Nonparametric Regression

Whitney Newey

October 2007

Nonparametric Regression

Often interested in a regression function.

Linear model:

$$E[Y|X] = X'\beta_0.$$

Nonparametric version allows for unknown functional form:

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

Note that for $\varepsilon = Y - g_0(X)$, which satisfies $E[\varepsilon|X] = 0$ by construction, we have

$$Y = g_0(X) + \varepsilon.$$

Generalizes the linear regression model $Y = X'\beta_0 + \varepsilon$ to allow for unknown function.

Important to allow for nonlinearity in estimating structural effects.

Linear model as approximation:

$$E[(Y - X'\beta)^2] = E[(\varepsilon + g_0(X) - X'\beta)^2] = E[\varepsilon^2] + E[\{g_0(X) - X'\beta\}^2].$$

Thus OLS has plim equal to $\arg\min_\beta E[\{g_0(X) - X'\beta\}^2]$.

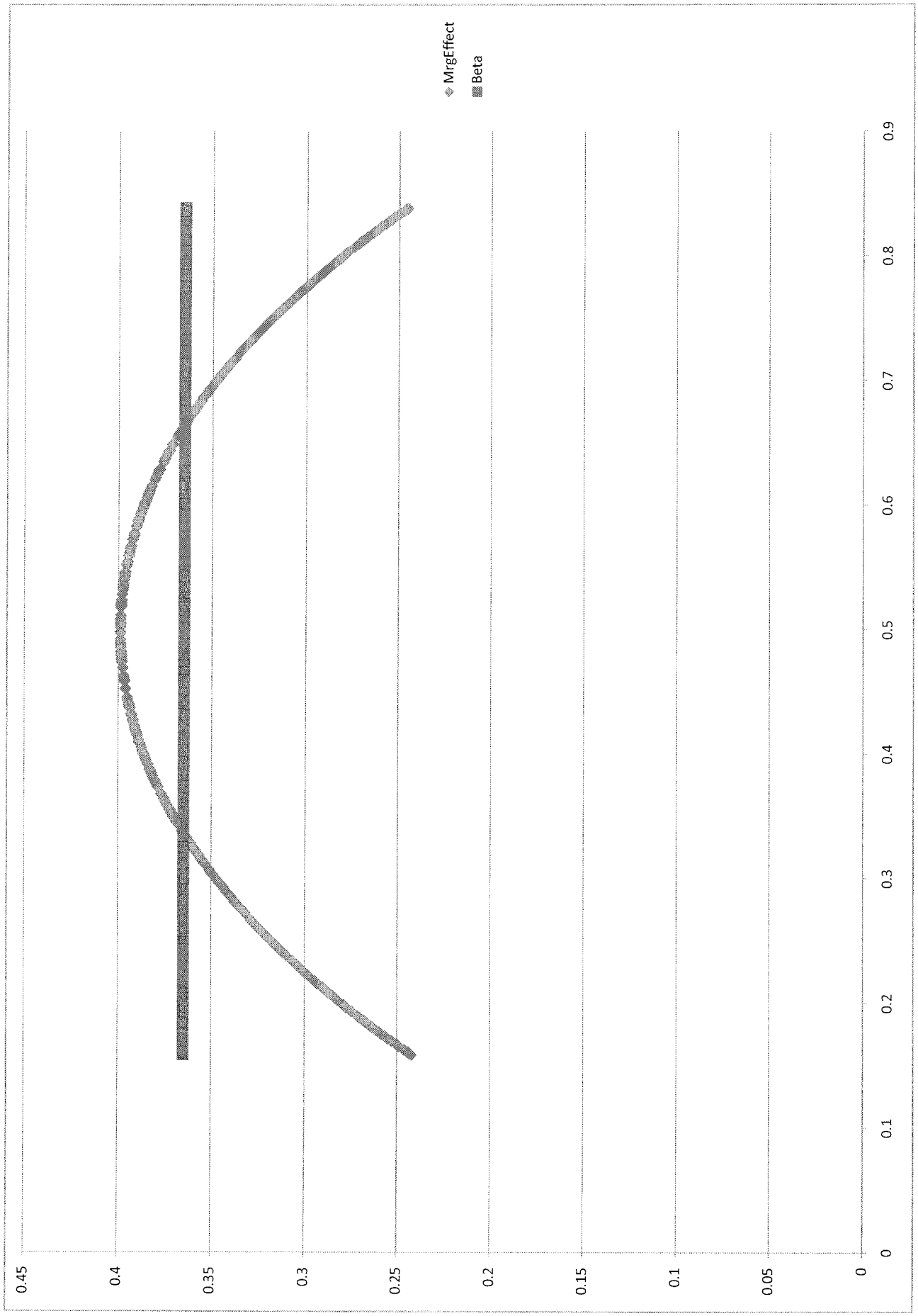
OLS is minimum MSE approximation to true, nonlinear regression $g_0(X)$.

In general OLS is NOT a weighted average of the derivative $\partial g_0(X)/\partial X$.

OLS coefficients depend on distribution of X, so do not give effect of changing X.

Ex: $Y \in \{0, 1\}$, $\Pr(Y = 1|X) = \Phi(\beta_1 + \beta_2 X)$, X is $U(0, 1)$, $\beta_2 = 1$

Plot true marginal effect $\partial \Pr(Y = 1|X)/\partial X$ and least squares approximation of slope.



Kernel Regression

One approach is to plug kernel density estimator into the formula

$$g_0(x) = \int yf(y|x)dy = \int yf(y,x)dy / \int f(y,x)dy$$

For scalar X and kernel $K(u)$,

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

$\sum_{i=1}^{n} w_i^h(x) = 1$, $w_i^h(x) \geq 0$ (for $K(u) \geq 0$).

Assume symmetric $K(u)$ with a unique mode at $u = 0$.

More weight for X_i closer to x.

Bandwidth h controls how fast the weights decline.

For smaller h, more weight given closer observations.

Reduces bias but increases variance.

For multivariate X, formula same, with multivariate kernel, such as

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

Kernel regression has big biases near boundaries.

Better approach is locally linear regression.

Locally Linear Regression

Let $K_h(u) = h^{-r}K(u/h)$ as before.

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

Interpretation: $\hat{g}_h(x)$ is constant in a weighted least squares of Y_i on $(1, x - X_i)$, with weights $K_h(x - X_i)$. Let e_1 a $(r+1) \times 1$ vector with 1 in first position and zeros elsewhere, and

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

Then

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

Locally linear fit of the data.

Predicted value from weighted least squares at x.

Weights smaller for observations farther from x.

Kernel regression is locally constant.

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

$\hat{\beta}(x) = \widehat{\partial g_0(x)/\partial x}$.

$\hat{g}_h(x)$ is weighted average of Y_i: Define

$$\begin{aligned}
\hat{S}_0(x) &= \sum_{i=1}^n K_h(x - X_i)/n,\ \hat{S}_1(x) = \sum_{i=1}^n K_h(x - X_i)(\frac{x - X_i}{h})/n, \\
\hat{S}_2(x) &= \sum_{i=1}^n K_h(x - X_i)(\frac{x - X_i}{h})(\frac{x - X_i}{h})'/n \\
\hat{m}_0(x) &= \sum_{i=1}^n K_h(x - X_i)Y_i/n, \hat{m}_1(x) = \sum_{i=1}^n K_h(x - X_i)(\frac{x - X_i}{h})Y_i/n.
\end{aligned}$$

Then

$$\hat{g}_h(x) = e_1' \begin{bmatrix} \hat{S}_0(x) & h\hat{S}_1(x)' \\ h\hat{S}_1(x) & h^2\hat{S}_2(x) \end{bmatrix}^{-1} \begin{pmatrix} \hat{m}_0(x) \\ h\hat{m}_1(x) \end{pmatrix}$$

$$= [\hat{S}_0(x) - \hat{S}_1(x)'\hat{S}_2(x)^{-1}\hat{S}_1(x)]^{-1}[\hat{m}_0(x) - \hat{S}_1(x)'\hat{S}_2(x)^{-1}\hat{m}_1(x)]$$

$$= \frac{\sum_{i=1}^n \hat{a}_i(x)Y_i}{\sum_{i=1}^n \hat{a}_i(x)},$$

$$\hat{a}_i(x) = K_h(x - X_i)[1 - \hat{S}_1(x)'\hat{S}_2(x)^{-1}(\frac{x - X_i}{h})]/n.$$

Asymptotic bias and variance of locally linear estimators for scalar x.

Bias: For $g_0 = (g_0(X_1), ..., g_0(X_n))'$,

$$E[\hat{g}_h(x)|X] = \frac{\sum_{i=1}^n \hat{a}_i(x) g_0(X_i)}{\sum_{i=1}^n \hat{a}_i(x)}.$$

Expand $g_0(X_i)$ around $g_0(x)$ to get

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

where $\bar{X}_i$ is between x and X_i. Note that

$$\sum_{i=1}^n \hat{a}_i(x)(X_i - x)' = -h\hat{S}_1(x)' + \hat{S}_1(x)'\hat{S}_2(x)^{-1} h \hat{S}_2(x) = 0.$$

Then substituting in the expansion we get

$$\begin{aligned} E[\hat{g}_h(x)|X] &= g_0(x) + \frac{h^2}{2} g_0''(x)\hat{C}_2 + \frac{h^3}{6}\hat{C}_3 \\ \hat{C}_2 &= \frac{\sum_{i=1}^n \hat{a}_i(x)(\frac{x - X_i}{h})^2}{\sum_{i=1}^n \hat{a}_i(x)} = \frac{\hat{S}_2(x) - \hat{S}_1(x)\hat{S}_2(x)^{-1}\hat{S}_3(x)}{\hat{S}_0(x) - \hat{S}_1(x)\hat{S}_2(x)^{-1}\hat{S}_1(x)}, \\ \hat{C}_3 &= \frac{\sum_{i=1}^n \hat{a}_i(x) g'''(\bar{X}_i)(\frac{x - X_i}{h})^3}{\sum_{i=1}^n \hat{a}_i(x)} \end{aligned}$$

Need limit of $\hat{C}_2$ and $\hat{C}_3$ to get asymptotic bias expression

Let $\mu_j = \int K(u)u^j du$ and $u = (x - X_i)/h$.

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

for $h \longrightarrow 0$. Also, for $nh \longrightarrow \infty$.

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

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

$$\hat{S}_j(x) = \mu_j f_0(x) + o_p(1).$$

$$\hat{S}_j(x) = \mu_j f_0(x) + o_p(1).$$

Assume $\mu_0 = \int K(u)du = 1$ and $\mu_1 = \int K(u)udu = 0$. Then the continuous mapping theorem gives

$$\hat{C}_2 = \frac{(\mu_2 - \mu_1\mu_3/\mu_2)f_0(x)}{\left(\mu_0 - \mu_1^2/\mu_2\right) f_0(x)} = \mu_2 + o_p(1).$$

Assuming $g_0'''(x)$ is bounded it can be shown similarly that $\hat{C}_3 = O_p(1)$. Then

$$E[\hat{g}_h(x)|X] = g_0(x) + \frac{h^2}{2} g_0''(x)\mu_2 + O_p(h^3)$$

Next, note that

$$Var(\hat{g}_h(x)|X) = \frac{\sum_{i=1}^n \hat{a}_i(x)^2 \sigma(X_i)}{\left[\sum_{i=1}^n \hat{a}_i(x)\right]^2}$$

Let $\nu_j = \int K(u)^2 u^j du$. By a similar argument to those above,

$$\begin{aligned} hE[K_h(x - X_i)^2 (\frac{x - X_i}{h})^j \sigma^2(X_i)] &= h \int K_h(x-t)^2 (\frac{x-t}{h})^j \sigma^2(t) f_0(t) dt \\ &= \int K(u)^2 u^j \sigma^2(x - hu) f_0(x - hu) du \\ &= v_j \sigma^2(x) f_0(x) + o(1). \end{aligned}$$

Then we will have

$$\frac{h}{n} \sum_{i=1}^n K_h(x - X_i)(\frac{x - X_i}{h})^j \sigma^2(X_i) = v_j \sigma^2(x) f_0(x) + o_p(1)$$

$$\frac{h}{n}\sum_{i=1}^{n} K_h(x-X_i)(\frac{x-X_i}{h})^j\sigma^2(X_i) = v_j\sigma^2(x)f_0(x) + o_p(1)$$

Then by $\hat{S}_1(x) \xrightarrow{p} \mu_1 f_0(x) = 0$,

$$\begin{aligned}\sum_{i=1}^{n}\hat{a}_i(x)^2\sigma(X_i) &= \hat{a}_i(x) = \frac{1}{nh}\sum_{i=1}^{n}\{hK_h(x-X_i)^2[1-\hat{S}_1(x)'\hat{S}_2(x)^{-1}(\frac{x-X_i}{h})]^2\sigma \\ &= \frac{1}{nh}\{\sum_{i=1}^{n} hK_h(x-X_i)^2\sigma(X_i)/n + o_p(1)\} \\ &= \frac{1}{nh}v_0\sigma^2(x)f_0(x) + o_p(\frac{1}{nh}).\end{aligned}$$

Then by $\sum_{i=1}^{n}\hat{a}_i(x) \xrightarrow{p} f_0(x)$ we have

$$Var(\hat{g}_h(x)|X) = \frac{1}{nh}\frac{v_0\sigma^2(x)}{f_0(x)} + o_p(\frac{1}{nh})$$

Summarizing we have

$$E[\hat{g}_h(x)|X] = g_0(x) + \frac{h^2}{2} g_0''(x)\mu_2 + o_p(h^2)$$

$$Var(\hat{g}_h(x)|X) = \frac{1}{nh}\frac{v_0\sigma^2(x)}{f_0(x)} + o_p(\frac{1}{nh})$$

Combining these results we have the MSE

$$\begin{aligned} E[\{\hat{g}_h(x) - g_0(x)\}^2|X] &= \frac{1}{nh}\frac{v_0\sigma^2(x)}{f_0(x)} + \frac{h^4}{4} g_0''(x)^2\mu_2^2 \\ &\quad + o_p(\frac{1}{nh} + h^4). \end{aligned}$$

In contrast, kernel regression MSE is

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

Kernel regression has boundary bias.

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.

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

Series Regression

Nonparametric estimator is predicted value from regressing on approximating functions.

Examples: Power series, regression splines.

Let $p^K(x) = (p_{1K}(x), ..., p_{KK}(x))'$, $Y = (Y_1, ..., Y_n)'$ and $P = [P^K(X_1), ..., P^K(X_n$
Then

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

A^- denotes generalized inverse, $AA^-A = A$.

Different than locally linear; global fit rather than local average.

Examples: x scalar

Power series:

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

$p^K(x)'\beta$ is a polynomial. Good global approximation of smooth functions. Not good when jump or kink; kink at one point makes whole approximation bad. Applications need orthogonal polynomials because of severe multicollinearity. .

Regression splines:

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

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 rate for approximating $g_0(x)$ by $p^K(x)'\beta$.,

Assume there exists $\gamma > 0$ and C such for each K there is $\bar{\beta}_K$ with

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

Here $K^{-\gamma}$ is root MSE rate for series approximation.

Comes from approximation theory.

For power series, X_i in a compact subset or $\Re^r$, $g_0(x)$ is continuously differentiable of order s, then $\gamma = s/r$.

For splines, $\gamma = \min\{4, s\}/r$.

For (X_i, Y_i) i.i.d. and $Var(Y_i|X_i) \leq \Delta$,

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

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

To show this let

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

Note Q idempotent, so $I - Q$ idempotent, hence $I - Q$ 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 $i \neq j$,

$$E[\varepsilon_i \varepsilon_j | X] = E[\varepsilon_i \varepsilon_j | X_i, X_j] = E[\varepsilon_i E[\varepsilon_j | X_i, X_j, \varepsilon_i] | X_i, X_j] = E[\varepsilon_i E[\varepsilon_j | X_j] | X_i, X_j]$$

Then for $\sigma_i^2 = Var(Y_i | X_i)$ and $\Sigma = \mathsf{diag}(\sigma_1^2, ..., \sigma_n^2)$ we have $E[\varepsilon\varepsilon' | X] = \Sigma$. It follows that for $tr(A)$ the trace of a square matrix A, by $rank(Q) \leq K$,

$$E\left[\varepsilon' Q \varepsilon | X\right] = \mathsf{tr}\left(QE\left[\varepsilon\varepsilon' | X\right]\right) = \mathsf{tr}(Q\Sigma) = \mathsf{tr}(Q\Sigma Q) \leq \Delta\,\mathsf{tr}(Q) \leq \Delta K.$$

Summarizing:

$$E\left[(g_0-\bar{g})\left(I-Q\right)(g_0-\bar{g})\right] \leq E\left[(g_0-\bar{g})'\left(g_0-\bar{g}\right)\right] \leq CnK^{-2\gamma}.E\left[\varepsilon' Q\varepsilon|X\right] \leq \Delta$$

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

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

Then by i.i.d. observations,

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

Choosing Bandwidth or Number of Terms

Data based choice operationalizes complete flexibility.

For series estimator, number of terms varies with data set.

Minimizing the sum of squared residuals does not lead to optimal choice of K or h. Why?

Explain choice for series estimator with $Var(Y_i|X_i) = \sigma^2$, $P'P$ nonsingular.

$$E\left[\frac{\sum_{i=1}^n \{\hat{g}(X_i) - g_0(X_i)\}^2}{n}|X\right] = \frac{E[\varepsilon' Q \varepsilon|X]}{n} + \frac{g_0{}'(I-Q)g_0}{n}$$
$$= \sigma^2 \frac{K}{n} + \frac{g_0'(I-Q)g_0}{n}.$$

To see problem with sum of squared residuals, since $Y = g_0 + \varepsilon$,

$$(Y - \hat{g})'(Y - \hat{g}) = Y'(I - Q)Y = \varepsilon'(I - Q)\varepsilon + 2\varepsilon'(I - Q)g_0 + g_0'(I - Q)g_0.$$

Taking expectations,

$$E\left[\frac{(Y - \hat{g})'(Y - \hat{g})}{n}\right] = \sigma^2 - \sigma^2\frac{K}{n} + g_0'(I - Q)g_0.$$

Expected sum of square residuals subtracts rather than adds $\sigma^2 K/n$. Fix?

Add penalty for number of terms. Let $\hat{\sigma}^2$ be some estimate of the variance of disturbances, like $\hat{\sigma}^2 = (Y - \hat{g})'(Y - \hat{g})/n$ for some fixed $\bar{K}$. Mallows criteria is to choose K to minimize.

$$\hat{M}(K) = \frac{(Y - \hat{g})'(Y - \hat{g})}{n} + 2\hat{\sigma}^2 \frac{K}{n}.$$

Estimates MSE up to constant. Replace $\hat{\sigma}^2$ by σ^2 (asymptotically valid):

$$\begin{aligned} E\left[\frac{(Y - \hat{g})'(Y - \hat{g})}{n} + 2\sigma^2 \frac{K}{n}\right] &= \sigma^2 - \sigma^2 \frac{K}{n} + g_0'(I - Q)g_0 + 2\sigma^2 \frac{K}{n} \\ &= \sigma^2 + \sigma^2 \frac{K}{n} + g_0'(I - Q)g_0. \end{aligned}$$

Equal to MSE plus σ^2.

Mallows criteria $\hat{M}(K)$. Imposes homoskedasticity. Cross-validation works with heteroskedasticity.

Let $\hat{g}_{-i,K}(X_i)$ predicted value for i^{th} observation using all the otherobservations, for kernel and series respectively.

$$\hat{g}_{-i,K}(X_i) = Y_i - \frac{Y_i - \hat{g}_K(X_i)}{1 - Q_{ii}}.$$

Criteria is

$$C\hat{V}(K) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{g}_{-i,K}(X_i)]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{g}(X_i)}{1 - Q_{ii}}\right]^2.$$

Choosing K to minimize $C\hat{V}(K)$ is asymptotically optimal:

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

Locally linear version:

$$\begin{aligned}
\hat{S}_{-i,0}(x) &= \frac{1}{n-1}\sum_{j\neq i} K_h(x-X_j),\ \hat{S}_{-i,1}(x) = \frac{1}{n-1}\sum_{j\neq i} K_h(x-X_j)(\frac{x-X_j}{h}) \\
\hat{S}_{-i,2}(x) &= \frac{1}{n-1}\sum_{j\neq i}^{n} K_h(x-X_j)(\frac{x-X_j}{h})(\frac{x-X_j}{h})', \\
\hat{a}_{-i,j}(x) &= K_h(x-X_j)\left[1-\hat{S}_{-i,1}(x)'S_{-i,2}(x)^{-1}(\frac{x-X_j}{h})\right]
\end{aligned}$$

The a locally linear estimator of $g_0(X_i)$ that uses all the observations except the i^{th} is

$$\hat{g}_{-i,h}(x) = \frac{\sum_{j\neq i}\hat{a}_{-i,j}(x)Y_j}{\sum_{j\neq i}\hat{a}_{-i,j}(x)}.$$

The cross-validation criteria is

$$\widehat{CV}(h) = \frac{\sum_{i=1}^{n}\left[Y_i-\hat{g}_{-i,h}(x)\right]^2}{n}.$$

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

The Curse of Dimensionality in Nonparametric Regression

The convergence rate for series estimators can be obtained by setting $K/n = K^{-2\gamma}$. The otpiml rate of growth of K is $n^{1/(1+2\gamma)}$. Plug that in take square roots to get

$$\sqrt{E[\{\hat{g}(X_i) - g_0(X_i)\}^2]} \leq \bar{C} n^{-\frac{\gamma}{1+2\gamma}}.$$

For power seris or splines with s low, $\gamma = s/r$, so that

$$\sqrt{E[\{\hat{g}(X_i) - g_0(X_i)\}^2]} \leq \bar{C} n^{-\frac{s}{r+2s}}$$

As r, the dimension of X, optimal convergence rate declines.

Rate increases as s grows, but does not help much in practice.

Curse of dimensionality is not so bad in restricted models.

Additive model: When $X = (x_1, ..., x_r)'$,

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

Here one dimenstional rate $n^{-\frac{s}{1+2s}}$ is attainable.

Series estimator is simple. Restrict approximating functions to depend on only one component.

Scalar u and $p_{\ell L}(u), \ell = 1, ..., L$ approximating functions,

$p^K(x) = (1, p^L(x_1)', ..., p^L(x_r)')', p^L(u) = (p_{1L}(u), ..., p_{LL}(u))', K = Lr+1.$

Additivity in log price and log income holds in Hausman Newey (1995) gasoline demand application.