to obtain consistency.

- Then

$$\mathcal{C}_\alpha = \widehat{f}(x) - \widehat{b}(x) \pm z_{\alpha/2}\widehat{v}(x)^{1/2}$$

  is a valid $1 - \alpha$ confidence set in the sense that

$$\Pr[f(x) \in \mathcal{C}_\alpha] \to 1 - \alpha.$$

- For regression. The asymptotic distribution contained in the above results can be used to calculate pointwise confidence intervals for the local constant and local linear estimators.

- In practice it is usual to ignore the bias term, since this is rather complicated, depending on higher derivatives of the regression function and perhaps on the derivatives of the density of X. This approach can be justified when a bandwidth is chosen that makes the bias relatively small.

- That is, we suppose that h^2 is small relative to $1/(nh)^{1/2}$, i.e., $h = o(n^{-1/5})$. In this case, the interval

  $$\mathcal{C}_\alpha = \widehat{m}(x) \pm z_{\alpha/2} \left(\widehat{\mathsf{var}}\left[\widehat{m}(x)\right]\right)^{1/2},$$

  where $\widehat{\mathsf{var}}[\widehat{m}(x)]$ is a consistent estimate of the asymptotic variance of $\widehat{m}(x)$, is a valid $1-\alpha$ confidence set in the sense that

  $$\mathsf{Pr}[m(x) \in \mathcal{C}_\alpha] \to 1 - \alpha.$$

- To get estimates of var$[\widehat{m}(x)]$ we exploit the linearity of the local constant and local linear estimates. That is, they both can be written in the form
$$\widehat{m}(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i,$$
where $\{w_{ni}(x)\}$ only depend on the design. This is also true of a much wider class of estimators than kernels or local linear.

- One could argue that the conditional distribution is the appropriate framework for inference here, since the covariates are ancillary. In this case the calculations leading to the asymptotic variance are particularly easy for any linear smoother of this type. We have
$$\text{var}\left\{\widehat{m}(x) \,|X_1, \ldots, X_n\right\} = \sum_{i=1}^{n} w_{ni}^2(x)\sigma_i^2,$$
where $\sigma_i^2 = E(\varepsilon_i^2|X_i)$. Note that this is exactly true for any linear smoother.

- If the error terms were normally distributed, then $\widehat{m}(x)$ itself is also normally distributed, conditional on the design.

- In general, although we may not be able to prove it, we can expect that $\widehat{m}(x)$ is asymptotically normal after location and scale adjustment. Thus we expect that under appropriate regularity conditions,
$$
\begin{aligned}
& \frac{\widehat{m}(x) - E\left\{\widehat{m}(x)\,|X_1, \ldots, X_n\right\}}{\left(\sum_{i=1}^n w_{ni}^2(x)\widehat{\varepsilon}_i^2\right)^{1/2}} \\
= \quad & \frac{\widehat{m}(x) - E\left\{\widehat{m}(x)\,|X_1, \ldots, X_n\right\}}{\left(\sum_{i=1}^n w_{ni}^2(x)\sigma_i^2\right)^{1/2}} + o_p(1) \\
\Longrightarrow \quad & N(0,1),
\end{aligned}
$$
where $\widehat{\varepsilon}_i = Y_i - \widehat{m}(X_i)$ are the nonparametric residuals.

- This result is the basis for confidence intervals for any linear smoother. This case includes splines, series, local polynomial, nearest neighbors and the many hybrid modifications thereof. It also includes multidimensional estimates and standard estimates of derivatives.

- Thus the confidence interval becomes

$$\mathcal{C}_\alpha = \widehat{m}(x) \pm z_{\alpha/2} \left(\widehat{v}_1(x)\right)^{1/2},$$

$$\widehat{v}_1(x) = \sum_{i=1}^{n} w_{ni}^2(x) \widehat{\varepsilon}_i^2.$$

- One could instead separate out the estimator of variance from the weights,

$$\widehat{v}_2(x) = \widehat{\sigma}^2(x) \sum w_{ni}^2(x),$$

$$\widehat{\sigma}^2(x) = \sum_{i=1}^{n} w_{ni}(x) \widehat{\varepsilon}_i^2.$$

- Some authors also impose homoskedasticity in construction of confidence intervals since there are a number of good estimators of σ^2 in that case like

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2.$$

- In special cases there are additional estimators based on the specific structure of the limiting distribution. So for kernel and local linear estimators we might consider

$$\widehat{v}_3(x) = \frac{1}{nh}\frac{\widehat{\sigma}^2(x)}{\widehat{f}(x)}||K||_2^2,$$

where $\widehat{f}(x)$ is the standard kernel density estimate.

- Typically one finds that

$$\widehat{v}_1(x) \geq \widehat{v}_2(x) \geq \widehat{v}_3(x)$$

although the difference is not large.

- In many cases, we have the additional condition that $\sum_{i=1}^{n} w_{ni}(x) = 1$ and so

$$\begin{aligned} & \widehat{m}(x) - m(x) \\ = & \sum_{i=1}^{n} w_{ni}(x)\varepsilon_i \\ & + \sum_{i=1}^{n} w_{ni}(x)\{m(X_i) - m(x)\}. \end{aligned}$$

- We have

$$\begin{aligned} \text{var}[\widehat{m}(x)] &= E(\text{var}[\widehat{m}(x)\,|X_1, \ldots, X_n]) \\ &\quad +\text{var}(E[\widehat{m}(x)\,|X_1, \ldots, X_n]), \end{aligned}$$

where the second term is of smaller order when $\sum_{i=1}^{n} w_{ni}(x) = 1$.

- It follows that

$$\begin{aligned} \text{var}[\widehat{m}(x)] &\simeq \text{var}[\widehat{m}(x)\,|X_1, \ldots, X_n] \\ &= \sum_{i=1}^{n} w_{ni}^2(x)\sigma^2(X_i), \end{aligned}$$

so that conditional and unconditional variance are approximately the same.

- Some estimators that do not satisfy $\sum_{i=1}^{n} w_{ni}(x) = 1$ have var$(E[\widehat{m}(x)\,|X_1, \ldots, X_n])$ of the same magnitude as $E($var$[\widehat{m}(x)\,|X_1, \ldots, X_n])$ and so the asymptotics are different. One might argue that one should only care about the var$[\widehat{m}(x)\,|X_1, \ldots, X_n]$ because of the ancillarity of the covariate.

- Can also subtract off the mean from the residuals $\widehat{\varepsilon}_i$ since they are not guaranteed to have mean zero and do not. This does not affect the consistency of the standard error estimates but it can affect the bias.

Uniform confidence bands

- The main use of the limit theorem for the supremum error is in setting uniform confidence intervals. The confidence intervals we have provided
$$\widehat{m}(x) \pm z_{\alpha/2}\,(\mathsf{var}[\widehat{m}(x)])^{1/2}$$
have been valid for a single point. However, we are usually interested in the function m at a number of different points in which case simply plotting out the above interval for each x will not give the right level.

- There are two main approaches to providing correct confidence intervals. One is to use Bonferroni type inequalities to correct the level [see Savin (1984) and Härdle (1991) for further discussion of this] and the second approach is to treat the function $\widehat{m}(\cdot)$ as a random variable and use stochastic process limit theory.

- In other words, we find a set of functions $\mathcal{C}(\widehat{m})$ with the property that

$$\Pr\left[m \in \mathcal{C}(\widehat{m})\right] = 1 - \alpha$$

  for large n. This is provided by the limit theory by letting

$$\mathcal{C}(\widehat{m}) = \left\{m(\cdot) : a_n(T_n - b_n) \leq c_\alpha\right\},$$

  where c_α solves $\exp(-2\exp(-c_\alpha)) = 1-\alpha$, which leads to bands of the form

$$\widehat{m}(x) \pm (b_n + \frac{c_\alpha}{a_n})\,(\widehat{\text{var}}[\widehat{m}(x)])^{1/2} \quad \text{all } x,$$

  where $\widehat{\text{var}}[\widehat{m}(x)]$ is some estimate of $\text{var}[\widehat{m}(x)]$.

- This intervals has the correct coverage. In practice, these intervals do not work terribly well for the reasons discussed in Hall (1993). A better approach is based on the bootstrap, which we will cover later on.

Optimality

- Stone (1982) established what is the optimal rate of convergence for nonparametric regression under certain conditions. In particular for a class of distributions for (Y, X) he found the sequence b_n such that for positive constants c

  $$\lim_{n\to\infty} \inf_{\widehat{m}\in\mathcal{M}} \sup_{m\in M} \Pr\left[||\widehat{m} - m||_q \geq cb_n\right] = 1.$$

  Here, $q \in (0, \infty]$ and the norm is taken over a compact set $D \subseteq R^d$.

- The set $\mathcal{M}$ includes all estimators.

- The set M determines the difficulty of the estimation problem.

  - When M includes just functions from a particular parametric class, one can usually obtain rate $n^{-1/2}$.

  - When M includes d-dimensional functions that are p times continuously differentiable on D, the optimal rate

    * For $q < \infty$, is $n^{-p/(2p+d)}$

    * For $q = \infty$, is $(n/\log n)^{-p/(2p+d)}$

- This bound is achievable if there exists an estimator $\widehat{m}$ such that

$$\lim_{n\to\infty} \sup_{m\in M} \Pr\left[||\widehat{m} - m||_q \geq c' b_n\right] = 0$$

for some constant c'. Stone exhibited a rate optimal estimator.

- Fan (1993) has investigated optimality under a mean squared error criterion. Let

  $$R_n(M, \mathcal{M}) = \inf_{\widehat{m} \in \mathcal{M}} \sup_{m \in M} E\left[(\widehat{m}(x) - m(x))^2\right]$$

  be the pointwise MSE optimal bound for an interior point x.

- He showed that the (best) local linear estimator comes within 0.896^2 of the bound asymptotically when $\mathcal{M}$ is chosen to include all estimators.

- He also showed that when $\mathcal{M}$ is restricted to the class of estimators linear in Y, the (best modified) local linear estimator achieves the bound. In this sense the local linear estimator is Best Linear Asymptotic Minimax (BLAM). Fan modified the local linear by the inclusion of a trimming factor of order n^{-2} in the denominator to ensure that the moments existed.

- We consider some non-asymptotic results. Suppose that we consider the criterion

$$Q = \sum_{i=1}^{n} E\left[(\widehat{m}(X_i) - m(X_i))^2 \, | X_1, \ldots, X_n \right]$$

otherwise known as the trace mean squared error criterion associated with the $n \times 1$ vector $\widehat{m} = (\widehat{m}(X_1), \ldots, \widehat{m}(X_n))^\top$.

- Suppose that $\widehat{m}$ is linear, i.e.,

$$\widehat{m} = Wy,$$

where $y = (Y_1, \ldots, Y_n)^\top$ and W is an $n \times n$ matrix just depending on the covariates.

- Write the regression model as

$$y = m + \varepsilon,$$

where $m = (m(X_1), \ldots, m(X_n))^\top$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ and suppose that $E[\varepsilon | X] = 0$ and $E\left[\varepsilon\varepsilon^\top | X\right] = \sigma^2 I_n$.

- Then

$$Q = \sigma^2 \mathsf{tr}(WW^\top) + \mathsf{tr}(bb^\top),$$

where the bias is $b = (W - I_n)m$ and the first term is the variance.

- Define the symmetric matrix

$$W_c = I_n - \left[(W - I_n)^\top (W - I_n)\right]^{1/2}.$$

Cohen (1966) showed that the estimator

$$\widehat{m}_c = W_c y$$

has smaller Q - in particular, its bias is the same but its variance is smaller unless W is symmetric, specifically

$$\mathsf{tr}(W_c W_c^\top) \leq \mathsf{tr}(WW^\top).$$

- This follows because

$$\begin{aligned} & W_c^\top W_c - W^\top W \\ = \; & I_n + \Big[(W-I_n)^\top (W-I_n)\Big] \\ & -2\Big[(W-I_n)^\top (W-I_n)\Big]^{1/2} - W^\top W \\ = \; & 2(I_n - \widetilde{W}) \\ & -2\Big[(W-I_n)^\top (W-I_n)\Big]^{1/2}, \end{aligned}$$

  where $\widetilde{W} = (W + W^\top)/2$.

- Then

$$\begin{aligned} & \mathrm{tr}((I_n - \widetilde{W})^2) \\ \leq \; & \mathrm{tr}((W-I_n)^\top (W-I_n)) \end{aligned}$$

  is equivalent to showing that

$$\mathrm{tr}(\widetilde{W}^2 - WW^\top) \leq 0.$$

- This follows because we can write

$$\begin{aligned} & \mathrm{tr}(\widetilde{W}^2 - WW^\top) \\ = \; & -\mathrm{tr}((W - W^\top)^\top (W - W^\top))/4 \\ \leq \; & 0. \end{aligned}$$

- This says that any estimator of m of the linear form for which W is not symmetric is inadmissible according to the trace mean squared error criterion and gives a concrete way of improving estimators.

- Kernel and local polynomial estimators have asymmetric W matrices, and are inadmissible, although asymptotically this inadmissibility disappears as we know.

- Only spline estimators amongst the commonly used estimators have symmetric W.

Bandwidth Selection

- In this lecture we describe several methods of bandwidth selection for nonparametric regression estimation. We first define some performance criteria for an estimate $\widehat{m}(\cdot)$ of the function $m(\cdot)$. In the sequel $\pi(\cdot)$ is some weighting function defined on the support of X.

1. Pointwise MSE

$$d_{MP}(\widehat{m}(x), m(x)) = E\left[\{\widehat{m}(x) - m(x)\}^2\right]$$

2. Integrated MSE

$$d_{MI}(\widehat{m}, m) = \int E\left[\{\widehat{m}(x) - m(x)\}^2\right] \pi(x) dx$$

3. Average S.E.

$$d_A(\widehat{m}, m) = \frac{1}{n} \sum_{i=1}^{n} \{\widehat{m}(X_i) - m(X_i)\}^2 \pi(X_i)$$

4. Integrated S.E.

$$d_I(\widehat{m}, m) = \int \{\widehat{m}(x) - m(x)\}^2 \pi(x) f(x) dx$$

5. Conditional MISE

$$d_C(\widehat{m}, m) = E\{d_I(\widehat{m}, m) | X_1, \ldots, X_n\}.$$

Let h_j be the bandwidth sequences that minimize the corresponding criterion d_j.

- The mean squared error criteria actually have explicit formulae for the optimal bandwidth. Recall that for univariate local linear regression, the asymptotic mean squared error at the point x is

$$\begin{aligned} & d_{MP}(\widehat{m}(x), m(x)) \\ \cong\ & \frac{1}{nh}\frac{\sigma^2(x)}{f(x)} \int K^2(u) du \\ & + \frac{h^4}{4}\left\{m''(x)\right\}^2 \left(\int u^2 K(u) du\right)^2 \end{aligned}$$

$$\equiv \frac{a(x)}{nh} + b(x)h^4.$$

- An optimal bandwidth can be defined as one that minimizes this criterion; this bandwidth will satisfy the following first order condition

$$\frac{a(x)}{nh^2} = 4b(x)h^3,$$

which solves to give

$$h_{MP}(x) = \left[\frac{a(x)}{4b(x)}\right]^{1/5} n^{-1/5}.$$

- So the optimal bandwidth depends on the unknown quantities: $\sigma^2(x), f(x)$, and $m''(x)$, and changes with each point x. Frequently, people work with an Integrated mean squared error criterion $d_{MI}(\widehat{m}(x), m(x))$, in which case the optimal bandwidth is

$$h_{MI} = \left[\frac{\int a(x)\pi(x)dx}{4\int b(x)\pi(x)dx}\right]^{1/5} n^{-1/5},$$

and the optimal bandwidth depends on only averages of $\sigma^2(x), f(x)$, and $m''(x)$.

Plug-in

- This involves nonparametrically estimating the unknown quantities in $a(x)$ and $b(x)$ by $\widehat{a}(x)$ and $\widehat{b}(x)$, say, and then let

$$\begin{aligned}\widehat{h}_{MP}(x) &= \left[\frac{\widehat{a}(x)}{4\widehat{b}(x)}\right]^{1/5} n^{-1/5} \\ \widehat{h}_{MI} &= \left[\frac{\int \widehat{a}(x)\pi(x)dx}{4\int \widehat{b}(x)\pi(x)dx}\right]^{1/5} n^{-1/5}.\end{aligned}$$

- Provided $\widehat{a}(x)P{\longrightarrow}a(x)$ and $\widehat{b}(x)P{\longrightarrow}b(x)$, then

$$\frac{|\widehat{h}_{MP}(x) - h_{MP}(x)|}{h_{MP}(x)} P{\longrightarrow}0,$$

while if

$$\sup_{x:\pi(x)>0} |\widehat{a}(x) - a(x)|\, P{\longrightarrow}0$$

$$\sup_{x:\pi(x)>0} \left|\widehat{b}(x) - b(x)\right| P{\longrightarrow}0,$$

Then

$$\sup_{x:\pi(x)>0} \frac{|\widehat{h}_{MI}(x) - h_{MI}(x)|}{h_{MI}(x)} P \longrightarrow 0.$$

- The disadvantage of this method is that one must estimate the derivatives of m and f, which are typically poorly behaved estimates. The variance of a kernel estimate of $m''(x)$ is of order $1/nh^5$ and the bias can be arbitrarily bad unless some additional smoothness is assumed.

- Silverman (1986) suggests a compromise method he called rule of thumb. This involves specifying an auxiliary parametric model for the data distribution and using this to infer a simple formula for the optimal bandwidths. In density estimation his approach yields the simple formula

$$\widehat{h} = 1.06\widehat{\sigma}n^{-1/5},$$

where $\widehat{\sigma}$ is the estimated standard deviation (this can be the sample standard deviation or the interquartile range divided by 1.3). This is based on a normal distribution model and a Gaussian kernel.

- In regression this approach is more complicated because one has to specify $m(.), f(.)$, and $\sigma^2(.)$. Fan and Gijbels (1996, p67) give a convenient formula for local linear regression. Suppose that we take

$$\pi(x) = \pi_0(x) f(x)$$

and parametric regression function $m_\theta(x)$, with $m_\theta''(x) \neq 0$, and assume that the error is i.i.d. with variance σ^2. Then the optimal bandwidth can be estimated by

$$\widehat{h}_{opt} = C_{0,1}(K) \left[\frac{\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2 \pi_0(X_i)}{\frac{1}{n}\sum_{i=1}^n m_{\widehat{\theta}}''(X_i)\pi_0(X_i)} \right]^{1/5} n^{-1/5},$$

where

$$\widehat{\varepsilon}_i = Y_i - m_{\widehat{\theta}}(X_i)$$

are the parametric residuals and $\widehat{\theta}$ is an estimate of θ, while

$$C_{0,1}(K) = \left(\frac{\int K^2(t)dt}{[\int t^2 K(t)dt]^2} \right)^{1/5}.$$

For the Gaussian kernel, $C_{0,1}(K) = 0.776$. It is convenient to take $m_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2/2$, in which case $m_\theta''(x) = \theta_2$.

- Krieger and Picklands (1981) show in the context of pointwise density estimation that the resulting plug-in estimator is asymptotically efficient. Specifically, they assumed only twice continuously differentiable density and showed that for any consistent (in the relative sense) estimator $\widehat{h}_{opt}$ of h_{opt} the resulting estimator $\widehat{f}_{\widehat{h}_{opt}}(x)$ asymptotically has the same mean squared error as $\widehat{f}_{h_{opt}}(x)$. They also constructed a consistent bandwidth sequence $\widehat{h}_{opt}$.

- Their arguments were based on weak convergence of the local in bandwidth empirical process. See Einmahl and Mason (2005) for a recent extension of this theory.

Cross Validation

- This approach is based on an approximation to ASE or ISE. Thus

$$\begin{aligned} d_A(\widehat{m}, m) &= \frac{1}{n}\sum_{i=1}^{n}\{\widehat{m}(X_i) - m(X_i)\}^2\,\pi(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\widehat{m}(X_i)^2\pi(X_i) \\ &\quad -\frac{2}{n}\sum_{i=1}^{n}\widehat{m}(X_i)m(X_i)\pi(X_i) \\ &\quad +\frac{1}{n}\sum_{i=1}^{n}m(X_i)^2\pi(X_i). \end{aligned}$$

- The last term does not depend on the bandwidth, so we drop it from consideration. The first term just depends on the data and so can be computed easily. The problem arises with the second term, and in particular

$$\sum_{i=1}^{n}\widehat{m}(X_i)m(X_i)\pi(X_i)/n.$$

- We clearly can't just substitute $m(X_i)$ by $\widehat{m}(X_i)$. However, we might replace it by an unbiased estimator [in the conditional distribution], which is Y_i. This is the equivalent to taking

  $$R(h) = \frac{1}{n}\sum_{i=1}^{n} \{Y_i - \widehat{m}(X_i)\}^2 \pi(X_i)$$

  as the bandwidth criterion. Unfortunately, this method will lead us to select $h = 0$ always, because then $\widehat{m}(X_i) = Y_i$ for all i.

- What has gone wrong? The problem is that $\widehat{m}(X_i)$ depends on all the $Y's$ in the sample, i.e.,

  $$\begin{aligned} \widehat{m}(X_i) &= \sum_{j=1}^{n} w_{ij}Y_j \\ &= w_{ii}Y_i + \sum_{j\neq i} w_{ij}Y_j, \end{aligned}$$

  where w_{ij} are the smoother weights,

- so that

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(X_i) Y_i \pi(X_i) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} w_{ii} Y_i^2 \pi(X_i) \\
& + \frac{1}{n}\sum\sum_{i \neq j} w_{ij} Y_j Y_i \pi(X_i).
\end{aligned}
$$

- We have

$$
\begin{aligned}
& E\left[\frac{1}{n}\sum_{i=1}^{n} w_{ii} Y_i^2 \pi(X_i) | X_1, \ldots, X_n\right] \\
= \; & \frac{1}{n}\sum_{i=1}^{n} w_{ii}\{m^2(X_i) + \sigma^2(X_i)\}\pi(X_i) \\
\simeq \; & \frac{K(0)}{nh}\frac{1}{n}\sum_{i=1}^{n} \frac{\{m^2(X_i) + \sigma^2(X_i)\}\pi(X_i)}{f(X_i)},
\end{aligned}
$$

which is the same magnitude as the variance effect we are trying to pick up. Therefore, $R(h)$ is a downward biased estimator of $d_A(\widehat{m}, m)$.

- There are two solutions to this problem. First, we can estimate

$$\sum_{i=1}^{n} \widehat{m}(X_i)^2 \pi(X_i)/n$$
$$\sum_{i=1}^{n} \widehat{m}(X_i) m(X_i) \pi(X_i)/n$$

by

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{m}_i(X_i) Y_i \pi(X_i)$$
$$\frac{1}{n}\sum_{i=1}^{n} \widehat{m}_i^2(X_i) \pi(X_i),$$

where $\widehat{m}_i(X_i)$ is the leave-out-"i" estimator. In the local constant case this is:

$$\begin{aligned} \widehat{m}_i(x) &= \frac{\frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x-X_j}{h}\right) Y_j}{\widehat{f}_i(x)} \\ \widehat{f}_i(x) &= \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x-X_j}{h}\right). \end{aligned}$$

- In conclusion, let

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widehat{m}_i(X_i)\}^2 \, \pi(X_i).$$

- Choose $\widehat{h}_{cv} \in H_n$ to minimize $CV(h)$ for some set H_n, and then let $\widehat{m}_{\widehat{h}_{cv}}(\cdot)$.

- An equivalent method which has some advantages computationally, is to let

$$\begin{aligned} CV(h) &= \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widehat{m}(X_i)\}^2 \, \pi(X_i) \\ &\quad + \frac{2}{n} \sum_{i=1}^{n} w_{ii} Y_i^2 \pi(X_i). \end{aligned}$$

  This latter approach is similar in spirit to the model selection ideas of time series.

- We next give a theorem due to Härdle and Marron (1985) [see also Stone (1984) for density estimation], which established the optimality of this method for local constants.

Theorem 0.1 *Suppose that the following assumptions are satisfied:*

1. $H_n = [\underline{h}(n), \bar{h}(n)]$

   $$\underline{h}(n) \geq C^{-1} n^{\delta-1},\ \bar{h}(n) \leq C n^{-\delta},\ C, \delta > 0$$

2. K *is Hölder continuous, i.e.,*

   $$|K(x_1) - K(x_2)| \leq c|x_1 - x_2|^{\xi},\ \xi > 0$$

   and $\int |u|^{\xi} K(u) du < \infty$*.*

3. *The regression function* m *and the marginal density* f *are Hölder continuous.*

4. *The conditional moments are bounded by constants* C_i

   $$E(|Y|^i | X = x) \leq C_i \quad \text{for all } x, \text{ for } i = 1, 2, \ldots$$

5. *The marginal density $f(x)$ of x is compactly supported and is bounded from below on the support of w.*

   Then the bandwidth $\widehat{h}_{cv}$ is asymptotically optimal with respect to distances d_A, d_I and d_C, in the sense that with probability one

$$\frac{d(\widehat{m}_{\widehat{h}_{cv}}, m)}{\inf_{h \in H_n} d(\widehat{m}_h, m)} \to 1$$

$$\frac{\widehat{h}_{cv}}{\widehat{h}_{opt}} \to 1.$$

- The conditions of this theorem are very weak in some respects. Specifically, the amount of smoothness assumed for m and f is almost nil. This means that the bandwidth selection method is automatically adapting to the amount of smoothness. In the full proof one must take account of an general magnitude for $d(\widehat{m}_h, m)$ and of a 'parameter' set that is much larger than the one we considered, which is why the theorem is stated in this fashion. Finally, in the special case we worked with one can also establish the stronger result

$$n^{2/5}\left\{\widehat{m}_{\widehat{h}_{cv}}(x)-\widehat{m}_{\widehat{h}_{opt}}(x)\right\} P\longrightarrow 0.$$

The Bootstrap

- The bootstrap is a very popular method for obtaining confidence intervals or performing hypothesis tests. There can be computational reasons why this method is preferred to the usual approach based on estimating the unknown quantities of the asymptotic distribution. There can also be statistical reasons why the bootstrap is better than the asymptotic plug-in approach. The bootstrap has been shown to work in a large variety of situations, we are just going to look at the simplest i.i.d. cases.

- Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution function F. We have a statistic (root)

  $$R_n(\tau; X_1, \ldots, X_n; F),$$

  which is a function of the data $X_1, \ldots, X_n$ and a parameter value τ. For example R_n could be an estimator or a test statistic. Let

  $$H_n(x, F) = \Pr(R_n \leq x),$$

where the probability is calculated under the true distribution F. The question is, how to estimate $H_n(x, F)$ and functions thereof.

- The 'asymptotic' approach uses the fact that

  $H_n(x,\, F) \longrightarrow H(x,\, F)$ as $n \to \infty$ by CLT or other met

  then estimate $H_n(x, F)$ by

  $$\widehat{H}_A(x) = H(x,\, F_n),$$

  where F_n is some estimate of F like the empirical distribution. For example,

  $$R_n = n^{1/2}(\bar{X} - \mu) \Longrightarrow N(0,\, \sigma^2),$$

  and we approximate the distribution of R_n by $N(0, \widehat{\sigma}^2)$, where $\widehat{\sigma}^2$ is some consistent estimate of σ^2. In some cases H does not depend on F; then R_n is a pivot or asymptotic pivot.

- The Bootstrap approach is based on

  $$\widehat{H}_B(x) = H_n(x,\, F_n).$$

In fact, we make a further approximation by using Monte-Carlo methods to find H_n.

- The probability measure of the data $X_1, \ldots, X_n$ is denoted P_n, this is discrete with probability $1/n$ at each sample point. Let $X_1^*, \ldots, X_m^*$ be a sample from P_n and let $R_n^* = R_n(\hat{\tau}_n; X_1^*, \ldots, X_m^*; F_n)$. Then

$$\mathcal{L}(R_n^* | X_1, \ldots, X_n) = \widehat{H}_B.$$

Actually use T replications to approximate this distribution by an 'empirical'. Usually, $m = n$.

- Note that F_n could be the empirical distribution, i.e., $F_n(x) = \sum_{i=1}^n \{X_i \leq x\}/n$ or an estimate parametric c.d.f. $F_{\widehat{\theta}}$. In the latter case, the resampling is from the distribution $F_{\widehat{\theta}}$.

Theorem 0.2 *(Bickel and Freedman 1986) Suppose that $X_1, \ldots, X_n$ are i.i.d. with finite mean μ and positive variance σ^2. Along almost all sample sequences $\{X_1, \ldots, X_n\}$, as $n, m \to \infty$*

$$\mathcal{L}\left\{m^{1/2}(\mu_m^* - \mu_n)|X_1, \ldots, X_n\right\} \Longrightarrow N(0, \sigma^2),$$

where $\mu_m^* = m^{-1}\sum_{i=1}^m X_i^*$ *and* $\mu_n = n^{-1}\sum_{i=1}^n X_i$.

- In conclusion, we have found an alternative way to approximate the distribution of $n^{1/2}(\mu_n - \mu)$: just tabulate the distribution of $m^{1/2}(\mu_m^* - \mu_n)$ conditional on $X_1, \ldots, X_n$. In some cases this can be done exactly, but more often one approximates this distribution by a further step based on resampling. This idea can be used to obtain confidence intervals or to obtain critical values for tests.

Confidence interval for the mean

- The asymptotic approach:

$$C_{n,A} = \{t : R_n(X_1, \ldots, X_n; t) \leq \widehat{H}_A^{-1}(\alpha)\}$$

$$\widehat{H}_A^{-1}(\alpha) = \mathsf{sup}\{x : \widehat{H}_A(x) \leq \alpha\}.$$

  Then $\mathsf{Pr}(\tau \in C_{n,A}) \to \alpha$.

- For the Bootstrap, we let

$$C_{n,B} = \{t : R_n(X_1, \ldots, X_n; t) \leq \widehat{H}_B^{-1}(\alpha)\}$$

$$\widehat{H}_B^{-1}(\alpha) = \mathsf{sup}\{x : \widehat{H}_B(x) \leq \alpha\}$$

  Then $\mathsf{Pr}(\tau \in C_{n,B}) \to \alpha$.

- To carry out a test of $\tau = \tau_0$ reject if $\widehat{\tau} \notin C_{n,B}$ or $C_{n,A}$ or let $\hat{C}_B(\alpha)$ satisfy

$$H_B^{-1}(\alpha) = \hat{C}_B(\alpha)$$

- For asymptotic tests and confidence intervals you get the same result if you use $n^{1/2}(\mu_n - \mu)$ or $n^{1/2}(\mu_n - \mu)/s_n$, where s_n is the sample standard deviation. Not true for bootstrap. There is a difference between using pivotal or non-pivotal statistics $m^{1/2}(\mu^*_m - \mu_n)$ or $m^{1/2}(\mu^*_m - \mu_n)/s_n$ or $m^{1/2}(\mu^*_m - \mu_n)/s^*_m$. In either case we find $\widehat{H}_B(x)$ and hence $\widehat{H}_B^{-1}(\alpha)$.

Nonparametric Density Estimation

- Suppose that $X_1, \ldots, X_n$ are i.i.d. with twice continuously differentiable density f . Consider the kernel estimator

  $$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

  with bandwidth sequence $h \propto n^{-1/5}$, so that

  $$n^{2/5}\{\widehat{f}(x) - f(x)\} \Longrightarrow N(b(x),\, v(x))$$

  for some $b(x), v(x)$.

- We now investigate a bootstrap algorithm for approximating the distribution of the root

  $$R_n = n^{2/5}\{\widehat{f}(x) - f(x)\}.$$

  Ideally, we would like to take account of both bias and variance; the usual asymptotic approach ignores the bias.

- Suppose that we resample with replacement from $\{X_1, \ldots, X_n\}$, obtaining the sample $\{X_1^*, \ldots, X_n^*\}$, where X_i^* puts mass $1/n$ at each X_i. Then let

$$\widehat{f}^*(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i^*)$$

$$R_n^* = n^{2/5}(\widehat{f}^*(x) - \widehat{f}(x)).$$

- Following the previous work we might take $\mathcal{L}\{R_n^*|X_1, \ldots$ as an approximation to $\mathcal{L}\{R_n\}$. Unfortunately,

$$\begin{aligned}
& E\{\widehat{f}^*(x)|X_1, \ldots, X_n\} \\
= \; & \frac{1}{nh}\sum_{i=1}^{n} E\left[K\left(\frac{x - X_i^*}{h}\right)\middle| X_1, \ldots, X_n\right] \\
= \; & \frac{1}{h}E\left[K\left(\frac{x - X_i^*}{h}\right)\middle| X_1, \ldots, X_n\right] \\
= \; & \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \widehat{f}(x).
\end{aligned}$$

- In other words, $\widehat{f}^*(x)$ is a conditionally unbiased estimate of $\widehat{f}(x)$. This sounds good but since $\widehat{f}(x)$ is biased it means that $\widehat{f}^*(x)$ does a poor job of estimating that bias.

- However, the variance is correct, i.e.,

$$
\begin{aligned}
& \text{var}\{\widehat{f}^*(x)|X_1,\ldots,X_n\} \\
= \; & \frac{1}{n^2h^2}\sum_{i=1}^n \text{var}\left[K\left(\frac{x-X_i^*}{h}\right)\middle| X_1,\ldots,\, X_n\right] \\
= \; & \frac{1}{nh^2}\text{var}\left[K\left(\frac{x-X_i^*}{h}\right)\middle| X_1,\ldots,\, X_n\right] \\
= \; & \frac{1}{nh^2}\left[\frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)^2 - \left\{\frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\right\}\right. \\
= \; & \frac{1}{nh}\left[\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)^2 - h\widehat{f}^2(x)\right] \\
= \; & \frac{1}{nh}f(x)\int K(u)^2du + O(n^{-1}),
\end{aligned}
$$

which is the asymptotic variance of $\widehat{f}(x)$.

- The central limit theorem is also valid because you have independent random variables.

- There are two obvious ways of correcting the bias problem:

- We can work instead with bandwidths $h = o(n^{-1/5})$ so that the bias is not present in the limiting distribution of $\widehat{f}(x)$.

- The second approach is to make an explicit bias correction to $\widehat{f}$; this requires estimation of f''.

- We consider a more appealing approach that is correct but does not require explicit estimation of the higher derivatives of f. The proposal is to resample from a smoothed version of f, e.g., $\widehat{f}(x)$.

Generate a sample $\{U_1, \ldots, U_n\}$ of $U[0,1]$'s, and then let

$$X_1^* = \widehat{F}^{-1}(U_1), \ldots, X_n^* = \widehat{F}^{-1}(U_n),$$

where $\widehat{F}(x) = \int_{-\infty}^{x} \widehat{f}(z)dz$.

- Now let

$$\widehat{f}^*(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i^*}{n}\right)$$

as before.

- However, now we have

$$\begin{aligned} & E[\widehat{f}^*(x)|X_1, \ldots, X_n] \\ = \; & \frac{1}{h}EK\left(\frac{x - X_i^*}{h}\right) \\ = \; & \frac{1}{h}\int K\left(\frac{x-z}{h}\right)\widehat{f}(z)dz \\ = \; & \int K(u)\widehat{f}(x - uh)du \\ \cong \; & \widehat{f}(x) + \frac{h^2}{2}\mu_2(K)\widehat{f}''(x), \end{aligned}$$

which implies that

$$E[\widehat{f}^*(x) - \widehat{f}(x)|X_1, \ldots, X_n] \cong \frac{h^2}{2}\mu_2(K)f''(x),$$

provided

$$\widehat{f}''(x) \to \widehat{f}''(x)a.s.$$

- The problem here is that for the consistency of $\widehat{f}''(x)$ we would require that $nh^5 \to \infty$, which rules out the optimal bandwidth $h \propto n^{-1/5}$. Therefore, we resample from

$$\widehat{F}_g(x) = \int_{-\infty}^{x} \widehat{f}_g(z)dz,$$

  where $\widehat{f}_g(x)$ is a kernel density estimate constructed from the bandwidth g.

- This gives

$$\begin{aligned} & E[\widehat{f}^*(x)|X_1, \ldots, X_n] \\ = \ & \frac{1}{n}\int K\left(\frac{x-X}{h}\right)\widehat{f}_g(x)dx \\ \cong \ & \widehat{f}_g(x) + \frac{h^2}{2}\mu_2(K)\widehat{f}_g''(x), \end{aligned}$$

which includes $g = 0$ and $g = h$ as special cases.

- Now take

$$\mathcal{L}\{n^{2/5}(\widehat{f}^*(x) - \widehat{f}_g(x))|X_1, \ldots, X_n\}$$

as an "estimate" for

$$\mathcal{L}\{n^{2/5}(\widehat{f}_h(x) - f(x))\}.$$

Provided f'' is continuous at x and $ng^5 \to \infty$, $\widehat{f}_g''(x) \longrightarrow f''(x)$ $a.s.$

- Therefore,

$$E[\widehat{f}^*(x)|X_1, \ldots, X_n] - \widehat{f}_g(x) \cong \frac{h^2}{2}\mu_2(K)f''(x),$$

which is the same as the asymptotic mean of $\widehat{f}_h(x) - f(x)$.

Nonparametric Regression

- Suppose that

$$Y_i = m(X_i) + \varepsilon_i,$$

where either:

Model 1. X_i are fixed in repeated samples but become dense on their support as sample size tends to infinity, while ε_i are i.i.d. mean zero variance σ^2.

Model 2. (Y_i, X_i) are i.i.d. with $m(x) = E(Y|X = x)$ and var$(Y|X = x) = \sigma^2(x)$.

- The main difference is that in model 1 the errors are homoskedastic and indeed i.i.d. In model 1 we can use the following algorithm

Residual resampling

1. Calculate residuals $\hat{\varepsilon}_i = Y_i - \widehat{m}_h(X_i)$, $i = 1, \ldots, n$

2. Recenter $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \overline{\hat{\varepsilon}}$, where $\overline{\hat{\varepsilon}} = n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i \neq 0$.

3. Resample $\{\varepsilon_i^*, \ldots, \varepsilon_n^*\}$ drawn with replacement from $\{\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n\}$.

4. Let $Y_i^* = \widehat{m}_g(X_i) + \varepsilon_i^*$, $i = 1, \ldots, n$. Create bootstrap observations; required that $g/h \to \infty$.

5. Calculate bootstrap nonparametric estimate

$$\widehat{m}_h^*(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i^*$$

6. To approximate the distribution of any functional of $\widehat{m}_h(\cdot) - m(\cdot)$ use the computable conditional distribution of $\widehat{m}^*_h(\cdot) - \widehat{m}_g(\cdot)$.

Theorem 0.3 *Suppose that $h \propto n^{-1/5}$, $g/h \to \infty$, $g \to 0$, K is bounded support and symmetric about zero, m is twice continuously differentiable. Then,*

$$\begin{aligned} \sup_{-\infty<t<\infty} |\,\text{Pr}[(nh)^{1/2}(\widehat{m}^*_h(x) - \widehat{m}_g(x)) &\leq t|\mathcal{X}_n] \\ - \text{Pr}[(nh)^{1/2}(\widehat{m}_h(x) - \widehat{m}(x)) &\leq t]| \\ &\to 0. \end{aligned}$$

- When the errors are not identically distributed we can use the 'wild bootstrap'. In this case you draw ε^*_i from a distribution with mean zero and variance $\tilde{\varepsilon}^2_i$. For example, a normal distribution or a discrete distribution. I

- In the second model, we propose to use i.i.d. resampling with oversmoothing.

i.i.d. resampling

1. Resample (X_i^*, Y_i^*), $i = 1, \ldots, n$

2. Compute $w_{ni}^*(x)$ from X_i^*, $i = 1, \ldots, n$ in the same way as $w_{ni}(x)$ was computed from X_i, $i = 1, \ldots, n$

$$\widehat{m}_h^*(x) = \sum_{i=1}^{n} w_{ni}^*(x) Y_i^*.$$

3. To approximate the distribution of any functional of $\widehat{m}_h(\cdot) - m(\cdot)$ use the computable conditional distribution of $\widehat{m}_g^*(\cdot) - \widehat{m}_g(\cdot)$, where g is another bandwidth.