Nonparametric and Semiparametric Estimation

Whitney K. Newey

Fall 2007

Introduction

Function form misspecification error is important in elementary econometrics. Flexible functional forms; e.g. translog

$$y = \beta_1 + \beta_2 \ln(x) + \beta_3 [\ln(x)]^2$$

Fine for simple nonlinearity, e.g. diminishing returns. Economic theory does not restrict form. Nonparametric methods allow for complete flexibility. Good for graphs. Good for complete flexibility with a few dimensions.

An Empirical Example

An example illustrates. Deaton (1989); effect of rice prices on the distributions of incomes in Thailand.

p price of rice; q amount purchased; y amount sold.

Change in benefits from dp is $dB = (q - y)dp = p(q - y)d\ln(p)$.

Elasticity form:

$$\frac{dB}{x}/d\ln(p) = (w - py/x),$$

w budget share of rice purchases; x total expenditure. Benefit/expenditure measure is the negative of right-hand side.

Empirical Distribution Function

Simple nonparametric estimation problem. The CDF of Z is $F_Z(z) = \Pr(Z \leq z)$. Let $Z_1, ..., Z_n$ be i.i.d. data, $1(A)$ indicator of A, so $F_Z(z) = E[1(Z_i \leq z)]$.

$$\hat{F}_Z(z) = \frac{\#\{i | Z_i \leq z\}}{n} = \frac{1}{n}\sum_{i=1}^{n} 1(Z_i \leq z).$$

Empirical CDF.

Probability weight $1/n$ on each observation.

Consistent and asymptotically normal.

Nonparametrically efficient.

No good for density estimation.

Kernel Density Estimator

Add a little continuous noise to smooth out empirical CDF.

$\bar{Z}_n$ have empirical CDF.

U a continuous random variable with pdf $K(u)$, indep of $\bar{Z}_n$

h a positive scalar.

Define

$$\tilde{Z} = \bar{Z}_n + hU$$

Empirical CDF plus noise. Kernel density estimator is density of $\tilde{Z}$.

Derivation: Let $F_U(u) = \int_{-\infty}^{u} K(t)dt$ be CDF of U.

By iterated expectations

$$E[1(\tilde{Z} \leq z)] = E[E[1(\tilde{Z} \leq z)|\bar{Z}_n]],$$

so by $1(\tilde{Z} \leq z) = 1(U \leq (z - \bar{Z}_n)/h)$,

$$\begin{aligned} F_{\tilde{Z}}(z) &= \Pr(\tilde{Z} \leq z) = E[1(\tilde{Z} \leq z)] \\ &= E[[1(U \leq \frac{z - \bar{Z}_n}{h})|\bar{Z}_n]] \\ &= E[F_U(\frac{z - \bar{Z}_n}{h})] = \sum_{i=1}^{n} F_U(\frac{z - Z_i}{h})/n. \end{aligned}$$

Differentiating gives pdf

$$\begin{aligned} \hat{f}_h(z) &= dF_{\tilde{Z}}(z)/dz = \sum_{i=1}^{n} K_h(z - Z_i)/n; \\ K_h(u) &= h^{-1}K(u/h). \end{aligned}$$

This is a *kernel density estimator*. The function $K(u)$ is the *kernel* and the scalar h is the *bandwidth.*

$$\begin{aligned} \hat{f}_h(z) &= dF_{\tilde{Z}}(z)/dz = \sum_{i=1}^{n} K_h(z - Z_i)/n; \\ K_h(u) &= h^{-1}K(u/h). \end{aligned}$$

Bandwidth h controls the amount of smoothing. As h increases, density smoother, but more "noise" from U, i.e. more bias. As $h \longrightarrow 0$ get rough density, spikes at data points, but bias shrinks. Choosing h important in practice; see below. $\hat{f}_h(z)$ will be consistent if $h \longrightarrow 0$ and $nh \longrightarrow \infty$.

Examples:

Gaussian kernel: $K(u) = (2\pi)^{-1/2}e^{-u^2/2}$.

Epanechnikov: $K(u) = 1(|u| \leq 1)(1 - u^2)(3/4)$.

Choice of K does not matter as much as choice of h. Epanechnikov kernel has slightly smaller mean square error, and so optimal.

Bias and Variance of Kernel Estimators

$(Z_1, ..., Z_n)$ are i.i.d..

Bias: $f_0(z)$ is pdf of Z_i. Expectation of kernel estimator; with

$$\begin{aligned} E[\hat{f}_h(z)] &= \int K_h(z-t)f_0(t)dt = \frac{1}{h}\int K(\frac{z-t}{h})f_0(t)dt \\ &= \int K(u)f_0(z-hu)du, \end{aligned}$$

for change of variables $u = (z-t)/h$. Taylor expand $f_0(z-hu)$ around $h = 0$,

$$\begin{aligned} f_0(z-hu) &= f_0(z) - f_0'(z)hu + \Gamma(h,u,z)h^2, \\ \Gamma(h,u,z) &= f_0''(z+\bar{h}(z,u)u)u^2/2, \end{aligned}$$

where $|\bar{h}(z,u)| \leq |h|$. For $\int K(u)u^2du < \infty$, $\int K(u)udu = 0$, assuming $f_0''(z)$ continuous and bounded,

$$\int K(u)\Gamma(h,u,z)du \longrightarrow [\int K(u)u^2du]f_0''(z)/2.$$

Then for $o(h^2) = a(h)$ with $\lim_{h\longrightarrow 0}a(h)/h^2 = 0$,

$$\begin{aligned} h^2\int K(u)\Gamma(h,u,z)du &= h^2[\int K(u)u^2du]f_0''(z)/2 \\ &\quad +o(h^2) \end{aligned}$$

Then multiplying the expansion

$$f_0(z-hu) = f_0(z) - f_0'(z)hu + \Gamma(h,u,z)h^2$$

by $K(u)$ and integrating gives

$$E[\hat{f}_h(z)] = f_0(z) + h^2f_0''(z)\int K(u)u^2du/2 + o(h^2).$$

We can summarize these calculations in the following result:

PROPOSITION 1: *If $f_0(z)$ is twice continuously differentiable with bounded second derivative, $\int K(u)du = 1, \int K(u)udu = 0, \int u^2K(u)du < \infty$, then*

$$E[\hat{f}_h(z)] - f_0(z) = h^2f_0''(z)\int K(u)u^2du/2 + o(h^2).$$

Variance: From Proposition 1, $E[K_h(z-Z_i)] = E[\hat{f}_h(z)]$ is bounded as $h \longrightarrow 0$. Let $O(1/n)$ denote $(a_n)_{n=1}^{\infty}$ such that na_n is bounded. Then by $\hat{f}_h(z)$ a sample of average of $K_h(z-Z_i)$, for $h \longrightarrow 0$,

$$\begin{aligned} Var(\hat{f}_h(z)) &= \{E[K_h(z-Z_i)^2] - \{E[K_h(z-Z_i)]\}^2\}/n \\ &= \frac{1}{h^2}\int K(\frac{z-t}{h})^2 f_0(t)dt/n + O(1/n) \\ &= \frac{1}{h}\int K(u)^2 f_0(z-hu)du/(nh) + O(1/n). \end{aligned}$$

For $f_0(z)$ continuous and bounded and $\int K(u)^2du < \infty$,

$$\int K(u)^2 f_0(z-hu)du \longrightarrow f_0(z)\int K(u)^2du.$$

By $h \longrightarrow 0$, it follows that $nhO(1/n) \longrightarrow 0$, so that $O(1/n) = o(1/nh)$. Plugging in above variance formula we find,

$$Var(\hat{f}_h(z)) = f_0(z)\int K(u)^2du/(nh) + o(1/(nh)).$$

We can summarize these calculations in the following result:

PROPOSITION 2: *If $f_0(z)$ is continuous and bounded, $\int K(u)^2du < \infty$, $h \longrightarrow 0$, and $nh \to \infty$ then*

$$Var[\hat{f}_h(z)] = f_0(z)\int K(u)^2du/(nh) + o(1/(nh)).$$

Consistency and Convergence Rate of Kernel Estimators

For consistency implied by

$$\begin{aligned} h &\longrightarrow 0; \text{ bias goes to zero.} \\ nh &\longrightarrow \infty; \text{ variance goes to zero.} \end{aligned}$$

Bandwidth shrinks to zero slower than $1/n$.

Intuition for the $h \longrightarrow 0$: Smoothing "noise" must go away asymptotically to remove all bias.

Intuition for $nh \longrightarrow \infty$: For Epanechnikov kernel; $K((z-Z_i)/h) > 0$ if and only if $|z-Z_i| < h$. If h shrinks as fast as or faster than $1/n$, the number of observations with $|z-Z_i| < h$ will not grow, so averaging over a finite number of observations, hence variance does not go to zero.

Explicit form for (MSE) under $h \longrightarrow 0, nh \longrightarrow \infty$.

$$\begin{aligned} MSE(\hat{f}_h(z)) &= Var(\hat{f}_h(z)) + Bias^2(\hat{f}_h(z)) \\ &= f_0(z) \int K(u)^2 du/(nh) \\ &\quad + h^4 \{ f_0''(z) \int K(u) u^2 du/2 \}^2 \\ &\quad + o(h^4 + 1/(nh)). \end{aligned}$$

By $h \longrightarrow 0$, MSE vanishes slower than $1/n$. Thus, kernel estimator converges slower than $n^{-1/2}$. Avoidance of bias by $h \longrightarrow 0$ means fraction of the observations used goes to zero.

Bandwidth Choice for Density Estimation:

Graphical: Choose one that looks good, report several.

Minimize asymptotic integrated MSE. Integrating over z,

$$\begin{aligned} \int MSE(\hat{f}_h(z)) dz &= \int K(u)^2 du/(nh) \\ &\quad + \int f_0''(z)^2 dz [\int K(u) u^2 du/2]^2 h^4 \\ &\quad + o(h^4 + 1/(nh)). \end{aligned}$$

Min over h has first-order conditions

$$\begin{aligned} 0 &= -n^{-1} h^{-2} C_1 + h^3 \int f_0''(z)^2 dz C_2, \\ C_1 &= \int K(u)^2 du, C_2 = [\int K(u) u^2 du]^2. \end{aligned}$$

Solving gives

$$h = [C_1 / \{ n C_2 \int f_0''(z)^2 dz \}]^{1/5}.$$

$$h = [C_1 / \{ n C_2 \int f_0''(z)^2 dz \}]^{1/5}.$$

Asymptotically optimal bandwidth. Could be estimated by "plugging-in" estimator for $f_0''(z)$. This procedure depends on initial bandwidth, but final estimator not as sensitive to choice of bandwidth for $f_0''(z)$ as choice of bandwidth for $f_0(z)$.

Silverman's rule of thumb: Optimal bandwidth when $f_0(z)$ is Gaussian and $K(u)$ is a standard normal pdf.

$$h = 1.06 \sigma n^{1/5}, \ \sigma = Var(z_i)^{1/2}.$$

Use estimator of the standard deviation σ.

Estimate directly the integrated MSE:

$$\begin{aligned}\int MSE(\hat{f}_h(z))dz &= \int E[\{\hat{f}_h(z) - f_0(z)\}^2]dz \\ &= E[\int \{\hat{f}_h(z) - f_0(z)\}^2] \\ &= E[\int \hat{f}_h(z)^2] - 2E[\int \hat{f}_h(z)f_0(z)] \\ &\quad + f_0(z)^2.\end{aligned}$$

Unbiased estimator of $E[\int \hat{f}_h(z)^2]$ is

$$\int \hat{f}_h(z)^2 dz = \sum_{i,j} \int K_h(Z_i - t)K_h(t - Z_j)dt/n^2$$

To find unbiased estimator of second, note that

$$E[\int \hat{f}_h(z)f_0(z)dz] = \int \int K_h(z - t)f_0(z)f_0(t)dzdt.$$

By observations independent, we can average over pairs to estimate this term. Last term in MSE does not depend on h, so we can drop. Combining estimates of first two terms gives criterion.

$$C\hat{V}(h) = \int \hat{f}_h(z)^2 dz - \frac{2}{n(n-1)} \sum_{i \neq j} K_h(Z_i - Z_j).$$

$$C\hat{V}(h) = \int \hat{f}_h(z)^2 dz - \frac{2}{n(n-1)} \sum_{i \neq j} K_h(Z_i - Z_j).$$

Choosing h by minimizing $C\hat{V}(h)$ is called *cross-validation.* Motivation for terminology is second term is

$$\begin{aligned}&-2\sum_{i=1}^{n} \hat{f}_{-i,h}(Z_i)/n, \\ \hat{f}_{-i,h}(z) &= \sum_{j \neq i} K_h(z - Z_j)/(n-1).\end{aligned}$$

Here $\hat{f}_{-i,h}(z)$ is estimator that uses all observations but the ith, so that $\hat{f}_{-i,h}(Z_i)$ is "cross-validated."

Multivariate Density Estimation:

Multivariate density estimation can be important as in example. Let z be $r \times 1$, $K(u)$ denote a pdf for a $r \times 1$ random vector, e.g. $K(u) = \Pi_{j=1}^r k(u_j)$ for univariate pdf $k(u)$. Let $\hat{\Sigma}$ the sample covariance matrix of Z_i. For a bandwidth h let

$$K_h(u) = h^{-r} \det(\hat{\Sigma})^{-1/2} K(\hat{\Sigma}^{-1/2}u/h),$$

$A^{-1/2}$ is inverse square root of positive definite A. Often $K(u) = \rho(u'u)$ for some ρ, so

$$K_h(u) = h^{-r} \det(\hat{\Sigma})^{-1/2} \rho(u'\hat{\Sigma}^{-1}u/h).$$

Multivariate kernel estimator, with scale normalization

$$\hat{f}_h(z) = \sum_{i=1}^{n} K_h(z - Z_i)/n.$$

Gaussian kernel: $K(u) = (2\pi)^{-r/2} \exp(-u'u/2)$.
Epanechnikov kernel: $K(u) = C_r(1 - u'u)1(u'u \leq 1)$.

The Curse of Dimensionality for Kernel Estimation

Difficult to nonparametrically estimate pdf's of high dimensional Z_i. Need many observations within a small distance of a point. As dimension rises with distance fixed, the proportion of observations that are close shrinks very rapidly. Mathematically, with one dimension $[0, 1]$ can be covered with $1/h$ balls of radius h, while it requires $\frac{1}{h^r}$ balls to cover $[0, 1]^r$. Thus, if data equally likely to fall in one ball, tend to be many fewer data points in any one ball for high dimensional data.

Silverman (1986, Density Estimation) famous table. Multivariate normal density at zero, the sample size required for MSE to be 10 percent of the density

r	5	6	7	8	9	10
n	768	2790	10,700	43,700	187,000	842,000

Curse of dimensionality shows up in convergence rate. Expand again,

$$\begin{aligned} f_0(z - hu) &= f_0(z) - h[\partial f_0(z)/\partial z]'u \\ &\quad + h^2 u'[\partial^2 f_0(z + \bar{h}(z, u)u)/\partial z \partial z']u/2, \end{aligned}$$

so bias asymptotically $C_3 h^2$ no matter how big r. For the variance, setting $u = (z - t)/h$

$$\begin{aligned} \int K_h(z - t)^2 f_0(t) d &= h^{-2r} \int K((z - t)/h)^2 f(t) dt \\ &= h^{-r} \int K(u)^2 f_0(z - hu) du. \end{aligned}$$

Integrating, with $\hat{\Sigma} = I$,

$$\int E[\{\hat{f}_h(z) - f_0(z)\}^2] dz \approx C_1/(nh^r) + C_3^2 h^4.$$

First-order conditions for the optimal h : $0 = -C_1 r n^{-1} h^{-r-1} + 4C_2 h^3$. Solving gives $h = C' n^{-1/(r+4)}$. Plugging back in gives

$$\int E[\{\hat{f}_h(z) - f_0(z)\}^2] dz \approx C' n^{-4/(r+4)}.$$

The convergence rate declines as r increases.

Nonparametric Regression

Often in econometrics the object of estimation is a regression function. A classical formulation is $Y = X'\beta + \varepsilon$ with $E[\varepsilon|X] = 0$. A more direct way to write this model is

$$E[Y|X] = X'\beta.$$

A nonparametric version of this model, that allows for unknown functional form in the regression, is

$$E[Y|X] = g_0(X),$$

where $g_0(x)$ is an unknown function.

Kernel Regression

To estimate $g_0(x)$ nonparametrically, one can start with kernel density estimate and plug it into the formula

$$\begin{aligned} g_0(x) &= \int yf(y|x)dy \\ &= \int yf(y,x)dy / \int f(y,x)dy \end{aligned}$$

Assume that X is a scalar. Let $k(u_1, u_2)$ be a bivariate kernel, with $\int t \cdot k(t, u_2)dt = 0$. Data are $(Y_1, X_1), ..., (Y_n, X_n)$. Let $K(u_2) = \int k(t, u_2)dt$. By change of variables $t = (y - Y_i)/h$,

$$\begin{aligned} \int y\hat{f}_h(y,x)dy &= n^{-1}h^{-2}\sum_{i=1}^n \int yk(\frac{y-Y_i}{h}, \frac{x-X_i}{h})dy \\ &= n^{-1}h^{-1}\sum_{i=1}^n \int (Y_i + ht)k(t, \frac{x-X_i}{h})dt \\ &= n^{-1}h^{-1}\sum_{i=1}^n Y_i \int k(t, \frac{x-X_i}{h})dt \\ &= n^{-1}h^{-1}\sum_{i=1}^n Y_i K(\frac{x-X_i}{h}), \\ \int \hat{f}_h(y,x)dy &= n^{-1}h^{-2}\sum_{i=1}^n \int k(\frac{y-Y_i}{h}, \frac{x-X_i}{h})dy \\ &= n^{-1}h^{-1}\sum_{i=1}^n K(\frac{x-X_i}{h}). \end{aligned}$$

Plugging in $\hat{f}_h(y,x)$ in formula for $g_0(x)$,

$$\widehat{g}_h(x) = \frac{\int y\hat{f}_h(y,x)dy}{\int \hat{f}_h(y,x)dy} = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}.$$

Also, kernel regression estimator $\hat{g}(x)$ just defined by second equality.

This regression estimator is a weighted average, with

$$\hat{g}_h(x) = \sum_{i=1}^{n} w_i^h(x) Y_i, w_i^h(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}$$

By construction, $\sum_{i=1}^{n} w_i^h(x) = 1$, while $w_i^h(x)$ is nonnegative by $K(u)$ being nonnegative. For symmetric $K(u)$ with a unique mode at $u = 0$, more weight will be given to observations with X_i that is closer to x. Bandwidth h controls how fast the weights decline. As h declines, more weight given closer observations. Reduces bias but increases variance.

For multivariate X, formula is the same, with $K(u)$ replaced by multivariate kernel, such as

$$K(u) = \det(\hat{\Sigma})^{-1/2} k(\hat{\Sigma}^{-1/2} u)$$

for some kernel $k(u)$. Consistency and convergence rate results are similar. Details are not reported here because the calculations are complicated by the ratio form of the estimator. Bandwidth choice below.

Series Regression

Another approach to nonparametric regression flexible functional forms with complete flexibility by letting the number of terms grow with sample size. Think of approximating $g_0(x)$ by linear combination $\sum_{j=1}^{K} p_{jK}(x)\beta_j$ of approximating functions $p_{jK}(x)$, e.g. polynomials or splines. Estimator of $g_0(x)$ is predicted value from regressing Y_i on $p^K(X_i)$ for $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))'$. Consistent as K grows with n.

Let $Y = (Y_1, ..., Y_n)'$ and $P = [P^K(X_1), ..., P^K(X_n)]'$. Then

$$\hat{g}(x) = p^K(x)'\hat{\beta}, \hat{\beta} = (P'P)^{-} P'Y,$$

A^{-} denotes generalized inverse, $AA^{-}A = A$. Different than kernal; global fit rather than local average.

Examples:

Power series: x scalar

$$p_{jK}(x) = x^{j-1}, (j = 1, 2, ...)$$

$p^K(x)'\beta$ is a polynomial. Such approximations good for global approximation of smooth functions. Not when most variation in narrow range or when jump or kink. Using orthogonal polynomials with respect to density can mitigate multicollinearity.

Regression splines: x is scalar,

$$\begin{aligned} p_{jK}(x) &= x^{j-1}, (j = 1, 2, 3, 4), \\ p_{jK}(x) &= 1(x > \ell_{j-4,K})(x - \ell_{j-4,K})^3, (j = 5, ..., K). \end{aligned}$$

The $\ell_1, ..., \ell_{K-4}$ are "knots," $p^K(x)'\beta$ is cubic in between ℓ_{jK}, twice continuously differentiable everywhere. Picks up local variation but still global fit. Need to place knots. B-splines can mitigate multi-collinearity.

Extend both to multivariate case by including products of individual components.

Convergence Rate for Series Regression

Depends on approximation rate, i.e. bias.

ASSUMPTION A: *There exists* $\gamma > 0$, $\bar{\beta}_K$, *and* C *such that*

$$\{E[\{g_0(X_i) - p^K(X_i)'\bar{\beta}_K\}^2]\}^{1/2} \leq CK^{-\gamma}.$$

Comes from approximation theory. For power series, X_i in a compact set, $g_0(x)$ is continuously differentiable of order s, then $\gamma = s/r$. For splines, $\gamma = \min\{4, s\}/r$. For multivariate, approximation depends on the order of the included terms (e.g. on power) which grows more slowly with K when x is higher dimensional.

PROPOSITION 3: *If Assumption A is satisfied and* $Var(Y_i|X_i) \leq \Delta$ *then*

$$E[\{\hat{g}_K(X_i) - g_0(X_i)\}^2] \leq \Delta K/n + C^2K^{-2\gamma}.$$

Proof: Let

$$\begin{aligned} Q &= P(P'P)^-P', \hat{g} = (\hat{g}(X_1), ..., \hat{g}(X_n))' = QY, \\ g_0 &= (g_0(X_1), ..., g_0(X_n))', \varepsilon = Y - g_0, \bar{g} = P\bar{\beta}_K. \end{aligned}$$

Q idempotent, so $I-Q$ idempotent, hence has eigenvalues that are zero or one. Therefore, by Assumption A,

$$\begin{aligned} & E\left[(g_0 - \bar{g})(I - Q)(g_0 - \bar{g})\right] \\ \leq\ & E\left[(g_0 - \bar{g})'(g_0 - \bar{g})\right] \\ \leq\ & nE\left[\{g_0(X_i) - p^K(X_i)'\bar{\beta}_K\}^2\right] \\ \leq\ & CnK^{-2\gamma}. \end{aligned}$$

Also, for $X = (X_1, ..., X_n)$, by independence and iterated expectations, for $i \neq j$,

$$\begin{aligned} E[\varepsilon_i\varepsilon_j|X] &= E[\varepsilon_i\varepsilon_j|X_i, X_j] \\ &= E[\varepsilon_i E[\varepsilon_j|X_i, X_j, \varepsilon_i]|X_i, X_j] \\ &= E[\varepsilon_i E[\varepsilon_j|X_j]|X_i, X_j] = 0. \end{aligned}$$

Then for $\Lambda_{ii} = Var(Y_i|X_i)$ and $\Lambda = \text{diag}(\Lambda_{11}, ..., \Lambda_{nn})$ we have $E[\varepsilon\varepsilon'|X] = \Lambda$. It follows that for $tr(A)$ the trace of a square matrix A, by $rank(Q) \leq K$,

$$\begin{aligned} E\left[\varepsilon' Q \varepsilon | X\right] &= \operatorname{tr}\left(Q E\left[\varepsilon\varepsilon' | X\right]\right) = \operatorname{tr}(Q\Lambda) \\ &= \operatorname{tr}(Q\Lambda Q) \leq \Delta \operatorname{tr}(Q) \leq \Delta K. \end{aligned}$$

Then by iterated expectations, $E[\varepsilon' Q \varepsilon] \leq CK$. Also,

$$\begin{aligned} & \sum_{i=1}^{n} \{\hat{g}(X_i) - g_0(X_i)\}^2 \\ = \; & (\hat{g} - g_0)'(\hat{g} - g_0) \\ = \; & (Q\varepsilon - (I-Q)g_0)'(Q\varepsilon - (I-Q)g_0) \\ = \; & \varepsilon' Q \varepsilon + g_0'(I-Q)g_0 \\ = \; & \varepsilon' Q \varepsilon + (g_0 - \bar{g})'(I-Q)(g_0 - \bar{g}). \end{aligned}$$

Then by i.i.d. observations,

$$\begin{aligned} & E[\{\hat{g}(X_i) - g_0(X_i)\}^2] \\ = \; & E[(\hat{g}_i - g_{0i})^2] \\ = \; & E[(\hat{g} - g_0)'(\hat{g} - g_0)]/n \\ \leq \; & \Delta \frac{K}{n} + C^2 K^{-2\gamma}. \end{aligned}$$

Choosing Bandwidth or Number of Terms

Data based choice operationalizes complete flexibility. Number of terms or bandwidth adjusts. Cross-validation is common. Let $\hat{g}_{-i,h}(X_i)$ and $\hat{g}_{-i,K}(X_i)$ be predicted values for i^{th} observation using all the others, for kernel and series respectively.

$$\begin{aligned} \hat{g}_{-i,h}(X_i) &= \frac{\sum_{j \neq i} Y_j K\left(\frac{X_i - X_j}{h}\right)}{\sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)}, \\ \hat{g}_{-i,K}(X_i) &= Y_i - \frac{Y_i - \hat{g}_K(X_i)}{1 - P^K(X_i)'(P'P)^- P^K(X_i)}, \end{aligned}$$

Second equality by recursive residuals. Criteria are

$$\begin{aligned} C\hat{V}(h) &= \sum_{i=1}^{n} v(X_i)[Y_i - \hat{g}_{-i,h}(X_i)]^2, \\ C\hat{V}(K) &= \sum_{i=1}^{n} [Y_i - \hat{g}_{-i,K}(X_i)]^2, \end{aligned}$$

$v(x)$ is a weight function, equal to zero near support boundary$_i$. Choose h or K to minmize.

(weighted) sample MSE is

$$MSE(h) = \sum_{i=1}^{n} v(X_i)\{\hat{g}_h(X_i) - g_0(X_i)\}^2/n$$

Cross-validation criteria optimal,

$$\frac{\min_h MSE(h)}{MSE(\hat{h})} \xrightarrow{p} 1.$$

Series too. $\hat{h}$ converges slowly.

Another Empirical Example:

Hausman and Newey (1995, Nonparametric Estimation of Exact Consumers Surplus, Econometrica), kernel and series estimates, Y is log of gasoline purchased, function of price, income, and time and location dummies. Six cross-sections of individuals from Energy Department, total of 18,109 observations. Cross-validation criteria are

Kernel		Spline		Power	
h	CV	Knots	CV	Order	CV
1.6	4621	1	4546	1	4534
1.9	4516	2	4543	2	4539
2.0	4508	3	4546	3	4512
2.1	4700	4	4551	4	4505
		5	4545	5	4507
		6	4552	6	4507
		7	4546	7	4500
		8	4551	8	4493
		9	4552	9	4494

Locally Linear Regression:

There is another local method, locally linear regression, that is thought to be superior to kernel regression. It is based on a locally fitting a line rather than a constant. Unlike kernel regression, locally linear estimation would have no bias if the true model were linear. In general, locally linear estimation removes a bias term from the kernel estimator, that makes it have better behavior near the boundary of the x's and smaller MSE everywhere.

To describe this estimator, let $K_h(u) = h^{-r}K(u/h)$ as before. Consider the estimator $\hat{g}(x)$ given by the solution to

$$\min_{g,\beta} \sum_{i=1}^{n} (Y_i - g - (x - X_i)'\beta)^2 K_h(x - X_i).$$

That is $\hat{g}(x)$ is the constant term in a weighted least squares regression of Y_i on $(1, x-X_i)$, with weights $K_h(x-X_i)$. For

$$\begin{aligned} Y &= \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \tilde{X} = \begin{pmatrix} 1 & (x-X_1)' \\ \vdots & \vdots \\ 1 & (x-X_n)' \end{pmatrix} \\ W &= \text{diag}\left(K_h(x-X_1), ..., K_h(x-X_n)\right) \end{aligned}$$

and e_1 a $(r+1)\times 1$ vector with 1 in first position and zeros elsewhere, we have

$$\hat{g}(x) = e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'WY.$$

This estimator depends on x both through the weights $K_h(x-X_i)$ and through the regressors $x-X_i$.

This estimator is a locally linear fit of the data. It runs a regression with weights that are smaller for observations that are farther from x. In contrast, the kernel regression estimator solves this same minimization problem but with β constrained to be zero, i.e., kernel regression minimizes

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

Removing the constraint $\beta = 0$ leads to lower bias without increasing variance when $g_0(x)$ is twice differentiable. It is also of interest to note that $\hat{\beta}$ from the above minimization problem estimates the gradient $\partial g_0(x)/\partial x$.

Like kernel regression, this estimator can be interpreted as a weighted average of the Y_i observations, though the weights are a bit more complicated. Let

$$\begin{aligned} S_0 &= \sum_{i=1}^{n} K_h(x-X_i), \ S_1 = \sum_{i=1}^{n} K_h(x-X_i)(x-X_i), S_2 = \sum_{i=1}^{n} K_h(x-X_i)(x-X_i)(x-X_i)' \\ \hat{m}_0 &= \sum_{i=1}^{n} K_h(x-X_i)Y_i, \hat{m}_1 = \sum_{i=1}^{n} K_h(x-X_i)(x-X_i)Y_i. \end{aligned}$$

Then, by the usual partitioned inverse formula

$$\begin{aligned} \hat{g}(x) &= e_1' \begin{bmatrix} S_0 & S_1' \\ S_1 & S_2 \end{bmatrix}^{-1} \begin{pmatrix} \hat{m}_0 \\ \hat{m}_1 \end{pmatrix} = (S_0 - S_1'S_2^{-1}S_1)^{-1}(\hat{m}_0 - S_1'S_2^{-1}\hat{m}_1) \\ &= \frac{\sum_{i=1}^{n} a_i Y_i}{\sum_{i=1}^{n} a_i}, \ a_i = K_h(x-X_i)[1 - S_1'S_2^{-1}(x-X_i)] \end{aligned}$$

It is straightforward though a little involved to find asymptotic approximations to the MSE. For simplicity we do this for scalar x case. Note that for $g_0 = (g_0(X_1), ..., g_0(X_n))'$

$$\hat{g}(x) - g_0(x) = e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W(Y-g_0) + e'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'Wg_0 - g_0(x).$$

Then for $\Sigma = diag(\sigma^2(X_1), ..., \sigma^2(X_n))$,

$$\begin{aligned} E\left[\{\hat{g}(x) - g_0(x)\}^2 \mid X_1, ..., X_n\right] &= e'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W\Sigma W\tilde{X}(\tilde{X}'W\tilde{X})^{-1}e \\ &\quad + \left\{e'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'Wg_0 - g_0(x)\right\}^2 \end{aligned}$$

An asymptotic approximation to MSE is obtained by taking the limit as n grows. Note that we have

$$n^{-1}h^{-j}S_j = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)[(x - X_i)/h]^j$$

Then, by the change of variables $u = (x - X_i)/h$,

$$E\left[n^{-1}h^{-j}S_j\right] = E[K_h(x - X_i)\{(x - X_i)/h\}^j] = \int K(u)u^j f_0(x - hu)du = \mu_j f_0(x) + o(1).$$

for $\mu_j = \int K(u)u^j du$ and $h \longrightarrow 0$. Also,

$$\begin{aligned} \text{var}\left(n^{-1}h^{-j}S_j\right) &\leq n^{-1}E\left[K_h(x - X_i)^2[(x - X_i)/h\right]^{2j}] \leq n^{-1}h^{-1}\int K(u)^2u^{2j}f_0(x - hu)du \\ &\leq Cn^{-1}h^{-1} \longrightarrow 0 \end{aligned}$$

for $nh \longrightarrow \infty$. Therefore, for $h \longrightarrow 0$ and $nh \longrightarrow \infty$

$$n^{-1}h^{-j}S_j = \mu_j f_0(x) + o_p(1).$$

Now let $H = \text{diag}(1, h)$. Then by $\mu_0 = 1$ and $\mu_1 = 0$ we have

$$n^{-1}H^{-1}\tilde{X}'W\tilde{X}H^{-1} = n^{-1}\begin{bmatrix} S_0 & h^{-1}S_1 \\ h^{-1}S_1 & h^{-2}S_2 \end{bmatrix} = f_0(x)\begin{bmatrix} 1 & 0 \\ 0 & \mu_2 \end{bmatrix} + o_p(1).$$

Next let $\nu_j = \int K(u)^2u^j du$. then by a similar argument we have

$$h\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)^2\left[(x - X_i)/h\right]^j \sigma^2(X_i) = \nu_j f_0(x)\sigma^2(x) + o_p(1).$$

It follows by $\nu_1 = 0$ that

$$n^{-1}hH^{-1}\tilde{X}'W\Sigma W\tilde{X}H^{-1} = f_0(x)\sigma^2(x)\begin{bmatrix} \nu_0 & 0 \\ 0 & \nu_2 \end{bmatrix} + o_p(1).$$

Then we have, for the variance term, by $H^{-1}e_1 = e_1$,

$$\begin{aligned} & e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W\Sigma W\tilde{X}(\tilde{X}'W\tilde{X})^{-1}e_1 \\ = \; & n^{-1}h^{-1}e_1'H^{-1}\left(\frac{H^{-1}\tilde{X}'W\tilde{X}H^{-1}}{n}\right)^{-1}\frac{hH^{-1}\tilde{X}'W\Sigma W\tilde{X}H^{-1}}{n}\left(\frac{H^{-1}\tilde{X}'W\tilde{X}H^{-1}}{n}\right)^{-1}H^{-1}e_1 \\ = \; & n^{-1}h^{-1}\left[\left(e_1'\begin{bmatrix} 1 & 0 \\ 0 & \mu_2 \end{bmatrix}^{-1}\begin{bmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & \mu_2 \end{bmatrix}^{-1}e_1\right)\frac{\sigma^2(x)}{f(x)} + o_p(1)\right]. \end{aligned}$$

It then follows that

$$e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W\Sigma W\tilde{X}(\tilde{X}'W\tilde{X})^{-1}e_1 = n^{-1}h^{-1}\left(\nu_0\frac{\sigma^2(x)}{f(x)} + o_p(1).\right)$$

For the bias consider an expansion

$$g(X_i) = g_0(x) + g_0'(x)(X_i - x) + \frac{1}{2}g_0''(x)(X_i - x)^2 + \frac{1}{6}g_0'''(\bar{X}_i)(X_i - x)^3.$$

Let $r_i = g_0(X_i) - g_0(x) - [dg_0(x)/dx](X_i - x)$. Then by the form of $\tilde{X}$ we have

$$g = (g_0(X_1), ..., g_0(X_n))' = g_0(x)We_1 - g_0'(x)We_2 + r$$

It follows by $e_1'e_2 = 0$ that the bias term is

$$\begin{aligned} &e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'Wg - g_0(x) = e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W\tilde{X}e_1g_0(x) - g_0(x) \\ &+e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W\tilde{X}e_2g_0'(x) + e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'Wr = e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'Wr. \end{aligned}$$

Recall that

$$n^{-1}h^{-j}S_j = \mu_j f_0(x) + o_p(1).$$

Therefore, by $\mu_3 = 0$,

$$\begin{aligned} &n^{-1}h^{-2}H^{-1}\tilde{X}'W((x - X_1)^2, ..., (x - X_n)^2)'\frac{1}{2}g_0''(x) \\ = &\begin{pmatrix} n^{-1}h^{-2}S_2 \\ n^{-1}h^{-3}S_3 \end{pmatrix}\frac{1}{2}g_0''(x) = f_0(x)\begin{pmatrix} \mu_2 \\ 0 \end{pmatrix}\frac{1}{2}g_0''(x) + o_p(1). \end{aligned}$$

Assuming that $g_0'''(\bar{X}_i)$ is bounded, bounded

$$\begin{aligned} &\left\|n^{-1}h^{-2}H^{-1}\tilde{X}'W\left((x - X_1)^3g_0'''(\bar{X}_1), ..., (x - X_n)^3g_0'''\left(\bar{X}_n\right)\right)'\right\| \\ \leq\ & C\max\left\{n^{-1}h^{-2}\sum_i K_h(x - X_i)\left|x - X_i\right|^3,\ n^{-1}h^{-2}S_4\right\} \longrightarrow 0. \end{aligned}$$

Therefore, we have

$$\begin{aligned} e_1'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'Wr &= h^2e_1'H^{-1}\left(\frac{H^{-1}\tilde{X}'W\tilde{X}H^{-1}}{n}\right)^{-1}\frac{h^{-2}H^{-1}\tilde{X}'Wr}{n} \\ &= \frac{h^2}{2}\left[g_0''(x)e_1'\begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix}^{-1}\begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} + o_p(1)\right] \\ &= \frac{h^2}{2}[g_0''(x)\mu_2 + o_p(1)]. \end{aligned}$$

Exercise: Apply analogous calculation to show kernel regression bias is

$$\mu_2 h^2 \left(\frac{1}{2} g_0''(x) + g_0'(x) \frac{f_0'(x)}{f_0(x)} \right)$$

Notice bias is *zero* if function is linear.

Combining the bias and variance expression, we have the following form for asymptotic MSE:

$$\frac{1}{nh} \nu_0 \frac{\sigma^2(x)}{f_0(x)} + \frac{h^4}{4} g_0''(x)^2 \mu_2^2.$$

In contrast, the kernel MSE is

$$\frac{1}{nh} \nu_0 \frac{\sigma^2(x)}{f_0(x)} + \frac{h^4}{4} \left[g_0''(x) + 2 g_0'(x) \frac{f_0'(x)}{f_0(x)} \right]^2 \mu_2^2.$$

Bias will be much bigger near boundary of the support where $f_0'(x)/f_0(x)$ is large. For example, if $f_0(x)$ is approximately x^α for $x > 0$ near zero, then $f_0'(x)/f_0(x)$ grows like $1/x$ as x gets close to zero. Thus, locally linear has smaller boundary bias. Also, locally linear has no bias if $g_0(x)$ is linear but kernel obviously does.

Simple bandwidth choice method is to take expected value of MSE.

One could use a plug in method to minimize integrated asymptotic MSE, integrated over $\omega(x) f_0(x)$ for some weight.

Reducing the Curse of Dimensionality

Idea: Restrict form of regression so that it only depends on low dimensional components. Additive model has regression additive in lower dimension components. Index model has regression depending only on a linear combination. Additive model, $X = (x_1, ..., x_r)'$,

$$E[Y|X] = \sum_{j=1}^{r} g_{j0}(X_j).$$

One dimensional rate. Series estimator is simplest. Restricts approximating functions to depend on only on component. Scalar u and $p_{\ell L}(u), \ell = 1, ..., L$ approximating functions, $p^L(u) = (p_{1L}(u), ..., p_{LL}(u))'$, $K = Lr + 1$ let $p^K(x) = (1, p^L(x_1)', ..., p^L(x_r)')'$. Regress Y_i on $p^K(X_i)$. For $\hat{\beta}^0$ equal to constant and $\hat{\beta}^j, (j = 1, ..., r)$ the coefficient vector for $p^L(x_j)/$

$$\hat{g}(X) = \hat{\beta}^0 + \sum_{j=1}^{r} p^L(x_j)' \hat{\beta}^j.$$

PROPOSITION 4: *If Assumption A is satisfied with $g_0(X)$ equal to each component $g_{j0}(X_j)$ and $p^K(X)$ replaced by $(1, p^L(X_j)')'$, where the constants C, γ do not depend on j, and $Var(Y|X) \leq \Delta$ then*

$$\begin{aligned} & E[\{\hat{g}_K(X_i) - g_0(X_i)\}^2] \\ \leq \ & \Delta/n + r[\Delta L/n + C^2(L+1)^{-2\gamma}]. \end{aligned}$$

Proof: Let $\tilde{p}^L(u) = (1, p^L(u)')'$. By Assumption A, is C and γ so for each j and L is a $(L+1) \times 1$ vector $\tilde{\beta}^{jL} = (\tilde{\beta}_1^{jL}, \tilde{\beta}_2^{jL\prime})'$,

$$E[\{g_{j0}(X_j) - \tilde{p}^L(X_j)'\tilde{\beta}^{jL}\}^2] \leq C^2(L+1)^{-2\gamma}.$$

Let $\bar{\beta} = (\sum_{j=1}^r \tilde{\beta}_1^{jL}, \tilde{\beta}_2^{1L\prime}, ..., \tilde{\beta}_2^{rL\prime})'$, so that

$$\begin{aligned} & E[\{g(X) - p^K(X)'\bar{\beta}\}^2] \\ = \ & \sum_{j=1}^r E[\{g_{j0}(X_j) - \tilde{p}^L(X_j)'\tilde{\beta}^{jL}\}^2] \\ \leq \ & rC^2(L+1)^{-2\gamma}. \end{aligned}$$

Then as in the proof of Propostion 3, we have

$$\begin{aligned} & E[\{\hat{g}_K(X) - g_0(X)\}^2] \\ \leq \ & \Delta K/n + rC^2(L+1)^{-2\gamma} \\ = \ & \Delta/n + r[\Delta L/n + C^2(L+1)^{-2\gamma}]. \text{ Q.E.D.} \end{aligned}$$

Convergence rate does not depend on r, although does affect. Here r could even grow with sample size at some power of n. Additivity condition satisfied in Hausman and Newey (1995).

Semiparametric Models

Data: $Z_1, Z_2, ...$ i.i.d.

Model: $\mathcal{F}$ a set of pdfs.

Correct specification: pdf f_0 of Z_i in $\mathcal{F}$.

Semiparametric model: $\mathcal{F}$ has parametric θ and nonparametric components.

Ex: Linear model $E[Y|X] = X'\beta_0$; parametric component is β, everything else nonparametric.

Ex: Probit, $Y \in \{0, 1\}$, $\Pr(Y = 1|X) = \Phi(X'\beta_0)$ is parametric component, nonparametric component is distribution of X.

Binary Choice with Unknown Disturbance Distribution: $Z = (Y, X)$, $v(x, \beta)$ a known function,

$$Y = 1(Y^* > 0), Y^* = v(X, \beta_0) - \varepsilon, \varepsilon \text{ independent of } X,$$

This model implies

$$\Pr(Y = 1|X) = G(v(X, \beta_0)),$$

Parameter β, everything else, including $G(u)$, is nonparametric. The $v(x, \beta)$ notation allows location and scale normalization, e.g. $v(x, \beta) = x_1 + x_2'\beta$, $x = (x_1, x_2')'$, x_1 scalar.

Censored Regression with Unknown Disturbance Distribution: $Z = (Y, X)$,

$$Y = \max\{0, Y^*\}, Y^* = X'\beta_0 + \varepsilon, \varepsilon \text{ independent of } X;$$

Parameter β, everything else, including distribution of ε, is nonparametric.

Binary choice and censored regression are limited dependent variable models. Semiparametric models are important here because misspecifying the distribution of the disturbances leads to inconsistency of MLE.

Partially Linear Regression: $Z = (Y, X, W)$,

$$E[Y|X, W] = X'\beta_0 + g_0(W).$$

Parameter β, everything else nonparametric, including additive component of regression. Can help with curse of dimensionality, with covariates X entering parametrically. In Hausman and Newey (1995) W is log income and log price, and X includes about 20 time and location dummies. X may be variable of interest and $g_0(Z)$ some covariates, e.g. sample selection.

Index Regression: $Z = (Y, X)$, $v(x, \beta)$ a known function,

$$E[Y|X] = \tau(v(X, \beta_0)),$$

where the function $\tau(\cdot)$ is unknown. Binary choice model has $E[Y|X] = \Pr(Y = 1|X) = \tau(v(X, \beta_0))$, with $\tau(\cdot)$. If allow conditional distribution of ε given X to depend (only) on $v(X, \beta_0)$, then binary choice model becomes index model.

Semiparametric Estimators

Estimators of β_0. Two kinds; do and do not require nonparametric estimation. Really model specific, but beyond scope to say why. One general kind of estimator:

$$\hat{\beta} = \arg\min_{\beta \in B} \sum_{i=1}^{n} q(Z_i, \beta)/n, \ \beta_0 = \arg\min_{\beta \in B} E[q(Z_i, \beta)],$$

B set of parmaeter values. Extremum estimator. Clever choices of $q(Z, \beta)$ in some semiparametric models.

Conditional median estimators use two facts:

Fact 1: The median of a monotonic transformation is transformation of the median.

Fact 2: The median minimizes the expected absloute deviation, i.e. $med(Y|X)$ minimizes $E[|Y - m(X)|]$ over functions $m(\cdot)$ of X.

Binary Choice: $v(x, \beta)$ include constant, ε has zero median, then $med(Y^*|X) = v(X, \beta_0)$. $1(y > 0)$ is montonic tranformation, Fact 1 implies $med(Y|X) = 1(med(Y^*|X) > 0) = 1(v(X, \beta_0) > 0)$. Fact 2, β_0 minimizes $E[|y - 1(v(x, \beta) > 0)|]$, so

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |Y - 1(v(X_i, \beta) > 0)|]$$

Maximum score estimator of Manski (1977). Only requires $med(Y^*|X) = v(X, \beta_0)$; allows for heteroskedasticity.

Censored Regression: $x'\beta$ includes a constant, ε has median zero, then $med(Y^*|X) = X'\beta_0$. $\max\{0, y\}$ is a monotonic transformation, so Fact 1 says $med(Y|X) = \max\{0, X'\beta_0\}$. By Fact 2, β_0 minimizes $E[|y - \max\{0, x'\beta\}|]$, so

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |Y_i - \max\{0, X_i'\beta\}|.$$

Censored least absolute deviations estimator of Powell (1984). Only requires $med(Y^*|X) = X'\beta$, allows for heteroskedasticity.

Generalize: $med(Y^*|X) = v(X, \beta_0)$ and $Y = T(Y^*)$ for monotonic transformation $T(y)$. By Fact 1 $med(Y|X) = T(v(X, \beta_0))$. Use Fact 2 to form

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |Y_i - T(v(X_i, \beta))|.$$

Global minimization, rather than solving first-order conditions, is important. For maximum score no first-order conditions. For censored LAD first-order conditions are zero whenever $x_i'\beta < 0$ for all $(i = 1, ..., n)$.

Approach provides estimates parameters and conditional median predictions, not conditional means. Generalizes to conditional quantiles.

Consistency and Asymptotic Normality of Minimization Estimators

A consistency result:

PROPOSITION 5: *If i)* $E[q(Z,\beta)]$ *has a unique minimum at* β_0*, ii)* $\beta_0 \in B$ *and* B *is compact; iii)* $q(Z,\beta)$ *is continuous at* β *with probability one; iv)* $E[\sup_{\beta\in B}|q(Z_i,\beta)|] < \infty$*; then*

$$\hat{\beta} = \arg\min_{\beta\in B} \sum_{i=1}^{n} q(Z_i,\beta) \xrightarrow{p} \beta_0.$$

Well known. Allows for $q(Z,\beta)$ to be discontinuous. Binary choice model above, assumption iii) satisfied if $v(x,\beta) = x'\beta$ and X_i includes continuously distributed regressor with corresponding component of β bounded away from zero on B. All the conditions are straightforward to check.

An asymptotic normality result, Van der Vaart (1995).

PROPOSITION 6: *If* $\hat{\beta} \xrightarrow{p} \beta_0$*,* β_0 *is in the interior of* B*, and i)* $E[q(Z_i,\beta)]$ *is twice differentiable at* β_0 *with nonsingular Hessian* H*; ii) there is* $d(z)$ *such that* $E[d(Z)^2]$ *exists and for all* $\beta,\tilde{\beta} \in B$*,* $|q(Z,\tilde{\beta}) - q(Z,\beta)| \leq d(Z)\|\tilde{\beta}-\beta\|$*; iii) with probability one* $q(Z,\beta)$ *is differentiable at* β_0 *with derivative* $m(Z)$*, then*

$$n^{1/2}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, H^{-1}E[m(Z)m(Z)']H^{-1}).$$

Straightforward to check for censored LAD. Do not hold for maximum score. Instead $n^{1/3}(\hat{\beta}-\beta_0) \xrightarrow{d}$.

Estimators with Nonparametric Components

Some models require use of nonparametric estimators. Include the partially linear and index regressions. We discuss least squares estimation when there is a nonparametric component in the regression. Basic idea is to "concentrate out" nonparametric component, to find a "profile" squared residual function, by substituting for nonparametric component an estimator.

Partially linear model as in Robinson (1988). Know $E[Y|X,W]$ minimizes $E[(Y - G(X,W))^2]$ over G, so that

$$(\beta_0, g_0(\cdot)) = \arg\min_{\beta, g(\cdot)} E[\{Y_i - X_i'\beta - g(W)\}^2].$$

Do minimization in two steps. First solve for minimum over g for fixed β, substituting that minimum into the objective function, then minimize β. The minimizer over g for fixed β is

$$E[Y_i - X_i'\beta|Z] = E[Y_i|Z_i] - E[X_i|Z_i]'\beta.$$

Substituting

$$\beta_0 = \arg\min_{\beta} E[\{Y_i - E[Y_i|Z_i] - (X_i - E[X_i|Z_i])'\beta\}^2].$$

Estimate using $\hat{E}[Y_i|Z_i]$ and $\hat{E}[X_i|Z_i]$ and replacing the outside expectation by a sample average,

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \{Y_i - \hat{E}[Y_i|Z_i] - (X_i - \hat{E}[X_i|Z_i])'\beta\}^2/n,$$

Least squares of $Y_i - \hat{E}[Y_i|Z_i]$ on $X_i - \hat{E}[X_i|Z_i]$. Kernel or series fine.

Index regression, as in Ichimura (1993). By $E[Y|X] = \tau_0(v(X, \beta_0))$,

$$(\beta_0, \tau_0(\cdot)) = \arg\min_{\beta, \tau(\cdot)} E[\{Y_i - \tau(v(X_i, \beta))\}^2].$$

Concentrating out the τ, $\tau(X, \beta) = E[Y|v(X, \beta)]$. Let $\hat{\tau}(X_i, \beta)$ a nonparametric estimator of $E[Y|v(X, \beta)]$, estimator is

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \{Y_i - \hat{\tau}(X_i, \beta)\}^2.$$

Generalize to log-likelihood and other objective functions. Let $q(z, \beta, \eta)$ depend on parametric component β and nonparametric component η. True values minimize $E[q(Z, \beta, \eta)]$, Estimator $\hat{\eta}(\beta)$ of $\eta(\beta) = \arg\min_{\eta} E[q(Z, \beta, \eta)]$,

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} q(Z_i, \beta, \hat{\eta}(\beta)).$$

Asymptotic theory difficult because of presence of nonparametric estimator. Know that often $n^{-1/2}$ rate, asymptotically normal, and even estimation of $\eta(\beta)$ does not affect asymptotic distribution.

Other estimators that depend on nonparametric estimators may have affect on limiting distribution, e.g. average derivative estimator of Stoker (1987) and Powell, Stock, and Stoker (1989).

Joint maximization possible but can be difficult because of need to smooth. Cannot allow $\hat{\eta}(\beta)$ to be n-dimensional.

An empirical example is Hausman and Newey (1995). Graphs are actually those for $\hat{g}(w)$ from a partially linear model.

Example of theory, series estimator of partially linear model. Let $p^K(w)$ be a $K \times 1$ vector of approximating functions, such as power series or splines. Also let

$$\begin{aligned} Y &= (Y_1, ..., Y_n)', X = [X_1, ..., X_n]', \\ P &= [p^K(W_1), ..., p^K(W_n)]', Q = P(P'P)^- P', \end{aligned}$$

Let $\hat{E}[Y_i|W_i] = p^K(W_i)'(P'P)^- P'Y$ and $\hat{E}[X_i'|W_i] = p^K(W_i)'(P'P)^- P'X$ be series estimators. Residuals are $(I - Q)Y$ and $(I - Q)X$, respectively, so by $I - Q$ idempotent,

$$\hat{\beta} = (X'(I - Q)X)^{-1} X'(I - Q)Y.$$

Here $\hat{\beta}$ is also least squares regression of Y on X and P. Result on $n^{1/2}$-consistency and asymptotic normality.

PROPOSITION 7: *If i) Y_i and X_i have finite second moments; ii) $H = E[Var(X_i|W_i)]$ is nonsingular; iii) $Var(Y_i|X_i, W_i)$ and $Var(X_i|W_i)$ are bounded; iv) there are C, γ_g,* and γ_x *such that for every K* there are α_K and β_K with $E[\{g_0(W_i) - \alpha_K p^K(W_i)\}^2] \leq K^{-\gamma_g}$ and $E[\|E[X|W_i] - \beta_K' p^K(W_i)\|^2] \leq CK^{-2\gamma_x}$ *v) $K/n \longrightarrow 0$ and $n^{1/2}K^{-(\gamma_g+\gamma_x)} \longrightarrow 0$, then for $\Sigma = E[Var(Y_i|X_i, W_i)\{X_i - E[X_i|W_i]\}\{X_i - E[X_i|W_i]\}']$,*

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, H^{-1}\Sigma H^{-1}).$$

Condition ii) is an identification condition that is essentially no perfect multicollinearity between X_i and any function of W_i. Intuitively, if one or more of the X_i variables were functions of W_i then we could not separately identify $g_0(W_i)$ and the coefficients on those variables. Furthermore, we know that necessary and sufficient conditions for identification of β from least squares objective function where $g(Z)$ has been partialled out are that $X_i - E[X_i|W_i]$ have a nonsingular second moment matrix. By iterated expectations that second moment matrix is

$$\begin{aligned} & E[(X_i - E[X_i|W_i])(X_i - E[X_i|W_i])'] \\ = \; & E[E[\{(X_i - E[X_i|W_i])(X_i - E[X_i|W_i])'\}|W_i]] = H. \end{aligned}$$

Thus, condition ii) is the same as the usual identification condition for least squares after partialling out the nonparametric component.

The requirement $K/n \longrightarrow 0$ is a small variance condition and $n^{1/2}K^{-(\gamma_g+\gamma_x)} \longrightarrow 0$ a small bias condition. The bias here is of order $K^{-(\gamma_g+\gamma_x)}$, which is of smaller order than just the bias in approximating $g_0(z)$ (which is only $K^{-\gamma_g}$). Indeed, the order of the bias of $\hat{\beta}$ is the product of the biases from approximating $g_0(z)$ and from approximating $E[x|z]$. So, one sufficient condition is that the bias in each of the nonparametric estimates vanish faster than $n^{-1/4}$. This faster than $n^{-1/4}$ condition is common to many semiparametric estimators.

Some amount of smoothness is required for root-n consistency. Existence of K satisfying the rate condition iii) requires that $\gamma_g + \gamma_x > r/2$. An analogous smoothness requirement (or even a stronger one) is generally needed for root-n consistency of any semiparametric estimator that requires estimation of a nonparametric component.

Proof of Propostion 7: For simplicity we give the proof when X_i is scalar. Let

$$\begin{aligned} M &= I - Q, \tilde{Z} = (W_1, ..., W_n)', \\ \bar{X} &= [E[X_1|W_1], ..., E[X_n|W_n]]', \\ V &= X - \bar{X}, g_0 = (g_0(z_1), ..., g_0(z_n))', \\ \varepsilon &= Y - X\beta_0 - g_0. \end{aligned}$$

Substituting $Y = X\beta_0 + g_0 + \varepsilon$ in the formula for $\hat{\beta}$, subtracting β_0, and multiplying by $n^{1/2}$ gives

$$\begin{aligned} & n^{1/2}(\hat{\beta} - \beta_0) \\ = \; & (X'MX/n)^{-1}(X'Mg_0/n^{1/2} + X'M\varepsilon/n^{1/2}) \end{aligned}$$

By the law of large numbers $V'V/n \xrightarrow{p} H$. Also, similarly to the proof of Proposition 3,

$$\begin{aligned} V'QV/n &= O_p(K/n), \bar{X}'M\bar{X}/n = O_p(K^{-2\gamma_x}), \\ g_0'Mg_0/n &= O_p(K^{-2\gamma_g}). \end{aligned}$$

Then $V'MV/n \xrightarrow{p} H$ and

$$\begin{aligned} & |V'M\bar{X}/n| \\ \leq \; & (V'MV/n)^{1/2}(\bar{X}'M\bar{X}/n)^{1/2} \xrightarrow{p} 0, \end{aligned}$$

so that

$$\begin{aligned} X'MX/n &= (\bar{X} + V)'M(\bar{X} + V)/n \\ &= \bar{X}'M\bar{X}/n \\ &\quad +2\bar{X}'MV/n + V'MV/n \xrightarrow{p} H. \end{aligned}$$

Next, similarly to the proof of Proposition 3 we have $E[VV'|W] \leq CI_n$ and $E[\varepsilon\varepsilon'|W, X] \leq CI_n$. Then

$$\begin{aligned} E[\{V'Mg_0/n^{1/2}\}^2] &= E[g_0'ME[VV'|W]Mg_0]/n \\ &\leq CE[g_0'Mg_0]/n \\ &= O(K^{-2\gamma_g}) \longrightarrow 0. \\ & |\bar{X}'Mg/n^{1/2}|^2 \\ &\leq n(\bar{X}'M\bar{X}/n)(g_0'Mg_0/n) \\ &= O_p(nK^{-2\gamma_x - 2\gamma_g}) \xrightarrow{p} 0, \end{aligned}$$

so that $X'Mg_0/n^{1/2} = \bar{X}'Mg_0/n^{1/2} + V'Mg_0/n^{1/2} \xrightarrow{p} 0$. Also,

$$\begin{aligned} E[\{\bar{X}'M\varepsilon/n^{1/2}\}^2] &= E[\bar{X}'ME[\varepsilon\varepsilon'|X,W]M\bar{X}]/n \\ &\leq CE[\bar{X}'M\bar{X}]/n \\ &= O(K^{-2\gamma_x}) \longrightarrow 0, \\ E[\{V'Q\varepsilon/n^{1/2}\}^2] &= E[V'QE[\varepsilon\varepsilon'|X,W]QV]/n \\ &\leq CE[V'QV]/n \\ &= O(K/n) \longrightarrow 0, \end{aligned}$$

and by the central limit theorem, $V'\varepsilon/n^{1/2} \xrightarrow{d} N(0,\Sigma)$. Therefore,

$$\begin{aligned} X'M\varepsilon/n^{1/2} &= \bar{X}'M\varepsilon/n^{1/2} \\ &\quad +V'\varepsilon/n^{1/2} - V'Q\varepsilon/n^{1/2} \xrightarrow{d} N(0,\Sigma). \end{aligned}$$

Then by the continuous mapping and Slutzky theorems it follows that

$$\begin{aligned} n^{1/2}(\hat{\beta} - \beta_0) &= (H + o_p(1))^{-1}[V'\varepsilon/n^{1/2} + o_p(1)]] \\ \xrightarrow{d} H^{-1}N(0,\Sigma) &= N(0, H^{-1}\Sigma H^{-1}). \end{aligned}$$