Nonparametric Regression

John Fox
Department of Sociology
McMaster University
1280 Main Street West
Hamilton, Ontario
Canada L8S 4M4
jfox@mcmaster.ca

February 2004
Abstract

Nonparametric regression analysis traces the dependence of a response variable on one or several predictors without specifying in advance the function that relates the predictors to the response. This article discusses several common methods of nonparametric regression, including kernel estimation, local polynomial regression, and smoothing splines. Additive regression models and semiparametric models are also briefly discussed.

Keywords: kernel estimation; local polynomial regression; smoothing splines; additive regression; semiparametric regression
Nonparametric regression analysis traces the dependence of a response variable \((y)\) on one or several predictors \((xs)\) without specifying in advance the function that relates the response to the predictors:

\[
E(y_i) = f(x_{1i}, \ldots, x_{pi})
\]

where \(E(y_i)\) is the mean of \(y\) for the \(i\)th of \(n\) observations. It is typically assumed that the conditional variance of \(y\), \(\text{Var}(y_i|x_{1i}, \ldots, x_{pi})\) is a constant, and that the conditional distribution of \(y\) is normal, although these assumptions can be relaxed.

Nonparametric regression is therefore distinguished from linear regression, in which the function relating the mean of \(y\) to the \(xs\) is linear in the parameters,

\[
E(y_i) = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}
\]

and from traditional nonlinear regression, in which the function relating the mean of \(y\) to the \(xs\), though nonlinear in its parameters, is specified explicitly,

\[
E(y_i) = f(x_{1i}, \ldots, x_{pi}; \gamma_1, \ldots, \gamma_k)
\]

In traditional regression analysis, the object is to estimate the parameters of the model — the \(\beta\)s or \(\gamma\)s. In nonparametric regression, the object is to estimate the regression function directly.

There are many specific methods of nonparametric regression. Most, but not all, assume that the regression function is in some sense smooth. Several of the more prominent methods are described in this article. Moreover, just as traditional linear and nonlinear regression can be extended to generalized linear and nonlinear regression models that accommodate non-normal error distributions, the same is true of nonparametric regression. There is a large literature on nonparametric regression analysis, both in scientific journals and in texts. For more extensive introductions to the subject, see in particular, Bowman and Azzalini [1], Fox [2, 3], Hastie and Tibshirani [4], Hastie, Tibshirani, and Freedman [5], and Simonoff [6].

The simplest use of nonparametric regression is in smoothing scatterplots. Here, there is a numerical
Figure 1: Female expectation of life by GDP per capita, for 154 nations of the world. The solid line is for a local-linear regression with a span of 0.5, while the broken line is for a similar fit deleting the four outlying observations that are labelled on the plot.

(response y and a single predictor x, and we seek to clarify visually the relationship between the two variables in a scatterplot. Figure 1, for example, shows the relationship between female expectation of life at birth and GDP per capita for 154 nations of the world, as reported in 1998 by the United Nations. Two fits to the data are shown, both employing local linear regression (described below); the solid line represents a fit to all of the data, while the broken line omits four outlying nations, labelled on the graph, which have values of female life expectancy that are unusually low given GDP per capita. It is clear that although there is a positive relationship between expectation of life and GDP, the relationship is highly nonlinear, levelling off substantially at high levels of GDP.

Three common methods of nonparametric regression are kernel estimation, local-polynomial regression (which is a generalization of kernel estimation), and smoothing splines. Nearest-neighbor kernel estimation proceeds as follows (as illustrated for the UN data in Figure 2):

1. Let $x_0$ denote a focal x-value at which $f(x)$ is to be estimated; in Figure 2 (a), the focal value is the 80th ordered x-value in the UN data, $x_{(80)}$. Find the $m$ nearest x-neighbors of $x_0$, where $s = m/n$ is called the span of the kernel smoother. In the example, the span was set to $s = 0.5$, and thus $m = 0.5 \times 154 = 77$. Let $h$ represent the half-width of a window encompassing the $m$ nearest neighbors.
Figure 2: How the kernel estimator works: (a) A neighborhood including the 77 observations closest to \( x_{(80)} \), corresponding to a span of 0.5. (b) The tricube weight function defined on this neighborhood; the points show the weights for the observations. (c) The weighted mean of the \( y \)-values within the neighborhood, represented as a horizontal line. (d) The nonparametric regression line connecting the fitted values at each observation. (The four outlying points are excluded from the fit.)
of \( x_0 \). The larger the span (and hence the value of \( h \)), the smoother the estimated regression function.

2. Define a symmetric unimodal weight function, centered on the focal observation, that goes to zero (or nearly zero) at the boundaries of the neighborhood around the focal value. The specific choice of weight function is not critical: In Figure 2 (b), the tricube weight function is used:

\[
W_T(x) = \begin{cases} 
1 - \left( \frac{|x - x_0|}{h} \right)^3 & \text{for } \frac{|x - x_0|}{h} < 1 \\
0 & \text{for } \frac{|x - x_0|}{h} \geq 1
\end{cases}
\]

A Gaussian (normal) density function is another common choice.

3. Using the tricube (or other appropriate) weights, calculate the weighted average of the \( y \)-values to obtain the fitted value

\[
\hat{y}_0 = \hat{f}(x_0) = \frac{\sum W_T(x_i)y_i}{\sum W_T(x_i)}
\]

as illustrated in Figure 2 (c). Greater weight is thus accorded to observations whose \( x \)-values are close to the focal \( x_0 \).

4. Repeat this procedure at a range of \( x \)-values spanning the data — for example, at the ordered observations \( x(1), x(2), \ldots, x(n) \). Connecting the fitted values, as in Figure 2 (d), produces an estimate of the regression function.

Local polynomial regression is similar to kernel estimation, but the fitted values are produced by locally weighted regression rather than by locally weighted averaging; that is, \( \hat{y}_0 \) is obtained in step 3 by the polynomial regression of \( y \) on \( x \) to minimize the weighted sum of squared residuals

\[
\sum W_T(x_i)(y_i - a - b_1x_i - b_2x_i^2 - \cdots - b_kx_i^k)^2
\]

Most commonly, the order of the local polynomial is taken as \( k = 1 \), that is, a local linear fit (as in Figure 1). Local polynomial regression tends to be less biased than kernel regression, for example at the boundaries of data — as in evident in the artificial flattening of the kernel estimator at the right of Figure 2 (d). More generally, the bias of the local-polynomial estimator declines and the variance increases with the order of
the polynomial, but an odd-ordered local polynomial estimator has the same asymptotic variance as the preceding even-ordered estimator: Thus, the local-linear estimator (of order 1) is preferred to the kernel estimator (of order 0), and the local-cubic (order 3) estimator to the local-quadratic (order 2).

Smoothing splines are the solution to the penalized regression problem: Find \( \hat{f}(x) \) to minimize

\[
S(h) = \sum (y_i - f(x_i))^2 + h \int \left[ f''(x) \right]^2 dx
\]

Here \( h \) is a roughness penalty, analogous to the span in nearest-neighbor kernel or local polynomial regression, and \( f'' \) is the second derivative of the regression function (taken as a measure of roughness). Without the roughness penalty, nonparametrically minimizing the residual sum of squares would simply interpolate the data. The mathematical basis for smoothing splines is more satisfying than for kernel or local polynomial regression, since an explicit criterion of fit is optimized, but spline and local polynomial regressions of equivalent smoothness tend to be similar in practice.

Local regression with several predictors proceeds as follows, for example. We want the fit \( \hat{y}_0 = \hat{f}(x_0) \) at the focal point \( x_0 = (x_{10}, \ldots, x_{p0}) \) in the predictor space. We need the distances \( D(x_i, x_0) \) between the observations on the predictors and the focal point. If the predictors are on the same scale (as, for example, when they represent coordinates on a map), then measuring distance is simple; otherwise, some sort of standardization or generalized distance metric will be required. Once distances are defined, weighted polynomial fits in several predictors proceed much as in the bivariate case. Some kinds of spline estimators can also be generalized to higher dimensions.

The generalization of nonparametric regression to several predictors is therefore mathematically straightforward, but it is often problematic in practice. First, multivariate data are afflicted by the so-called curse of dimensionality: Multidimensional spaces grow exponentially more sparse with the number of dimensions, requiring very large samples to estimate nonparametric regression models with many predictors. Second, although slicing the surface can be of some help, it is difficult to visualize a regression surface in more than three dimensions (that is, for more than two predictors).

Additive regression models are an alternative to unconstrained nonparametric regression with several predictors. The additive regression model is
\[ E(y_i) = \alpha + f_1(x_{1i}) + \cdots + f_p(x_{pi}) \]

where the \( f_j \) are smooth partial-regression functions, typically estimated with smoothing splines or by local regression. This model can be extended in two directions: (1) To incorporate interactions between (or among) specific predictors; for example

\[ E(y_i) = \alpha + f_1(x_{1i}) + f_{23}(x_{2i}, x_{3i}) \]

which is not as general as the unconstrained model \( E(y_i) = \alpha + f(x_{1i}, x_{2i}, x_{3i}) \). (2) To incorporate linear terms, as in the model

\[ E(y_i) = \alpha + \beta_1 x_{1i} + f_2(x_{2i}) \]

Such semiparametric models are particularly useful for including dummy regressors or other contrasts derived from categorical predictors.

Returning to the UN data, an example of a simple additive regression model appears in Figures 3 and 4. Here female life expectancy is regressed on GDP per capita and the female rate of illiteracy, expressed as a percentage. Each term in this additive model is fit as a smoothing spline, using the equivalent of four degrees of freedom. Figure 3 shows the two-dimensional fitted regression surface, while Figure 4 shows the partial regression functions, which in effect slice the regression surface in the direction of each predictor; because the surface is additive, all slices in a particular direction are parallel, and the two-dimensional surface in three-dimensional space can be summarized by two two-dimensional graphs. The ability to summarize the regression surface with a series of two-dimensional graphs is an even greater advantage when the surface is higher-dimensional.
Figure 3: Fitted regression surface for the additive regression of female expectation of life on GDP per capita and female illiteracy. The vertical lines represent residuals.
Figure 4: Partial regression functions from the additive regression of female life expectancy on GDP per capita and female illiteracy; each term is fit by a smoothing spline using the equivalent of four degrees of freedom. The “rug plot” at the bottom of each graph shows the distribution of the predictor, and the broken lines give a point-wise 95-percent confidence envelope around the fit.
A central issue in nonparametric regression is the selection of smoothing parameters — such as the span in kernel and local-polynomial regression or the roughness penalty in smoothing-spline regression (or equivalent degrees of freedom for any of these). In the examples in this article, smoothing parameters were selected by visual trial and error, balancing smoothness against detail. The analogous statistical balance is between variance and bias, and some methods (such as cross-validation) attempt to select smoothing parameters to minimize estimated mean-square error (i.e., the sum of squared bias and variance).
References


