Lecture 16
"Non-Parametric Kernel Regression"

Consider a general regression model

$$y_i = g(x_i) + u_i$$

where $g(x) = E(Y|X = x)$, in order to estimate g we may consider

$$E(Y|X = x) = \int y \frac{f(x,y)}{f(x)} dy$$

Now suppose we use Kernel density estimation to estimate both $f(x,y)$ and $f(x)$. This is the basic idea of Nadaraya-Watson estimator. In the numerator we have

$$\hat{f}(x,y) = n^{-1} \sum K_{h_1}(x - X_i) K_{h_2}(y - Y_i)$$

then

$$\begin{aligned}
\int y \hat{f}(x,y) dy &= n^{-1} \sum \int K_{h_1}(x - X_i) y K_{h_2}(y - Y_i) dy \\
&= n^{-1} \sum K_{h_1}(x - X_i) \int \frac{y}{h_2} K\left(\frac{y - Y_i}{h_2}\right) dy \\
&= n^{-1} \sum K_{h_1}(x - X_i) \int (s h_2 + Y_i) K(s) ds \\
&= n^{-1} \sum K_{h_1}(x - X_i) Y_i
\end{aligned}$$

Since $\int sK(s)ds = 0$ and $\int K(s)ds = 1$. Thus we have

$$\begin{aligned}
\hat{g}_h(x) &= n^{-1} \frac{\sum K_h(x - X_i) Y_i}{n^{-1} \sum K_h(x - X_i)} \\
&= \sum w_{h_1}(x) Y_i \\
\text{where} \quad w_{h_1}(x) &= \frac{(nh)^{-1} K((x - X_i)/h)}{\hat{f}_h(x)}.
\end{aligned}$$

So at x the estimate of $E(Y|X = x)$ is a weighted average of the Y_i "near x".

Performance of $\hat{g}$

Write $\hat{g}_h(x) = \hat{r}_h(x)/\hat{f}_h(x)$ and consider the numerator,

$$\begin{aligned}
E\hat{r}_h(x) &= E n^{-1} \sum K_h(x - X_i) Y_i \\
&= E K_h(x - X_i) Y_i \\
&= \int \int y K_h(x - u) f(y|u) f(u) dy du \\
&= \int K_h(x - u) f(u) (\int y f(y|u) dy) du \\
&= \int K_h(x - u) f(u) g(u) du \\
&= \int K_h(x - u) r(u) du
\end{aligned}$$

where we define $r(u) = f(u)g(u) = \int yf(y,u)dy$. Expanding, as in the kernel density case we have,

$$E\hat{r}_h(x) = r(x) + \frac{h^2}{2}r''(x)\mu_2(K) + o(h^2)$$

and whre the linear term in h vanishes due to the mean zero assumption on K just as in density estimation,

$$\begin{aligned} V\hat{r}_h(x) &= V(n^{-1}\sum K_h(x-X_i)Y_i) \\ &= n^{-1}V(K_h(x-X_i)Y_i) \\ &= n^{-1}\{\int K_h^2(x-u)\sigma^2(u)f(u)du - (\int K_h(x-u)r(u)du)^2\} \\ &= (nh)^{-1}\int K_h^2(u)\sigma^2(x+vh)f(x+vh)dv + o((nh)^{-1}) \\ &= (nh)^{-1}f(x)\sigma^2(x) \parallel K \parallel_2^2 + o((nh)^{-1}) \end{aligned}$$

where $\sigma(x) \equiv EY^2|X = x$. So,

$$MSE(\hat{r}_h(x)) = (nh)^{-1}f(x)\sigma^2(x) \parallel K \parallel_2^2 + \frac{h^4}{4}(r''(x)\mu_2(K))^2 + o(h^4) + o((nh)^{-1})$$

and we obtain the following result by Slutsky.
Theorem: $\hat{g}_h(x) \to g(x)$ if $h \to 0$ and $(nh) \to \infty$

Note these are the same conditions for convergence in probability for $\hat{f}_h(x)$ so Slutsky really applies here.

Bandwidth Selection

This is a large topic. Here is the simplest version of the theory:

$$\begin{aligned} \hat{g}_h(x) - g(x) &= \left(\frac{\hat{r}_h(x)}{\hat{f}_h(x)} - g(x)\right)\left(\frac{\hat{f}}{f} + \left(1 - \frac{\hat{f}}{f}\right)\right) \\ &= \frac{\hat{r} - g\hat{f}}{f} + (\hat{g} - g)\left(\frac{f - \hat{f}}{f}\right) \\ &= O_p(n^{-2/5}) + o_p(1)O_p(n^{-2/5}) = O_p(n^{-2/5}) \end{aligned}$$

where we have let $h = O(n^{-1/5})$ to compute the orders in the last line. So we can focus on the first term

$$\begin{aligned} f(x)^{-2}E(\hat{g} - g\hat{f})^2 &= (nhf)^{-2}E\sum\left[K\left(\frac{x-X_i}{h}\right)(Y_i - g)\right]^2 \\ &= \frac{1}{nh^2f^2}V\left(K\left(\frac{x-X}{h}\right)(Y-g)\right) \\ &\quad + \frac{1}{h^2f^2}E^2\left[K\left(\frac{x-X}{h}\right)(Y-g)\right] \end{aligned}$$

and this yields,

$$MSE(\hat{g}_h(x)) = (nh)^{-1}\frac{\sigma^2(x)}{f(x)} \parallel K \parallel_2^2 + \frac{h^4}{4}\mu_2^2(K)\left(g''(x) + 2\frac{g'(x)f'(x)}{f(x)}\right)^2 + o((nh)^{-1}) + o(h^4)$$

so

$$h = O(n^{-1/5}) \Rightarrow MSE(x) = O(n^{-4/5}).$$

Note that the first term in the bias involves the curvature and the second term is a first derivative effect which may dominate at (or near) inflection points.

Practicalities

There are many details we have neglected here, for example, the preceding formulas suggest that we should do adaptive bandwidth selection using larger bandwidth where the function g is smooth or the variability is large and smaller bandwidth when g is rough. In `R` kernel regression is implemented in the function `ksmooth( )`. See Härdle (1989, 1990) for many further details of theory and implementation.

Locally Polynomial Regression

A variation on the kernel regression ideas introduced above is locally polynomial kernel regression. Ordinary Nadaraya-Watson kernel regression may be viewed as a special, locally constant, case. A good reference on this in the econometric literature is Cleveland, Devlin and Gross (1988). A good reference in the statistics literature is Hastie and Loader, (1993). More complete treatments are provided by Fan and Gijbels (1996) and Ruppert, Wand and Carroll (2004).

The idea is very simple – we replace locally weighted means by locally weighted *regressions*.

Let $w_k(x) = W((x_k - x)/h)$ where $W(\ \ )$ plays the role of the kernel. A favorite kernel in [1] is

$$W(x) = \begin{cases} (1 - |x|^3)^3 & |x| \leq 1 \\ 0 & |x| > 1 \end{cases}$$

Now for any x consider the problem

$$\min_{b \in \Re^{p+1}} \sum w_k(x)(y_k - \beta_o - \beta_1(x - x_k) - \ldots - \beta_p(x - x_k)^p)^2$$

If we let

$$z_k = (1, (x - x_k), \ldots, (p!)^{-1}(x - x_k)^p)$$

we can write the model as

$$y_i = z_i'\beta(x) + u_i$$

and the problem is obviously just a garden variety WLS problem. Much of the standard LS theory carries over in a nice way. But there are some slightly unsettling features as well. For example, we may write the estimator as a linear operator, with $L = (Z'WZ)^{-1}Z'W$

$$\hat{g} = Ly$$

so

$$\hat{u} = (I - L)y$$

and

$$\hat{u}'\hat{u} = \sum \hat{u}_i^2 = u'(I - L)'(I - L)u$$

So far this is familiar and comforting until we realize that neither L nor $I - L$ are symmetric or idempotent as in the usual LS case. So $\hat{g}$ isn't really a projection. However, we plunge ahead

bravely anyway. Since

$$\begin{aligned} E\hat{u}'\hat{u} &= Ey'(I-L)'(I-L)y \\ &= E(g+u)'(I-L)'(I-L)(g+u) \\ &= g'(I-L)'(I-L)g + Eu'(I-L)'(I-L)u \\ &= g'(I-L)'(I-L)g + \sigma^2 \text{ Trace } [(I-L)(I-L)'] \end{aligned}$$

we have for any given choice of L (which obviously represents implicitly a degree of smoothing) a method of estimating σ^2. We will assume that the first term representing squared bias is negligible; note that it would be identically zero if we were in a correctly specified regression setting in which case g would be in the space spanned by L. (At this point you should be asking yourself; why are we assuming homoscedasticity? The answer is we want to try to understand crawling before we try to walk.) The trace term can be interpreted very much in the same way as in ordinary regression as the "effective degrees of freedom" of the fitted model, $\hat{g}$. Recall that in the usual case $L = P_X$ and so this trace is $n-q$, where obviously q depends on the bandwidth of the kernel. We will follow the lead of many researchers and call q the effective dimension of the fitted model. Oddly, since $I - L$ isn't idempotent q needn't be an integer. Nevertheless it is a number between p, the order of the locally polynomial model and ∞, since as bandwidth increases without bound the solution approaches the global polynomial model. On the other hand, as bandwidth shrinks the fitted function tends to interpolate all the observed points eventually leading to a hopelessly rough $\hat{g}$. In the next lecture we will relate these quantities to the standard AIC and BIC criteria and explain more about how they can be used and abused.

The CDG strategy for choosing a degree of smoothing is based on a variation of what is usually called "nearest neighbors." They suggest choosing h_i such that $r = [\phi n]$ and

$$r = \#\{k \mid |x_k - x_i| < h_i\}$$

This insures that the bandwidth at x_i contains approximately the same fraction of points, ϕ, for all i. They also suggest $\phi \approx 2/3$.

References

Green, P. and B. Silverman (1994). *NP Regression and GLM's*, Chapman-Hall

Härdle, W. (1989). *Applied NP-Regression*, Cambridge.

Härdle, W. (1990). *Smoothing Techniques*, Springer-Verlag.

Fan, J. and I. Gijbels, (1996). *Local Polynomial Fitting and Its Applications*, Chapman-Hall

Cleveland, W., S. Devlin and A. Grosse (1988). Regression by local fitting: Methods, properties and computational algorithms, *J. of Econometrics*, 37, 87-114.

Hastie T. and C. Loader (1993). Local regression, *Stat. Science*, 8, 120-129.

Ruppert, D., M. Wand and R. Carroll (2004), *Semiparametric Regression*, Cambridge, U. Press.