Nonparametric Methods in Economics and Finance Lecture 1

Oliver Linton

May 10, 2006

- The cumulative c.d.f.

$$F(x) = \text{Pr}(X \leq x).$$

Primitive. Many other quantities can be expressed as functionals of F.

  - The c.d.f. is bounded between zero and one and weakly increasing function.

  - The c.d.f. is of interest for many reasons. One application is to testing for stochastic dominance. In that case also it is of interest the integrated c.d.f.,

$$\begin{aligned} S(x) &= \int_{-\infty}^{x} F(x')dx' \\ T(x) &= \int_{-\infty}^{x} S(x')dx' \end{aligned}$$

- The density function $f(x)$ is defined as the Radon-Nikodym derivative of the c.d.f., i.e., it satisfies

$$F(x) = \int_{-\infty}^{x} f(x')dx'.$$

- When the c.d.f. is differentiable,

$$f(x) = F'(x),$$

but $f(x)$ is defined more generally. The density function is non-negative and integrates to one.

- The hazard function

$$\lambda(x) = \frac{f(x)}{1 - F(x)}$$

is of interest in many areas from mortality to unemployment. The hazard function is non-negative but otherwise unrestricted. We are also interested in the cumulated hazard function

$$\Lambda(x) = \int_{-\infty}^{x} \lambda(x')dx',$$

which is a weakly increasing function. The hazard function and density function are in one to one correspondence so that $f(x) = \lambda(x)\exp(-\Lambda(x))$.

- In many cases we have a vector of variables and are interested in the relationship between one variable and the others, denoted (Y, X). This can be generally described by the conditional c.d.f. and density function

$$\begin{aligned} F_{Y|X}(y|x) &= \Pr(Y \leq y|X = x) \\ F_{Y|X}(y|x) &= \int_{-\infty}^{y} f_{Y|X}(y'|x)dy' \end{aligned}$$

  as well as the conditional hazard function.

- The conditional density can be written as the ratio of joint to marginals

$$f_{Y|X}(y|x) = \frac{f_{Y,X}(y,x)}{f_X(x)},$$

  where the marginal density

$$f_X(x) = \int f_{Y,X}(y,x)dy.$$

- The regression function

$$E(Y|X=x) = \int y f_{Y|X}(y|x)dy$$

is an important quantity derived from the conditional density function. Its definition requires that $E(|Y|) < \infty$.

- The conditional variance

$$\text{var}(Y|X=x) = E(Y^2|X=x) - E^2(Y|X=x)$$

is often of interest in financial applications.

- The conditional quantile function is defined to be

$$Q_{Y|X}(\alpha|x) = \inf\left\{\lambda : F_{Y|X}(\lambda|x) \geq \alpha\right\}$$

for $\alpha \in (0, 1)$. Lower and upper quantiles.

- When $F_{Y|X}(\cdot|x)$ is strictly increasing around α,

$$Q_{Y|X}(\alpha|x) = F^{-1}_{Y|X}(\alpha|x).$$

- This is defined regardless of moments but does require strict monotonicity for a simple definition.

- The regression function and quantile function can be defined as minimizing functionals. Specifically, consider the problem

$$E\left[\{Y - g(X)\}^2\right],$$

where g is any measurable function. It follows immediately that $m(x) = E(Y|X = x)$ satisfies

$$\begin{aligned} & E\left[\{Y - g(X)\}^2 - \{Y - m(X)\}^2\right] \\ = \; & E\left[\{m(X) - g(X)\}^2\right] \\ \geq \; & 0 \end{aligned}$$

for all g. One can also prove this result using calculus of variations techniques for the general optimization problem

$$Q(g) = \int\int \{Y - g(X)\}^2 f(Y, X) dY dX,$$

where f is the joint density.

- Likewise, letting $M(x) = Q_{Y|X}(\alpha|x)$ and

$$Q(g) = E\left[\rho_\alpha(Y - g(X)) - \rho_\alpha(Y - M(X))\right],$$

$$\rho_\alpha(u) = u(\alpha - 1(u < 0).$$

- We have for all g

$$Q(M) \leq Q(g)$$

- In some cases one is also interested in estimating nonparametric regression with generated data. For example, the conditional variance can be estimated from either the relation

$$\begin{aligned} \text{var}(Y \quad &| \quad X = x) \\ &= \quad E(Y^2 \mid X = x) - E^2(Y \mid X = x) \end{aligned}$$

$$\text{var}(Y \mid X = x) = E(\varepsilon^2 \mid X = x),$$

where $\varepsilon = Y - m(X)$. In the former representation one estimates two nonparametric regressions $E(Y^2 \mid X = x), E(Y \mid X = x)$ and then takes a simple nonlinear function of them. In the second case one uses the first nonparametric regression as data.

- In semiparametric applications one often has to estimate 'generated' or 'profiled' regression functions

$$E\left[\tau_Y(Y;\theta)|\tau_X(X;\theta)\right],$$

where τ_Y and τ_X are transformations, either known, unknown upto a finite dimensional parameter θ, or unknown nonparametrically.

C.D.F. and Density Estimation

- Suppose that we have an i.i.d. sample $X_1, \ldots, X_n$ drawn from the distribution F. We estimate the c.d.f. by the empirical c.d.f.

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\left(X_i \leq x\right).$$

  This estimator obeys

  - $0 \leq F_n(x) \leq 1$
  - it is weakly increasing.
  - It is a step function with jumps of height $1/n$ (assuming no ties, which happens with probability zero for continuously distributed data). This is a CADLAG (Continue A Droite and Limite A Gauche) function.

- Note that the density function can be interpreted as the derivative of the c.d.f.

$$\begin{aligned} f(x) &= \lim_{h \to 0} \frac{1}{2h} \Pr\left[x - h \leq X_i \leq x + h\right] \\ &= \lim_{h \to 0} \frac{1}{2h} E\left[1(x - h \leq X_i \leq x + h)\right]. \end{aligned}$$

- However, the density function cannot be estimated by the derivative of $F_n(x)$, since this is a discontinuous function at the sample points and zero elsewhere.

- However, a numerical derivative with small h would be

$$\widehat{f}(x) = \frac{1}{2h}\left[F_n(x + h) - F_n(x - h)\right].$$

- This can be written in the form

$$\widehat{f}(x) = \frac{1}{2nh} \sum_{i=1}^{n} 1\left(|X_i - x| \leq h\right).$$

- We define now a more general class of estimators. Let h be a scalar bandwidth and $K(\cdot)$ a kernel satisfying $\int K(u)du = 1$ and $K_h(\cdot) = h^{-1}K(h^{-1}\cdot)$.

- A kernel K is said to be of order q if

$$\begin{aligned} \int K(u)du &= 1, \\ \int u^j K(u)du &= 0,\ j = 1, \ldots, q-1, \\ \int u^q K(u)du &< \infty. \end{aligned}$$

  The integrals here are over the support of the kernel which in general is some compact interval or the real line. Frequently, attention is restricted to K a probability density function symmetric about zero for which $q = 2$.

- In some cases we are interested in so-called boundary kernels that are functions of two arguments $K(u,t)$, where the parameter t controls the support of the kernel, thus $K(u,t)$ has support $[-1,t]$ and satisfies

$$\begin{aligned}\int_{-1}^{t} K(u,t)du &= 1,\\ \int_{-1}^{t} u^j K(u,t)du &= 0,\ j=1,\ldots,q-1,\\ \int_{-1}^{t} u^q K(u,t)du &< \infty\end{aligned}$$

as for regular kernels.

- Then let

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

- This estimate is non-negative and integrates to one in the special case where the support of X is the entire real line. Symmetric kernels.

- Otherwise not. If there are restrictions on the support of X it may be advisable to use a more complicated kernel that has two or more parameters, see Chen (1999).

- Let $\mathcal{K}_h(x) = \mathcal{K}(x/h) = \int K_h(x')dx'$ a smooth increasing c.d.f. and let

$$\begin{aligned}\widetilde{F}_n(x) &= \mathcal{K}_h * F_n \\ &= \int \mathcal{K}_h(x-y)dF_n(y) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_h(x-X_i)\end{aligned}$$

be a corresponding smoothed estimator of the c.d.f., where $*$ denotes convolution. Then

$$\begin{aligned}\widehat{f}_h(x) &= \widetilde{F}_n'(x) \\ &= K_h * F_n \\ &= \int K_h(x-y)dF_n(y).\end{aligned}$$

- This gives another interpretation of the kernel density and c.d.f. estimator: $\widetilde{F}_n(x), \widehat{f}(x)$ are the c.d.f. and density functions respectively of a sample of the random variables

$$Y_i = X_i + h\varepsilon_i$$

conditional on $X_1, \ldots, X_n$, when ε_i has density K.

- The tails of the kernel density estimator are like the tails of the kernel K.

- So for example, if K were standard normal, then the tails of $\widehat{f}(x)$ as $x \to \pm\infty$ behave likewise.

Regression Estimation

- Suppose that we observe a bivariate dataset $\{Y_i, X_i\}_{i=1}^n$ generated from

$$Y_i = m(X_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where ϵ_i is a random error independent over observations that satisfies

$$E(\epsilon_i \mid X_i = x) = 0.$$

Then $m(\cdot)$ is the regression function of Y on X. It is usual also to assume that

$$\text{var}(Y_i \mid X_i = x) = \sigma^2(x) < \infty.$$

- The smoothness of m determines how well it can be estimated.

Kernel Regression Estimators

- Recall that

$$m(x) = \frac{\int y f(x,y) dy}{\int f(x,y) dy},$$

where $f(x,y)$ is the joint density of (X,Y).

- A natural way to estimate $m(\cdot)$ is first to compute an estimate of $f(x,y)$ and then to integrate it according to this formula. A kernel density estimate $\widehat{f}_h(x,y)$ of $f(x,y)$ is

$$\widehat{f}_h(x,y) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) K_h(y - Y_i).$$

- We have (ignoring the limits of integration):

$$\begin{aligned} \int \widehat{f}_h(x,y)dy &= \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\ ; \\ \int y\widehat{f}_h(x,y)dy &= \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)Y_i. \end{aligned}$$

- Plugging these into numerator and denominator we obtain the Nadaraya–Watson kernel estimate

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x-X_i)Y_i}{\sum_{i=1}^{n} K_h(x-X_i)}.$$

- The estimator is well-defined when

$$\sum_{i=1}^{n} K_h(x-X_i) \neq 0,$$

  which happens with very high probability provided the covariate density is strictly positive at x

  - If $\sum_{i=1}^{n} K_h(x-X_i) = 0$, then define $\widehat{m}_h(x) = 0$ for example].

- The bandwidth h determines the degree of smoothness of $\widehat{m}_h$. This can be immediately seen by considering the limits for h tending to zero or to infinity, respectively. Indeed, at an observation X_i,

$$\widehat{m}_h(X_i) \to Y_i, \text{ as } h \to 0,$$

while at an arbitrary point x,

$$\widehat{m}_h(x) \to \frac{1}{n}\sum_{i=1}^{n} Y_i, \text{ as } h \to \infty.$$

- The Nadaraya-Watson estimator is linear in Y

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

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

depends only on the covariates $X_1, \ldots, X_n$.

- The weights satisfy

$$\sum_{i=1}^{n} w_{ni}(x) = 1.$$

When $K \geq 0$, the weights are probability weights since they also satisfy $w_{ni}(x) \in [0, 1]$.

- If

$$Y_i \mapsto a + bY_i, \widehat{m}(x) \mapsto a + b\widehat{m}(x).$$

- In practice, one estimates at a grid of points $x_1, \ldots, x_m$. Often one takes $m = n$ and $x_i = X_i$ the covariate value. In that case one obtains

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

where $\widehat{m}$ and y are the $n \times 1$ estimator and dependent variable vectors respectively, while W is the $n \times n$ smoother matrix with typical element $W_{ij} = w_{nj}(X_i)$. Let

$$K = \left[K\left(\frac{X_i - X_j}{h} \right) \right]_{i,j}$$

and let

$$W = K./Ki,$$

where ./ is the matrix division operation.

- We can interpret the kernel estimator as the minimizer of the local least squares criterion function

$$Q_n(\theta) = \sum_{i=1}^{n} K_h(x - X_i) \left(Y_i - \theta\right)^2,$$

that is,

$$\widehat{\theta}(x) = \arg\min_{\theta \in R} Q_n(\theta) = \widehat{m}_h(x).$$

- In fact, one can also set-up the global objective function

$$Q_n(\theta(.)) = \int \sum_{i=1}^{n} K_h(x - X_i) \left(Y_i - \theta(x)\right)^2 d\mu(x),$$

where now the 'parameter' is a function $\theta(.)$ and μ is any positive measure absolutely continuous with respect to Lebesgue measure on the support of X.

- It can be shown that the function $\widehat{\theta}(\cdot)$ that minimizes this criterion is exactly $\widehat{m}_h(x)$ for each x. Specifically, let $\theta_\epsilon(x) = \widehat{\theta}(x) + \epsilon g(x)$ for any function g, and compute the first order condition

$$\begin{aligned} & \frac{\partial}{\partial \epsilon} Q_n(\theta_\epsilon(.)) \Big|_{\epsilon=0} \\ = & -\int \sum_{i=1}^{n} K_h(x - X_i) \left(Y_i - \widehat{\theta}(x)\right) g(x) d\mu(x) \\ = & \ 0. \end{aligned}$$

- This must hold for every function g possessing certain regularity. It follows by the Euler-Lagrange Theorem that

$$\sum_{i=1}^{n} K_h(x - X_i) \left(Y_i - \widehat{\theta}(x)\right) = 0$$

for each x.

- The Nadaraya-Watson estimator can be written

$$\begin{aligned}\widehat{m}_h(x) &= \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\frac{Y_i}{\widehat{f}(x)} \\ &= \frac{1}{\widehat{f}(x)}\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)Y_i.\end{aligned}$$

- We should mention some related estimators. First, in some cases the design density might be known in which case one can use

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

  These estimators can be called externally normalized since the denominator factors $\widehat{f}(x)$ and $f(x)$ can be taken outside of the sum.

- An alternative approach is to use internal normalization, hence when f is known

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

  The advantages of this estimator are discussed in Jones, Davies, and Park (1994).

- When f is not known, Mack and Müller (1989) consider

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

k-Nearest Neighbor Estimators

- The k-nearest neighbor (k-NN) estimate is defined as a weighted average of the response variables in a varying neighborhood. This neighborhood is defined through those X-variables which are among the k-nearest neighbors of a point x.

- Let $\mathcal{N}(x) = \{i : X_i \text{ is one of the } k - NN \text{ to } x\}$ be the set of indices of the k-nearest neighbors of x. The k-NN estimate is the average of Y's with index in $\mathcal{N}(x)$,

$$\widehat{m}_k(x) = \frac{1}{k} \sum_{i \in \mathcal{N}(x)} Y_i.$$

- Connections to kernel smoothing can be made by considering this as a kernel smoother with uniform

kernel $K(u) = \frac{1}{2}\mathsf{I}(|u| \leq 1)$ and variable bandwidth $h = m(k)$, the distance between x and its furthest k-NN,

$$\widehat{m}_k(x) = \frac{\sum_{i=1}^n K_R(x - X_i)\, Y_i}{\sum_{i=1}^n K_R(x - X_i)}.$$

Note that for this specific kernel, the denominator is equal to k/nR the k-NN density estimate of $f(x)$. Formula provides sensible estimators for arbitrary kernels.

- One can also consider the similar estimator

$$\frac{\sum_{i=1}^n K_h(F_n(x) - F_n(X_i))Y_i}{\sum_{i=1}^n K_h(F_n(x) - F_n(X_i))},$$

where F_n is the covariate empirical distribution. Thus nearest neighbours can be interpreted as kernel smoothing in 'rank space'.

Local Polynomial Estimators

- The Nadaraya-Watson estimator can be regarded as the solution of the minimization problem

$$\widehat{m}_h(x) = \arg\min_{\theta \in R} \sum_{i=1}^{n} K_h(x - X_i) \{Y_i - \theta\}^2 .$$

- Let

$$P_\theta(t) = \theta_0 + \theta_1 t + \ldots + \theta_p t^p / p!$$

with $\theta = (\theta_0, \theta_1, \ldots, \theta_p)$ denote a polynomial. Let $\widehat{\theta}_0, \ldots, \widehat{\theta}_p$ minimize

$$\sum_{i=1}^{n} K_h(x - X_i) \{Y_i - P_\theta(X_i - x)\}^2$$

with respect to $\theta \in R^{p+1}$. Then, $\widehat{\theta}_0$ serves as an estimator of $m(x)$, while $\widehat{\theta}_j$ estimates the $j'th$ derivative of m.

Higher Dimensions

- All the above methods have generalizations to the case where $d > 1$. For example, in the kernel method we can replace the univariate K and h by multivariate kernel $\mathcal{K}$ and bandwidth matrix H, so that we replace $K\left((x - X_i)/h\right)$ by $\mathcal{K}(H^{-1/2}(x - X_i))$. A special case of this is where

$$\mathcal{K}(H^{-1/2}(x - X_i)) = K(\|x - X_i\|_H)$$

$$\|A\|_H = [\mathrm{tr}(A^\top H^{-1} A)]^{1/2}.$$

- In practice one does not want to choose an entire matrix of bandwidths without some structure. There are several simplifying approaches. First, let

$$H = h\Sigma,$$

where Σ is some fixed symmetric positive definite matrix (in practice estimated from the data) and h is a scalar bandwidth sequence.

- A second approach is based on 'product kernels' where

$$\mathcal{K}(x) = K(x_1) \cdots K(x_d)$$

$$H = \text{diag}\{h_1, \ldots, h_d\},$$

where again h_j reflect the scale of the j'th covariate.

- In the sequel we will adopt the simplest possible scheme so that we can use the same notation $K_h(x)$ for both univariate and multivariate cases. In the multivariate case, $K_h(x) = K(x/h)/h^d$.

Derivatives

- Derivatives can be estimated by differentiating the estimate of m the required number of times. This works provided the estimate of m is itself smooth enough, which can be achieved, for example, by taking K to be smooth like the Gaussian density function.

- The internal kernel method is particularly convenient for estimation of derivatives because in this case

$$\widehat{m}^{(\nu)}(x) = \frac{1}{nh^{d+|\nu|}} \sum_{i=1}^{n} K^{(\nu)} \left(\frac{x - X_i}{h} \right) \frac{Y_i}{\widehat{f}(X_i)}.$$

  The corresponding formula for the Nadaraya-Watson estimator is very complicated.

- The local polynomial method explicitly estimates the derivatives - the parameter estimate $\widehat{\theta}_j(x)$ estimates $m^{(j)}(x)$.

- The problem with this method is just that in high dimensions the number of local parameters to be estimated is very large.

  - For d dimensions and order p polynomial we have a total of $N = \sum_{\ell=0}^{p} N_\ell$ parameters, where $N_\ell = (\ell + d - 1)!/(d-1)!\ell!$

Some Nonlinear Estimators

The above estimators are all linear in the sense that

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

where $\{w_{ni}(x)\}$ only depend on the covariate $X_1, \ldots, X_n$. We now turn to some nonlinear smoothing methods.

Local Likelihood

- The principle underlying the local polynomial estimator can be generalized in a number of ways. Tibshirani (1984) introduced the local likelihood procedure in which an arbitrary parametric regression function $g(x; \theta)$ substitutes the polynomial. Fan, Heckman and Wand (1992) develop theory for a nonparametric estimator in a Generalized Linear Model (GLIM) in which, for example, a probit likelihood function replaces the polynomial.

- Suppose that $f(y|G(x))$ is the density function (or frequency function) of $Y|X$ where f is known and G is an unknown function related to the mean through a known function, i.e., for known F,

$$G(x) = F(m(x)).$$

Then let $\widehat{\theta}_0, \ldots, \widehat{\theta}_p$ minimize

$$\ell_n(\theta) = \sum_{i=1}^{n} K_h(x - X_i) \log f(Y_i|P_\theta(X_i - x)),$$

with respect to $\theta \in R^{p+1}$. Then, $\widehat{\theta}_0$ serves as an estimator of $G(x)$, while $\widehat{\theta}_j$ estimates the $j'th$ derivative of G.

- This includes the standard local polynomial estimator as a special case when f is the normal density function.

- Suppose that Y is binary then

$$f(y|G(x)) = \Phi(G(x))^y[1 - \Phi(G(x))]^{1-y}.$$

- Then

$$m(x) = \Phi(G(x)).$$

- The advantage of this method is that it imposes the restrictions implied by the data. Fan, Heckman, and Wand (1992) See also Gozalo and Linton (1999).

Local GMM

- One can instead have conditional moment restrictions of the form, for some unknown function g

$$E\left[\psi\left(Y, X, g(X)\right)|X = x\right] = 0,$$

where ψ is a vector of moment conditions.

  - For example, suppose that $E(Y|X = x) = m(x)$ and var$(Y|X = x) = m(x)$. Gagliardini and Gourieroux (2005), Gozalo and Linton (1999).

- Then estimation can proceed by minimizing

$$\begin{aligned} & \|G_n(\theta)\| \\ = & \left\| \sum_{i=1}^{n} K_h(x - X_i)\psi\left(Y_i, X_i, P_\theta(X_i - x)\right) \right\|, \end{aligned}$$

where $||A||$ is some vector norm, and letting $\widehat{\theta}_0$ serves as an estimator of $g(x)$.

Quantile Regression

- To estimate conditional quantiles we use a version of local likelihood. The main difference is that $m(x)$ is not interpreted as the conditional mean any more but some other location parameter. Also, the criterion function need not be smooth.

- Let $\widehat{m}(x) = \widehat{\theta}_0$, where $\widehat{\theta}$ is any minimizer of the following criterion function

$$\sum_{i=1}^{n} K_h(x - X_i)\rho_\alpha(Y_i - P_\theta(X_i - x))$$

with $\rho_\alpha(u) = u(\alpha - 1(u < 0)$. In general the solution is easy to compute but is not unique so some additional restriction has to be imposed to obtain a well defined solution. Chaudhuri (19?)

CDF and Density Estimation

Theorem 0.1 *As* $n \to \infty$,

$$\sup_{x \in R} |F_n(x) - F(x)| \longrightarrow 0 \ a.s.$$

- The only 'condition' is that X_i are i.i.d., although note that since F is a distribution function it has at most a countable number of discontinuities, and is right continuous. Note also that the supremum is over a non-compact set - in much subsequent work generalizing this theorem it has been necessary to restrict attention to compact sets. The proof of Theorem 1 exploits some special structure: specifically that for each x, $1(X_i \leq x)$ is Bernoulli with probability $F(x)$.

- Let B be the Brownian Bridge process. This is a Gaussian process with covariance function

$$\text{cov}(B(s), B(t)) = \min\{s, t\} - st$$

for every s, t. We next establish the weak convergence of the empirical c.d.f. i.e., a Functional Central Limit Theorem (FCLT).

Theorem 0.2 *As* $n \to \infty$,

$$n^{1/2} [F_n(\cdot) - F(\cdot)] \Longrightarrow B(F(\cdot)),$$

where B *is the Brownian Bridge process.*

- The limiting process is a time changed Gaussian process. The result can be established by first establishing the result for uniform $[0, 1]$ random variables and then employing the result that $X = F^{-1}(U)$, where U is a uniform random variable.

Theorem 0.3 *(Nadaraya, (1965)) Suppose that K is of bounded variation and integrates to one, that f is uniformly continuous on R, and that $h \to 0$ and $nh^2 \to \infty$. Then*

$$\sup_{x \in R} \left| \widehat{f}(x) - f(x) \right| \longrightarrow 0 \ a.s.$$

- This theorem places very weak assumptions on the kernel and density but somewhat stronger conditions on the bandwidth sequence. Note that the convergence is uniform over the entire real line. It is possible to establish uniform consistency of the kernel density estimator under weaker conditions on the bandwidth sequence like $nh/\log n \to \infty$ at the expense of stronger conditions on K.

Theorem 0.4 *(Silverman (1978)) Suppose that K is uniformly continuous with modulus of continuity w and of bounded variation, that $\int |K(x)|dx < \infty$ and $K(x) \to 0$ as $x \to \infty$, that $\int K(x)dx = 1$, and that $\int |x \log |x||^{1/2}|dK(x)| < \infty$. Suppose that f is uniformly continuous. Then, provided $h \to 0$ and $nh/\log n \to \infty$*

$$\sup_{x \in R} \left| \widehat{f}(x) - f(x) \right| \longrightarrow 0 \ a.s.$$

Suppose additionally that $\int_0^1 [\log(1/u)]^{1/2} \, d\gamma(u) < \infty$, where $\gamma(u) = \{w(u)\}^{1/2}$ and that

$nh(\log n)^{-2}\{\log(1/h)\} \to \infty$ and $\sum_{n=1}^{\infty} h_n^{\lambda} < \infty$ for some λ, then

$$\begin{aligned} & \sup_{x \in R} \left| \widehat{f}(x) - E[\widehat{f}(x)] \right| \\ = \ & O\left(\left(\frac{\log(1/h)}{nh} \right)^{1/2} \right) \ a.s. \end{aligned}$$

- These results use arguments that are special to the univariate case.

- The assumption that f is uniformly continuous is innocuous for densities of unbounded support but rather restrictive for those living on a bounded interval. For example, it rules out the uniform density. The assumption is needed for handling the bias term $E[\widehat{f}(x)] - f(x)$ and the results hold true for $\sup_{x \in R} |\widehat{f}(x) - E[\widehat{f}(x)]|$ without this assumption.

Theorem 0.5 *(Giné and Guillou (2002)) Suppose that the kernel K is a bounded function of bounded variation. Suppose that $h \to 0$ monotonically, such that $nh^d/|\log h| \to \infty$ and $|\log h|/\log\log n \to \infty$. Suppose further that the density f is bounded. Then*

$$\begin{aligned} & \sup_{x \in R} \left| \widehat{f}(x) - E[\widehat{f}(x)] \right| \\ = \; & O\left(\left(\frac{\log(1/h)}{nh^d} \right)^{1/2} \right) \; a.s. \end{aligned}$$

- This result is quite remarkable in terms of the weakness of the conditions. To establish consistency of $\widehat{f}(x)$ though we also need to analyze the term

$$\sup_{x \in R} |E[\widehat{f}(x)] - f(x)|.$$

  This requires additional conditions. For example, one might assume that f is uniformly continuous like Silverman (1978).

- To establish a rate one needs stronger conditions specifically smoothness. We note that uniform continuity is an appropriate condition for densities with unbounded support but does rule out many density with compact support for example the uniform density. For those cases different conditions are appropriate. The bias term is handled by making a change of variables. We have

$$E[\widehat{f}(x)] = \int K_h(x - X) f(X) dX$$

and if we transform

$$X \mapsto u = (x - X)/h$$

the integrand becomes $K(u)f(x - uh)du$. If the support of X is R then the range of integration of u is not affected.

- Then we have

$$\begin{aligned} \int K(u)f(x - uh)du &= f(x) \int K(u)du \\ &\quad - f'(x) \int K(u)udu \\ &\quad + f''(x)\frac{1}{2} \int K(u)u^2 du. \end{aligned}$$

- However, if the support of X is some interval $[\underline{x}, \overline{x}]$, then the range of integration of u becomes

$$[(x - \underline{x})/h, (\overline{x} - x)/h].$$

  - When x is an interior point this interval tends towards $(-\infty, \infty)$,
  - When $x = \underline{x}$, then the interval tends towards $(0, \infty)$.
  - When $x = \underline{x} + th$ for some t, then the interval tends to $[t, \infty)$.

- The limiting distribution of the density estimator at a point.

Theorem 0.6 *Suppose that K is bounded and satisfies $\int K(u)udu = 0$, $\int K(u)u^2du < \infty$ and $\int |K(u)|du < \infty$. Suppose also that f is twice continuously differentiable at the interior point x. Then*

$$(nh)^{1/2}\left[\widehat{f}(x) - f(x)\right] \Longrightarrow N(b(x), v(x)),$$

where

$$\begin{aligned} v(x) &= \|K\|_2^2 f(x) \\ b(x) &= \left(\lim_{n\to\infty}(nh^5)^{1/2}\right)\frac{\mu_2(K)}{2}f''(x). \end{aligned}$$

Regression Function

- Stone (1977) gave the following result for linear estimators

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

where $\{w_{ni}(x)\}$ only depend on the covariate $X_1, \ldots, X_r$

Theorem 0.7 *(Stone (1977). Let $\{w_{ni}(x)\}$ be a sequence of weights and let $X, X_1, \ldots, X_n$ be i.i.d. Suppose that the following conditions hold*
(1) There is a $C \geq 1$ such that for every nonnegative Borel measurable function f

$$E \sum_{i=1}^{n} w_{ni}(X) f(X_i) \leq C E f(X);$$

(2) There is a $D \geq 1$ such that

$$\Pr \left[\sum_{i=1}^{n} |w_{ni}(X)| \leq D \right] = 1;$$

(3) For all $a > 0$

$$\sum_{i=1}^{n} |w_{ni}(X)| \, 1 \left(||X_i - X|| > a \right) P \longrightarrow 0;$$

(4)

$$\sum_{i=1}^{n} |w_{ni}(X)| \, P \longrightarrow 1;$$

(5)

$$\max_{1 \leq i \leq n} |w_{ni}(X)| P \longrightarrow 0.$$

Then $\{w_{ni}(x)\}$ *are consistent in the sense that whenever* $E[|Y|^r] < \infty$,

$$E\left[|\widehat{m}(X) - m(X)|^r\right] \to 0.$$

- These are quite weak conditions. Note that for many regression estimators the sequence of weights $\{w_{ni}(x)\}$ are probability weights, i.e., they lie between zero and one and sum to one. In this case conditions (1), (2), and (4) are quite natural. A consequence of condition (1) is that

$$E[|\widehat{m}(X)|] < \infty \text{ whenever } E[|Y|] < \infty.$$

- Condition (3) is trivially satisfied for kernel estimators with kernels of bounded support. Condition (5) is also satisfied for many estimators - for nearest neighbor estimators for example it is trivial. Stone (1977) shows that these conditions can be satisfied for a range of estimators.

- The standard local linear estimator does not satisfy these conditions however. He shows how to modify local linear estimators to make them probability weights and thereby to satisfy the conditions. Kohler (2000) suggests an alternative way

of doing this by restricting the optimization to a bounded parameter space (that is allowed to expand slowly with sample size).

- Stone also shows how to apply these results to nonlinear estimators like conditional quantile estimators.

- Devroye and Wagner (1980) showed that the kernel estimator with non-negative bounded and bounded away from zero at the origin and compactly supported kernel K satisfies this provided only $h \longrightarrow 0$ and $nh^d \longrightarrow \infty$.

Theorem 0.8 *(Local Linear). Suppose that:*
(i) The marginal density f of the covariates is continuous at the interior point x and $f(x) > 0$;
(ii) The regression function $m(x) = E(Y|X = x)$ is twice differentiable and $m''(x)$ is continuous at x; the variance function $\sigma^2(x) = \mathsf{var}(Y|X = x)$ is continuous and positive at x;
(iii) The kernel K is continuous on its compact support;
(iv) $E(|Y|^{2+\delta}) < \infty$ for some $\delta > 0$;
(v) $h \to 0$ and $nh \to \infty$ such that $\lim_{n\to\infty} nh^5 = c_h < \infty$.
Then

$$(nh)^{1/2}\left[\widehat{m}_{LL}(x) - m(x)\right] \Longrightarrow N(b(x), v(x)),$$

where

$$\begin{aligned} v(x) &= \|K\|_2^2 \frac{\sigma^2(x)}{f(x)} \\ b(x) &= c_h \frac{\mu_2(K)}{2} m''(x). \end{aligned}$$

Corollary 0.9 *(Nadaraya-Watson) Suppose that (i)-(v) above hold and that also $f''(\cdot)$ exists and is continuous at x, and that K is a second order kernel. Then*

$$(nh)^{1/2}\left[\widehat{m}_{NW}(x) - m(x)\right] \Longrightarrow N(b_{NW}(x), v(x)),$$

where

$$b(x) = c_h \frac{\mu_2(K)}{2} \frac{(m \cdot f)'' - m \cdot f''}{f}(x).$$

- For both estimators the mean squared error for any interior point x is $O(h^4) + O(1/nh)$, and the best rate is given by taking $h \propto n^{-1/5}$ in which case the mean squared error is of order $n^{-4/5}$. This is larger than in the parametric case where the mean squared error declines like n^{-1}.

- The bias of the NW estimators depends on f and on its derivatives. The local linear estimator by contrast has bias uniformly of order h^2, and is "design adaptive", i.e., its bias only depends on $m''(x)$. Finally, the regularity conditions for the local linear estimator are weaker, since no derivatives of f are needed, just continuity.

- The asymptotic distribution for both estimators has a bias and is 'nuisance parameter dependent'. To obtain correct confidence intervals we would have to estimate $m''(x), f(x)$ and $\sigma^2(x)$ in the case of the local linear estimator, and $m'(x), m''(x), f(x)$

and $\sigma^2(x)$ in the case of the Nadaraya-Watson estimator, which in either case is an even more difficult than the problem we started out with. Therefore, in practice it is usual only to estimate the variance and to argue that the bias is of smaller order [which would be the case if $h = o(n^{-1/5})$]. This is called 'undersmoothing'.

- Boundary bias. When the evaluation point x is at the boundary or close to the boundary, the Nadaraya-Watson Estimator suffers badly from boundary bias. Specifically, in this case the bias is $O(h)$ for any point x_n that lies within h of the boundary. This is because the change of variables argument we use to obtain the bias no longer applies. The local linear estimator does not suffer so badly in the boundary region, and its bias is the same magnitude as at interior points, namely h^2.

- Suppose that $f(x) = 0$ or $f(x) = \infty$. Then we may still obtain consistency and asymptotic normality but at slower (faster) rates of convergence,

see Hengartner and Linton (1999). Likewise with $\sigma^2(x) = \infty$ or $\sigma^2(x) = 0$.

- Suppose that $\sigma^2(.)$ and or $f(.)$ is (boundedly) discontinuous at the point x but that m is continuous at x. Then the estimators are still consistent but the asymptotic variance changes to

$$\frac{\sigma^2(x_-)f(x_-)\int_{-\infty}^{0} K(u)^2du + \sigma^2(x_+)f(x_+)\int_{0}^{\infty} K(u)^2d}{\left(f(x_-)\int_{-\infty}^{0} K(u)du + f(x_+)\int_{0}^{\infty} K(u)du\right)^2}$$

  Under symmetry of the kernel this simplifies to $[\sigma^2(x_-)f(x_-)+\sigma^2(x_+)f(x_+)]/\left(f(x_-)+f(x_+)\right)^2/2$. Bias? If $m(x)$ is discontinuous at x, then the estimator converges to $(m(x_-)+m(x_+))/2$ under symmetry of the kernel.

- Can allow the marginal density and error variance (or distribution) to vary with i and n, denoted

$f_{ni}(.)$ and $\sigma^2_{ni}(.)$ provided these functions are uniformly smooth etc.. In this case the limiting variance is

$$\|K\|_2^2 \frac{\frac{1}{n}\sum_{i=1}^n f_{ni}(x)\sigma^2_{ni}(x)}{\left[\frac{1}{n}\sum_{i=1}^n f_{ni}(x)\right]^2}.$$

An extreme example is when $X_i = i/n$, i.e., purely deterministic in which case $f_{ni}(x) \to 1$ as $n \to \infty$. The theory is as above where f can be interpreted as a limiting or average density. This would be called a fixed design. A number of estimators have different properties depending on whether the design is fixed or random but the local linear and local constant have essentially identical properties in these two cases.

- Suppose that $E[|Y|^2] = \infty$ but $E[|Y|^{1+\alpha}] < \infty$ for some $\alpha \in (0, 1)$. Then we still have consistency but with slower rates and possibly non-normal limiting distributions (this is just as in the parametric case). Stute (1986) established the

almost sure convergence of the Nadaraya-Watson estimator under only weak moment conditions, namely $E(|Y|^{1+\delta}) < \infty$.

- The compact support condition on the kernel can be weakened to just K bounded and $\int |K(u)|u^2 du < \infty$.

- The bias and variance of the more general k-NN estimator is given in a theorem by Mack (1981). Stute (1984) proves asymptotic normality. In contrast to kernel smoothing, the variance of the k-NN regression smoother does not depend on f, the density of X. This makes sense since the k-NN estimator always averages over exactly k observations independently of the distribution of the X-variables. The bias constant $B_{nn}(x)$ is also different from the one for kernel estimators given in Theorem 2. An approximate identity between

k-NN and kernel smoothers can be obtained by setting

$$k = 2nhf(x),$$

or equivalently $h = k/2nf(x)$.

Some Heuristics

- We give some heuristics for the local constant approach. Write

$$\widehat{m}(x) = \frac{\widehat{r}(x)}{\widehat{f}(x)} \text{ and } m(x) = \frac{r(x)}{f(x)}$$

$$\begin{aligned} \widehat{r}(x) &= \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)Y_i \\ r(x) &= m(x)f(x) \end{aligned}$$

- First approach to this would be to apply the delta method. We know the properties of $\widehat{r}(x) - r(x)$ and $\widehat{f}(x) - f(x)$ and just Taylor expand. This turns out to be messy and easy to mess up.

- Instead

$$\begin{aligned}
& \widehat{m}(x) - m(x) \\
= & \frac{\widehat{r}(x) - m(x)\widehat{f}(x)}{\widehat{f}(x)} \\
= & \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\varepsilon_i}{\widehat{f}(x)} \\
& + \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\left[m(X_i) - m(x)\right]}{\widehat{f}(x)}.
\end{aligned}$$

- This decomposition is useful because we can replace $\widehat{f}(x)$ by $f(x)$ by the following argument. We have shown that for each interior point x,

$$\widehat{f}(x) = f(x) + O_p(h^2) + O_p(n^{-1/2}h^{-d/2}).$$

Therefore,

$$\begin{aligned}
& \widehat{m}(x) - m(x) \\
= & \frac{1}{f(x)}\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\varepsilon_i \\
& \times \left[1 + O_p(h^2) + O_p(n^{-1/2}h^{-d/2})\right]
\end{aligned}$$

$$+\frac{1}{f(x)}\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\left[m(X_i)-m(x)\right]$$
$$\times\left[1+O_p(h^2)+O_p(n^{-1/2}h^{-d/2})\right]$$

assuming that $f(x) > 0$.

- Writing

$$Z_{ni}(x) = K_h(x-X_i)\left(m(X_i)-m(x)\right)$$

we have

$$\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)\left(m(X_i)-m(x)\right)$$
$$= E[Z_{ni}(x)] + \frac{1}{n}\sum_{i=1}^{n} Z_{ni}(x) - E[Z_{ni}(x)].$$

- The properties of $E[Z_{ni}(x)]$ follow from a Taylor series expansion. We have

$$E[Z_{ni}(x)]$$
$$= \int K_h(x-X)\left(m(X)-m(x)\right) f(X)dX$$
$$= \int K(u)\left(m(x+uh)-m(x)\right) f(x+uh)du.$$

- The second term is a sum of mean zero independent and identically distributed random variables (for given n) and has variance

$$\begin{aligned} & \text{var}[Z_{ni}(x)] \leq E[Z_{ni}^2(x)] \\ = \; & h^{-2d} E\left[K\left(\frac{x - X_i}{h}\right)^2 (m(X_i) - m(x))^2\right] \\ = \; & h^{-d} \int K(u)^2 (m(x+uh) - m(x))^2 f(x+uh) du \\ = \; & O(h^{2-d}). \end{aligned}$$

- It follows that

$$\begin{aligned} & \widehat{m}(x) - m(x) \\ = \; & \frac{1}{f(x)} \left[\frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)\varepsilon_i + E[Z_{ni}(x)]\right] \\ & \times [1 + o_p(1)]. \end{aligned}$$

- Then we can apply the Lindeberg CLT to the random variable $\sum_{i=1}^{n} K_h(x - X_i)\varepsilon_i / n$.

- We next consider the properties of the internal estimator

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

Let also

$$\widetilde{m}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h (x - X_i) \frac{Y_i}{f(X_i)}.$$

Theorem 0.10 *Suppose that $X_1, Y_1 \ldots, X_n, Y_n$ are i.i.d. like X, Y and:*
(i) The marginal density f of the covariates is twice continuously differentiable at the interior point x and $f(x) > 0$;
(ii) The regression function $m(x) = E(Y|X = x)$ is twice continuously differentiable at x; the variance function $\sigma^2(x) = \mathsf{var}(Y|X = x)$ is continuous and positive at x;
(iii) The second order kernel K is Lipschitz continuous on its compact support;
(iv) $E(|Y|^{2+\delta}) < \infty$ for some $\delta > 0$;
(v) $h \to 0$ and $nh \to \infty$ such that $\lim_{n\to\infty} nh^5 = c_h < \infty$.
Then

$$(nh)^{1/2} \left[\widehat{m}(x) - m(x)\right] \Longrightarrow N(b(x), v(x)),$$

where, letting $T_K(s) = K(s) - (K * K)(s)$, *where* $K * K$ *denotes convolution:*

$$\begin{aligned} v(x) &= \frac{\sigma^2(x)}{f(x)}||K||_2^2 + \frac{m^2(x)}{f(x)}||T_K||_2^2 \\ b(x) &= c_h \frac{\mu_2(K)}{2}\left[m''(x) - \frac{m(x)f''(x)}{f(x)}\right]. \end{aligned}$$

- This should be compared with the limiting distribution of $\widetilde{m}(x)$ given in Mack and Müller (1989) where

$$\begin{aligned} v(x) &= \frac{\sigma^2(x) + m^2(x)}{f(x)} ||K||_2^2 \\ b(x) &= c_h \frac{\mu_2(K)}{2} m''(x). \end{aligned}$$

- We may further compare the bias and variance of $\widehat{m}(x)$ with some other estimators. The Nadaraya-Watson estimator has

$$\begin{aligned} v(x) &= \frac{\sigma^2(x)}{f(x)} ||K||_2^2 \\ b(x) &= c_h \frac{\mu_2(K)}{2} \frac{[(mf)'' - mf'']}{f}(x). \end{aligned}$$

- The external estimator with known density

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

has

$$\begin{aligned} v(x) &= \frac{\sigma^2(x) + m^2(x)}{f(x)} ||K||_2^2 \\ b(x) &= c_h \frac{\mu_2(K)}{2} \frac{(mf)''}{f}(x). \end{aligned}$$

The local linear (Fan (1992)) has

$$\begin{aligned} v(x) &= \frac{\sigma^2(x)}{f(x)} ||K||_2^2 \\ b(x) &= c_h \frac{\mu_2(K)}{2} m''(x). \end{aligned}$$

- Note that

$$||T_K||_2^2 \leq ||K||_2^2$$

and $||T_K||_2^2$ is actually very small for most kernels. For the Gaussian kernel, $||K||_2^2 = 0.282$ and $||T_K||_2^2 = 0.07$. It follows that the variance of $\widehat{m}(x)$ is less than the variance of $\widetilde{m}(x)$.

Bias reduction

- When the regression function has more derivatives, it is possible to reduce the magnitude of the bias by using higher order polynomials or in the case of the Nadaraya-Watson estimator, higher order kernels, that is, kernels for which

$$\int u^j K(u)du = 0, j = 1, \ldots, q-1$$

and $\int u^q K(u)du < \infty$ for $q > 2$.

- In either case we achieve bias of order h^q at interior points under corresponding smoothness conditions, in fact the bias of the Nadaraya-Watson estimator is

$$b(x) = c_h \frac{\mu_q(K)}{q!} \left\{ (m \cdot f)^{(q)} - m \cdot f^{(q)} \right\}(x),$$

and the variance is like above (although because the kernel is more wiggly, of necessity $||K||_2^2$ is larger for higher order kernels).

- For the local polynomial case, the form of the bias depends on whether the polynomial order is even or odd. If odd, then the bias is

$$b(x) = c_h C_q(K) m^{(q)}(x),$$

where $C_q(K)$ is a constant just depending on the kernel.

- In this case the optimal bandwidth is $h \propto n^{-1/(2q+1)}$ and the mean squared error is of order $n^{-2q/(2q+1)}$. Marron and Wand have shown that the small sample performance of higher order kernels is poor. There is no unbiased estimator.

Curse of dimensionality

- When there are many $X's$, the variance of the Nadaraya-Watson estimator is of order $1/nh^d$, where d is the dimensions and the bandwidth matrix $H = hI_d$. The reason for this high variance is because in high dimensions, observations are more spread out. In this case, the optimal rate of convergence is obtained by setting h^{2q} the same order as $1/nh$, i.e., $h \propto n^{-1/(2q+d)}$, and the resulting mean squared error is of order $n^{-2q/(2q+d)}$.

- It is also necessary to rule out functional dependence between the covariates.

Uniform Consistency

- In this section we discuss the uniform consistency for kernel regression estimators. We shall concentrate on the L_∞ distance, which is usually the most difficult to work with. These results are especially important for the analysis of semiparametric estimators which involve averages of nonparametric estimates evaluated at a large number of points. They are also relevant for many other estimation and testing problems.

Theorem 0.11 *(Local Linear, Masry (1996)). Suppose that:*
(i) The marginal density f of the covariates is continuous on the compact set $\mathcal{X}$ and $\inf_{x\in\mathcal{X}} f(x) > 0$;
(ii) The regression function $m''(.)$ is Lipschitz continuous on $\mathcal{X}$;
(iii) The kernel K is Lipschitz continuous on its compact support;
(iv) For some $\delta > 0$, $E(|Y|^{2+\delta}) < \infty$;
(v) $h \to 0$ and $nh \to \infty$ such that $n^{\delta/(2+\delta)}h/\log n\{\log n(\log\log n)^{1+\epsilon}\}^{2/(2+\delta)} \to \infty$ *for some $\epsilon > 0$. Then*

$$\begin{aligned} & \sup_{x\in\mathcal{X}} |\widehat{m}(x) - m(x)| \\ = \;& O\left(\left[\frac{\log n}{nh}\right]^{1/2}\right) + O(h^2) \text{ a.s.} \end{aligned}$$

- This is a simplified version of Masry (1996). By taking $h = O((\log n/n)^{1/5})$ we obtain the best possible rate of $(\log n/n)^{2/5}$. We need $n^{\delta/(2+\delta)}n^{-1/5} \longrightarrow \infty$, which requires that $\delta > 1/2$.

- Einhmahl and Mason (2000) establish more precise results for the stochastic part of $\widehat{m} - m$, obtaining the precise rate of almost sure convergence.

- Unlike in the density estimation case we must restrict our attention to compact sets. The reason is due to the presence of the marginal covariate density in the denominator. For unbounded support, $f(x) \to 0$ as $x \longrightarrow \infty$ and so $\sup_x 1/f(x) = \infty$.

- It is possible to extend these results to allow C_n to expand with sample size at some rate although this slows down the rate of convergence depending on the tails of the marginal distribution of the covariate. Andrews (199?). Further results for weighted norms. Law of the iterated logarithm. Results for $||\widehat{m} - m||_p$.

Functional Central Limit Theorem

- The first type of result is a local functional central limit theorem for the kernel estimator. Fix an interior point x and let

$$\begin{aligned} \nu_n(t) &= (nh)^{1/2} \left[\widehat{m}(x+th) - m(x+th)\right], \\ t &\in [-T, T] \end{aligned}$$

for some fixed T. Then we already have pointwise convergence in distribution of $\nu_n(t)$. Under additional conditions it can be shown that

$$\nu_n(\cdot) \Longrightarrow Z(\cdot),$$

where $Z(\cdot)$ is some Gaussian process. It follows that $\sup_{t\in[-T,T]} |\nu_n(t)|$ has the distribution of $\sup_{t\in[-T,T]} |Z(t)|$.

- A similar result can be established for the local in bandwidth process

$$\begin{aligned} \nu_n(t) &= \left(n^{1-\alpha}t\right)^{1/2} \left[\widehat{m}_t(x) - m(x)\right], \\ t &\in [-T, T], \end{aligned}$$

where $h = tn^{-\alpha}$ for some given α. We obtain likewise a functional CLT. These results have a number of applications from establishing the efficiency of plug-in estimators to testing theory.

- Let

$$\begin{aligned} T_n &= \sup_{x \in C} |T_n(x)| \\ T_n(x) &= \frac{\widehat{m}(x) - m(x)}{(\mathsf{var}[\widehat{m}(x)])^{1/2}}, \end{aligned}$$

where C is some compact set contained in the support of X, while $\mathsf{var}[\widehat{m}(x)]$ is the asymptotic variance or conditional variance.

- We know that (with undersmoothing) $T_n(x)$ is asymptotically standard normal for each x, but that $T_n = O_p((\log n)^{1/2})$.

- It can be shown that there exists increasing sequences a_n, b_n such that

$$\Pr\left[a_n(T_n - b_n) \leq t\right] \to e^{-2e^{-t}},$$

i.e., T_n is asymptotically Gumbel. This result was first proved in Bickel and Rosenblatt (1973) for the one dimensional density case and Rosenblatt (1976) for the d-dimensional density case, and Johnston (1982) for univariate local constant nonparametric regression.

- One application of this result is to the limiting distribution of estimates of nonparametric bounds for covariate effects in the presence of selection, see Manski (1994).

- The main use of this result is in setting uniform confidence intervals.

Theorem 0.12 *Suppose that*

1. *The functions* m, f, σ^2 *are all twice continuously differentiable on* C.

2. *The kernel is symmetric about zero and differentiable with bounded support* $[-A, A]$ *for some* A, *where* $K(\pm A) = 0$.

3. *For all* k, $E(|Y|^k | X = x) \leq C_k < \infty$.

4. $h = O(n^{-\delta})$ *with* $\frac{1}{5} < \delta < \frac{1}{3}$.

Then,

$$\Pr\left[a_n(T_n - b_n) \leq t\right] \to e^{-2e^{-t}}$$

with

$$\begin{aligned} b_n &= (-2\log h)^{1/2} + \frac{\log \frac{C}{\pi^2}}{(-2\log h)^{1/2}} \\ a_n &= (-2\log h)^{1/2}, \end{aligned}$$

where $C = \|K'\|^2 \Big/ \|K\|^2$.

Asymptotics For Nonlinear Kernel Smoothers

- E.g., Median kernel smoother,

$$\widehat{M}(x) = \arg\min_{\theta} \sum_{i=1}^{n} K_h\left(x - X_i\right) |Y_i - \theta| .$$

  Minimizer is not unique but any rule can to select $\widehat{M}(x)$ ok.

- More generally we can have that $\{w_{ni}(x)\}$ are smoother weights that satisfy $\sum_{i=1}^{n} w_{ni}(x) = 1$. Just like the usual median, the local median is a nonlinear function of the $Y's$.

- Note that $\widehat{M}(x)$ solves the first order condition

$$\widehat{M}(x) = \arg \text{zero}_{\theta}\ G_n(\theta) = 0,$$

$$G_n(\theta) = \sum_{i=1}^{n} w_{ni}(x) \left\{1(Y_i - \theta > 0) - 1(Y_i - \theta \leq 0)\right\} .$$

- What is important for that is the variance of the score function at the true parameter value and the derivative with respect to parameter values of the first order condition at the true value. The conditional variance of this score function [at the true $\theta = M(x)$] is precisely $\sum_{i=1}^{n} w_{ni}^2(x)$. Let

$$\begin{aligned} \overline{G}(\theta) &= E\left[G_n(\theta)\,|X_1, \ldots, X_n\right] \\ &= \sum_{i=1}^{n} w_{ni}(x)\left\{1 - 2F_i(\theta)\right\}, \end{aligned}$$

where $F_i(\theta) = \Pr(Y_i \leq \theta | X_i)$. Note that

$$\begin{aligned} \overline{G}'(M(x)) &= -2\sum_{i=1}^{n} w_{ni}(x) F_i'(M(x)) \\ &\simeq -2 f_x(M(x)), \end{aligned}$$

where $f_x = F_x'$ and $F_x = \Pr(Y \leq \theta | X = x)$, by a Taylor expansion and using the fact that $\sum_{i=1}^{n} w_{ni}(x) = 1$.

- Therefore,

$$\begin{aligned} & \widehat{M}(x) - M(x) \\ = & \frac{\sum_{i=1}^{n} w_{ni}(x) \{1(Y_i - \theta > 0) - 1(Y_i - \theta \leq 0)\}}{-2 f_x(M(x))} \\ & + O_p(h^2) \end{aligned}$$

by standard arguments. Therefore, we have [with undersmoothing]

$$\frac{\widehat{M}(x) - E\left\{\widehat{M}(x) \,|X_1, \ldots, X_n\right\}}{\left(\frac{\sum_{i=1}^{n} w_{ni}^2(x)}{4 f_x(M(x))^2}\right)^{1/2}} \Longrightarrow N(0,1).$$

See Jones and Marron (1990) for further discussion.

- This result can be extended to the class of smoothers that set

$$G_n(\theta) = \sum_{i=1}^{n} w_{ni}(x)\rho(Y_i, \theta)$$

equal to zero, where ρ is a function that satisfies

$$E[\rho(Y_i, \theta)|X_i = x] = 0$$

if and only if $\theta = \theta_0$. One can construct confidence intervals using this structure along the lines already discussed.