Economics 620, Lecture 19: Introduction to Nonparametric and Semiparametric Estimation

Nicholas M. Kiefer

Cornell University
Good when there are lots of data and very little prior information on functional form.

Examples:

\[ y = f(x) + \varepsilon \text{ (nonparametric)} \]
\[ y = z'\beta + f(x) + \varepsilon \text{ (partial linear)} \]
\[ y = f(z'\beta) + \varepsilon \text{ (index model)} \]

Have to have some restrictions on \( f \) to avoid a perfect fit.

Differentiability to some order, and bounded derivatives.
Assume the errors are iid and arrange the observations in order of the $x_i$.

Consider $y = f(x) + \varepsilon$.

Moving average estimator:
\[ \hat{f}(x_i) = k^{-1} \sum y_j \text{ for } k \text{ values of } j \text{ centered on } i.\]
Let $k$ increase with the sample size, but more slowly than $n$.

\[
\begin{align*}
\hat{f}(x_i) &= k^{-1} \sum y_j = k^{-1} \sum f(x_j) + k^{-1} \sum \varepsilon_j \\
&= f(x_i) + f'(x_i) k^{-1} \sum (x_j - x_i) \\
&\quad + \frac{1}{2} k^{-1} f''(x_i) \sum (x_j - x_i)^2 + k^{-1} \sum \varepsilon_j
\end{align*}
\]

$f'$ is multiplied by zero if $x$ is symmetric around $x_i$. 
\[ \hat{f}(x_i) = f(x_i) + 24^{-1}(k/n)^2 f'' + k^{-1} \sum \varepsilon_j \text{ approximately. Hence} \]

\[ \hat{f}(x_i) = f(x_i) + O((k/n)^2) + O_p(k^{-1/2}). \]

Note these “errors” are bias and variance

\[ (\hat{f}(x_i) - f(x_i))^2 = O((k/n)^4) + O_p(k^{-1}). \]

Consistent if \( k \to \infty \) and \( k/n \to 0 \).

“Best” trades off bias and variance at the same rate: \( (k/n)^4 \) looks like \( k^{-1} \). Or \( k = O(n^{4/5}) \), implying

\[ (\hat{f}(x_i) - f(x_i))^2 = O_p(n^{-4/5}). \]
$n^{-4/5}$ is the best possible rate.

However, this has asymptotic bias (proportional to $f''$) - so let $k$ go a little slower to infinity. Then the bias term disappears.

Interpretation????

JUST A TRICK!!!

Generalization: Kernel Regression

$$\hat{f}(x_i) = \sum w_j(x_i)y_j.$$ 

Really just weighted local averages.
Kernel: $K(u)$ bounded, symmetric around zero, integrates to 1 (a normalization).

Examples:

Uniform, Bartlett, Normal (not drawn)
Estimation 5

\[ w_i(x_i) = K((x_j - x_i)/\lambda)(\sum K((x_j - x_i)/\lambda) \]

\( \lambda \) is like \( k \); in fact \( k = 2\lambda n \).

Convergence rate is optimized (at \( n^{-4/5} \)) when \( \lambda = O(n^{-1/5}) \).

\( \lambda \) is the bandwidth.

Usually assume a faster rate to eliminate the bias term in constructing confidence intervals. (JUST A TRICK.)
Smoothness Restrictions:

Example: \( |f'(x)| < L \)

Solve \( \min \sum (y_i - \hat{y}_i)^2 \) s.t. \( |(\hat{y}_i - \hat{y}_j)/(x_i - x_j)| < L \).

Adding monotonicity adds the constraint

\[ \hat{y}_i < \hat{y}_j \text{ for } x_i < x_j. \]

Concavity adds another constraint.

Rates of convergence depend on the dimension of \( x \) (here 1) and on the number of derivatives. Maximal rate is \( n^{-2m/(2m+d)} \) where \( m \) is \# derivatives and \( d \) is the dimension of \( x \).
Selection of bandwidth?

Try a few and look at the results and residuals!!

Formally, use cross validation.

CV: fit \( \hat{f}_{-i} \) using all data except the ith observation, then predict \( \hat{f}_{-i}(x_i) \). Then calculate

\[
CV(\lambda) = n^{-1} \sum (y_i - \hat{f}_{-i}(x_i))^2.
\]

Choose \( \lambda \) to minimize this function.

Requires a lot of computation.
Partial Linear Model

\[ y = z \beta + f(x) + \varepsilon \]

The amazing result is that \( \beta \) can be estimated at the parametric rate.

\[
y - E(y|x) = y - E(z|x)\beta - f(x) = (z - E(z|x))\beta + \varepsilon.
\]

Suggests regressing \( y - E_y|x \) on \( z - E_z|x \).

Estimate these conditional expectations by nonparametric kernel regression.

Extends easily to higher dimensional \( z \) (estimate many conditional expectation functions and do the regression).
\[ y = f(x'\beta) + \varepsilon \]

Here \( x \) is \( k \)-dimensional - the linear index \( x'\beta \) affects \( y \) nonparametrically.

For fixed \( \beta \), \( f \) can be estimated, for example, with kernel regression, as \( \hat{f}_\beta \). Estimate \( \beta \) by minimizing

\[
n^{-1} \sum (y_i - \hat{f}_\beta(x'_i\beta))^2.\]

There is a lot of work on this problem. A basic result is that \( \beta \) can be estimated at the usual rate (variance like \( n^{-1} \)).

Binary \( y \) generalizes logit, probit.
Identification:

Note $f$ and $\beta$ are not separately identified. A normalization is necessary (typically one of the $\beta = 1$).

To estimate $f$ consistently, at least one of the regressors must be continuous.

(Think about it - we will use differentiability assumptions on $f$.)

Of course, also $\sum x_i x_i'$ must have full rank.

A little more is required.
NP estimation of residual variance:

\[ y_i = f(x_i) + \varepsilon_i \]

arranged in order of \( x \), with \( |f'| < L \)

\[
\begin{align*}
s^2 &= 1/2n^{-1} \sum (y_i - y_{i-1})^2 \\
E(s^2) &= 1/2n^{-1} \sum (f(x_i) - f(x_{i-1}))^2 \\
&\quad + 1/2En^{-1} \sum (\varepsilon_i - \varepsilon_{i-1})^2
\end{align*}
\]

First term looks like \( (f'[x_i - x_{i-1}])^2 < (L/n)^2 \)

(\( x \) cont. distributed) Cross product?
Second term is $\sigma^2$ so the estimator is consistent.

Asymptotic distribution:

$$n^{1/2}(s^2 - \sigma^2) \rightarrow N(0, \sigma^4).$$

To test against a parametric alternative, calculate $s_a^2$ from the alternative and consider

$$n^{1/2}(s_a^2 - s^2)/s^2 \rightarrow N(0, 1)$$

reject if large.
Higher Dimensions are Problems

Suppose \( y = f(x) = f(x_1, x_2) + \varepsilon \).

Estimate \( f(x_i) \) by taking an average of the \( y \) in a neighborhood of \( x_i \).
Suppose the neighborhood is a \( \lambda \times \lambda \) square?

\( W/ \) uniform \( x \) on the unit square, each neighborhood has about \( \lambda^2 n \) observations.

\[
\hat{f}(x_i) = (\lambda^2 n)^{-1} \sum y_j \\
= (\lambda^2 n)^{-1} \sum f(x_j) + (\lambda^2 n)^{-1} \sum \varepsilon_j \\
\geq f(x_i) + O(\lambda^2) + O_p(1/(\lambda n^{1/2})).
\]

Same arguments as before, but now have \( \lambda \) instead of \( \lambda^{1/2} \).
Consistency requires $\lambda \geq 0$ and $\lambda n^{1/2} \geq \infty$.

The optimal rate reduces bias and variance at the same rate. This implies $\lambda = O(n^{1/6})$. Then

$$(\hat{f} - f)^2 = O_p(n^{-2/3}).$$

This rate is optimal and is slower than the rate in the 1-dimensional model.

The same arguments work for kernel estimators in higher dimensions.

Many variations are available (different kernels, bandwidth choices, neighborhoods, etc.).
Want, say, 1% of data to form a local average (or local weighted average w/kernel).

Uniform observations, 1 dim, unit interval, local is a .01 length interval. 2 dim, unit square, local is $0.01^{1/2} = 0.1$ unit square - 1/10 the range in each dimension.

Generally, $0.01^{1/p}$ where $p$ is the dim. Gets nonlocal fast.

Picture?

Mean distance to origin increases w/dimension - most points are “near” the boundary.
The source of most of this lecture and a great reference on applied nonparametric and semiparametrics (like the partial linear model) is Adonis Yatchew (2003).

Semiparametric Regression for the Applied Econometrician, Cambridge University Press.