Advanced Statistical Modeling

Part 2. Nonparametric Modeling

Session 1:

Nonparametric regression model I

Pedro Delicado

Departament d'Estadística i Investigació Operativa

Universitat Politècnica de Catalunya

Introduction to nonparametric modeling

- Nonparametric statistical methods are techniques that do not require assuming parametric hypothesis about the data probability distribution.
- Classic nonparametric methods: From the middle of the XX century, nonparametric techniques, mainly hypothesis tests, that are based on the empirical distribution function and ranks.
- A few decades later there appeared a second generation of nonparametric methods, nonparametric function estimation or smoothing methods, with the aim of estimating a whole function related with the data probability distribution.
- This second kind of techniques is the object of the second part of the course ASM.

Contents of the course

Session 1: Nonparametric regression model I. 0. Introduction to nonparametric modeling. 1. Local polynomial regression.

Session 2: Nonparametric regression model II. 2. Kernel functions. 3. Theoretical properties. The bias-variance trade off. 4. Linear smoothers.

Session 3: Nonparametric regression model III. 5. Choosing the degree of the local polynomial. 6. Choosing the smoothing parameter: Cross validation, plug-in methods, varying windows.

Session 4: Generalized nonparametric regression model. 1. Nonparametric regression with binary response. 2. Generalized nonparametric regression model. 3. Estimation by maximum local likelihood.

Session 5: Inference with nonparametric regression. 1. Variability bands. 2. Testing for no effects. 3. Checking a parametric model. 4. Comparing curves.

Session 6: Spline smoothing. 1. Penalized least squares nonparametric regression. 2. Splines, cubic splines and interpolation. 3. Smoothing splines. 4. B-splines and P-splines. 5. Spline regression. 6. Fitting generalized nonparametric regression models with splines.

Session 7: Generalized additive models and Semiparametric models. 1. Multiple nonparametric regression. The curse of dimensionality. 2. Additive models. 3. Generalized additive models. 4. Semiparametric models.

Example 1. Density estimation. *CD rate data.*

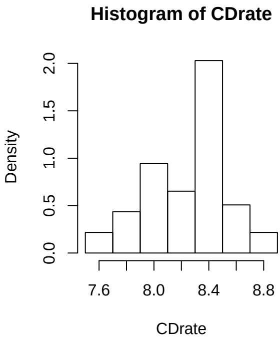
- This data set represents the three-month certificate of deposit (CD) rates for 69 Long Island banks and thrifts (saving and loan associations) in August 1989.
- Two types of institutions: banks and thrifts.

Stem-and-Leaf Plot:

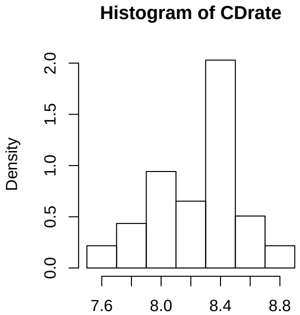
```
The decimal point is 1 digit(s) to the left of the |
  74 | 167
  76 | 15
  78 | 2200
  80 | 0000000000556157
  82 | 0550003334556
  84 | 0000000599000000000001257
  86 | 550158
```

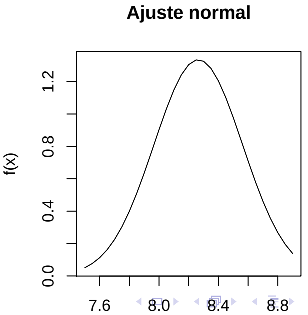
This graphic allows us to visualize the data distribution (is a kind of rotated histogram) without losing numerical information.

- A better graphical representation: the data histogram.
- The histogram was the first nonparametric density estimator
- It shows what sections of the real line gather more probability than others.
- We can see bimodality and left asymmetry.
- Drawbacks of the histogram: is a non-continuous step function.

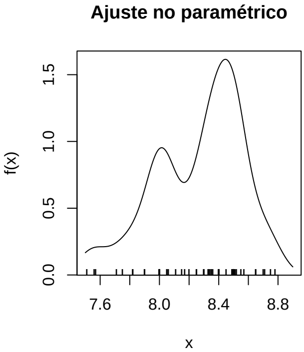


- An alternative way of density estimation: To assume a parametric model.
- We assume, for instance, normality. Then we only need to estimate the two parameters, mean and standard deviation, that characterize a particular normal distribution. We use the sample version of them.
- Drawbacks: The parametric model is too rigid. For instance, normality implies symmetry and unimodality. That is against the data histogram.



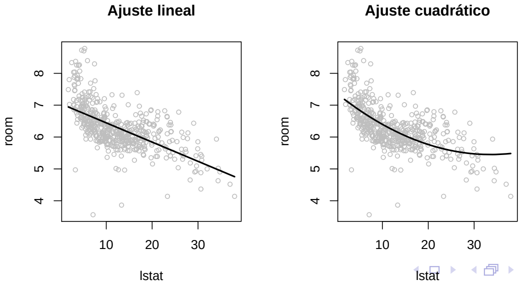


- The kernel density estimator is a nonparametric estimator that outperforms the histogram.
- It is smooth and it respects the data asymmetry and bimodality.

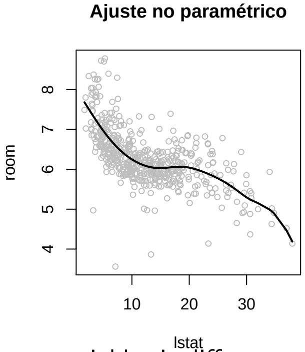


Example 2. Regression with continuous response.

- Boston House-price Data, 506 neighborhoods of Boston, 1978.
- `http://lib.stat.cmu.edu/datasets/boston_corrected.txt`
- The list of variables includes: `RM` average number of rooms per dwelling, `LSTAT` % of the population with the lower status in a social-class classification, `CRIM` per capita crime rate by town, `AGE` proportion of owner-occupied units built prior to 1940, `MEDV` Median value of owner-occupied homes in $1000's
- We study `RM` as a function of `LSTAT`. Parametric regression.



Nonparametric fit of `room` versus `lstat`



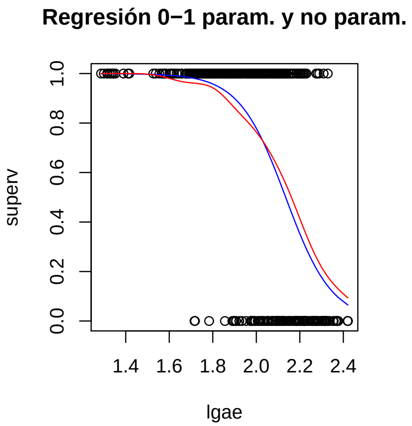
- The relation between variables is different when `lstat` is lower than 10%, when it is between 10% and el 20%, or when it is greater than 20%.
- In the middle range of `lstat` the values of `room` are almost constant. In the other two sections `room` is a decreasing function of `lstat`.
- The fall is steeper at the first section than at the third one.

Example 3. Regression with binary response

- Burn injuries data (Fan and Gijbels (1996)).
- Data from 435 adults (between ages 17 and 85) suffering from burn injuries.
- The binary response variable is taken to be 1 for those victims who survived their burn injuries and zero otherwise: `surv`.
- `lgae`, log(area of third degree burn + 1) is taken as a covariate.
- The conditional expectation of `surv` given a level of `lgae` is the conditional probability of survival given this particular value of `lgae`.

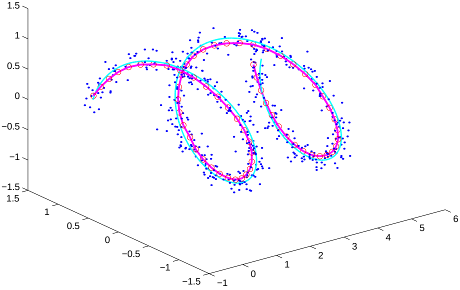
Parametric and nonparametric fits

We show the data and the estimated survival probability using the logistic and a nonparametric estimator.



Example 4. Principal curves.

- Principal curves are one of the nonlinear generalizations of principal components.
- They were first defined by Trevor Hastie and Werner Stuetzle as "self-consistent" smooth curves which pass through the "middle" of a d-dimensional probability distribution or data cloud.



Uses of smoothing methods

- Exploratory Data Analysis. Smoothing methods provide nice graphical representations of *density functions* or *regression functions* and their derivatives, among other.
- Modeling. Many times the inspection of an accurate graphical description of the observed data suggests to the researcher a tentative statistical model for them. For instance, a bimodal estimated density suggests the possibility of having data coming from a mixture of two subpopulation. Then a mixture of two parametric distributions is considered as a potential model for the data.
- Inference problems. Confidence bands for an unknown function, hypothesis testing involving functions (independence between two variables, equal distribution on two or more subpopulations, ...).

- Goodness-of-fit for a parametric model. Consider the random variable $X \sim f$. We want to test

$$H_0 : f \in \mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}, \text{ against } H_1 : f \notin \mathcal{F}_\Theta$$

A useful statistic for testing this hypothesis is $T = d(f_{\hat{\theta}}, \hat{f})$, where $\hat{\theta}$ is an estimator of θ (then $f_{\hat{\theta}}$ is a parametric estimator of f), $\hat{f}$ is a nonparametric estimator of f and $d(\cdot, \cdot)$ is a distance between density functions. Then $d(f_{\hat{\theta}}, \hat{f})$ is a kind of distance between the data and the null hypothesis.

- Parametric estimation. Assume that $X \sim f_{\theta_0}$, for some $\theta_0 \in \Theta$. The minimum distance estimator of θ is given by

$$\hat{\theta} = \arg\min_{\theta \in \Theta} d(f_\theta, \hat{f}).$$

- Defining new statistical methods. Many standard statistical procedures can be modified just changing $f_{\hat{\theta}}$ by $\hat{f}$. This modification usually allows the method to be applied to a wider range of situations because the parametric hypothesis is no longer required.

The regression function

- Let (X, Y) be random variables with continuous joint distribution.
- The best prediction (in the sense of minimum mean squared prediction error) of the *dependent variable* Y given that the predicting variable X takes the known value x, is the conditional expectation

$$m(x) = E(Y|X = x),$$

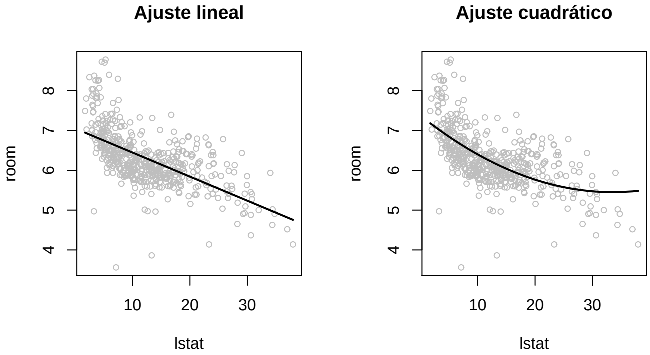
also known as regression function.
- The parametric regression models assume that the function $m(\cdot)$ is known except for a fixed finite number of unknown parameters.
- For instance, the simple linear regression model postulates that

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

So $m(x) = \beta_0 + \beta_1 x$ is known except for two parameters: β_0, β_1.

Example of parametric regression

Parametric fits of variable room as a function of variable lstat.



The nonparametric regression model

- We observe n pairs of data (x_i, y_i) coming from the nonparametric regression model

  $$y_i = m(x_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

  where $\varepsilon_1, \ldots, \varepsilon_n$ are independent r.v. with

  $$E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2 \text{ for all } i,$$

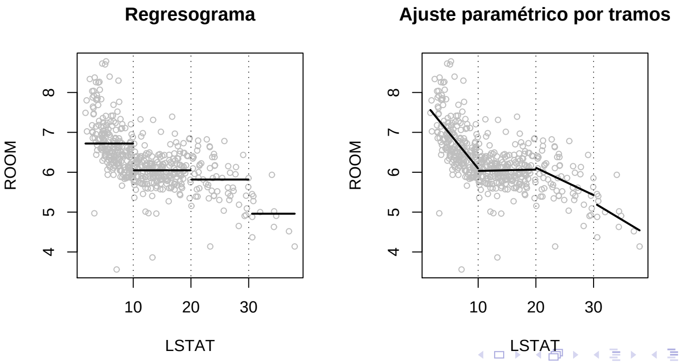
  and the input variable values $x_1, \ldots, x_n$ are known.
- The functional form of the regression function $m(x)$ is not specified.
- Certain regularity conditions on $m(x)$ are assumed. For instance, it is usually assumed that $m(x)$ has continuous second derivative.

What does it mean "fitting a nonparametric regression model"?

- To provide an estimator $\hat{m}(x)$ of $m(x)$ for all $x \in \mathbb{R}$.
  - This usually implies to draw the graphic of the pairs $(t_j, \hat{m}(t_j))$, $j = 1, \ldots, J$, where t_j, $j = 1, \ldots, J$ is a regular fine grid covering the range of observed values x_i, $i = 1, \ldots, n$.
  - An algorithm that computes $\hat{m}(t)$ for any input value t can be provided alternatively.
- To give an estimator $\hat{\sigma}^2$ of the residual variance σ^2.

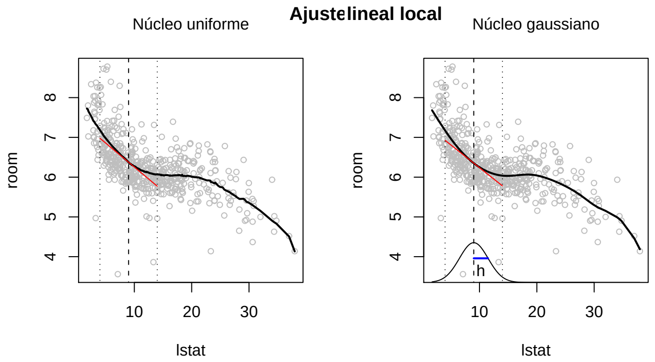
Example: Boston housing data

- The scatter plot of variables LSTAT and ROOM suggests that a unique linear model is not valid for the whole range of LSTAT.
- A first idea: To divide the range of LSTAT in several intervals, each of them showing an approximately linear relation between both variables.



Good results, but not entirely satisfactory. Two improvements:

- In order to estimate the regression function at a given value t, using data (x_i, y_i) such that x_i is in an interval centered at t.
- Assigning to each datum (x_i, y_i) a weight $w(x_i, t)$ being a decreasing function of distance $|t - x_i|$.



Local linear fitting.

- Weights are assigned by a kernel function K.
- The weight of (x_i, y_i) when estimating $m(t)$ is
$$w_i = w(t, x_i) = K\left(\frac{x_i - t}{h}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{x_j - t}{h}\right),$$
- The scale parameter h controls how the total weight is concentrated around t.
- For small values of h only the closest observations to t have a relevant weight. On the other hand, a large h allows data distant from t to be taken into account when estimating $m(t)$.
- h is called smoothing parameter or bandwidth.
- The final estimate is significantly affected by changes in the choice of smoothing parameter, so this task is crucial in nonparametric estimation.

- Once the weights $w_i = w(t, x_i)$ have been calculated, the following weighted least squares problem is solved:

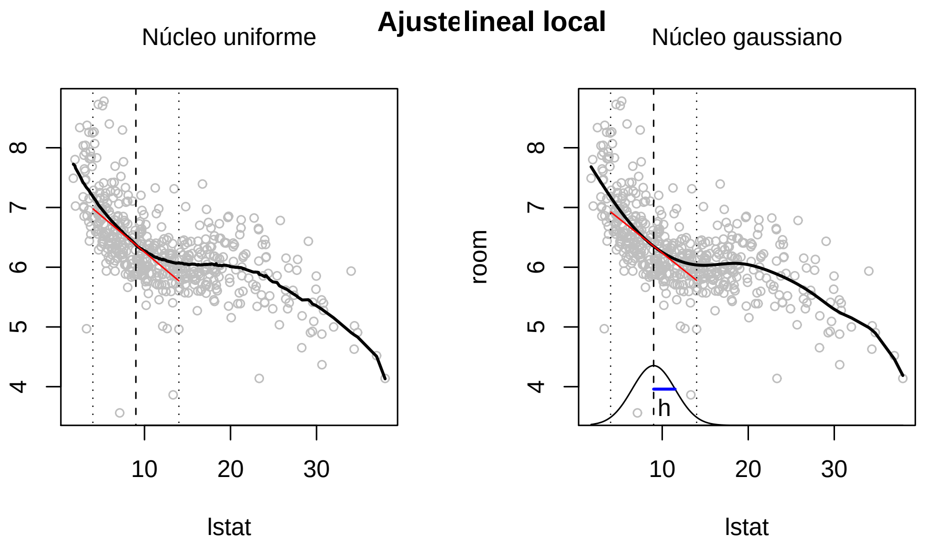
$$\min_{a,b} \sum_{i=1}^{n} w_i \left(y_i - (a + b(x_i - t))\right)^2 .$$

- The optimal parameters a and b depend on t, because the weights $w(t, x_i)$ depend on t: $a = a(t), b = b(t)$.
- The regression line fitted around t is

$$l_t(x) = a(t) + b(t)(x - t).$$

- Finally, the regression function estimation at point t is the value that $l_t(x)$ takes when $x = t$:

$$\hat{m}(t) = l_t(t) = a(t).$$


Ajuste lineal local
Núcleo uniforme
Núcleo gaussiano
room
4
5
6
7
8
10
20
30
h
lstat
lstat

Practice:

Write your own local linear regression function

Local polynomial fitting

- Consider the weighted polynomial regression problem

$$\min_{\beta_0,\ldots,\beta_q} \sum_{i=1}^{n} w_i \left(y_i - (\beta_0 + \beta_1(x_i - t) + \cdots + \beta_q(x_i - t)^q)\right)^2 .$$

- Observe that the estimated coefficients depend on t, the point for which the regression function is being estimated: $\hat{\beta}_j = \hat{\beta}_j(t)$.
- Finally, the proposed estimate for $m(t)$ is the value of the locally fitted polynomial $P_{q,t}(x) = \sum_{j=0}^{p} \hat{\beta}_j (x-t)^j$ evaluated at $x = t$:

$$\hat{m}_q(t) = P_{q,t}(t) = \hat{\beta}_0(t).$$

- Moreover the estimated polynomial $P_{q,t}(x)$ allows us to estimate the first q derivatives of m at t:

$$\hat{m}_q^{(s)}(t) = \frac{d^s}{dx^s} \left(P_{q,t}(x)\right)\Big|_{x=t} = s! \hat{\beta}_s(t).$$

Particular case: Nadaraya-Watson estimator

- When the degree of the polynomial locally fitted is $q = 0$ (that is, a constant) the resulting nonparametric estimator of $m(t)$ is known as Nadaraya-Watson estimator or, simply, kernel estimator:

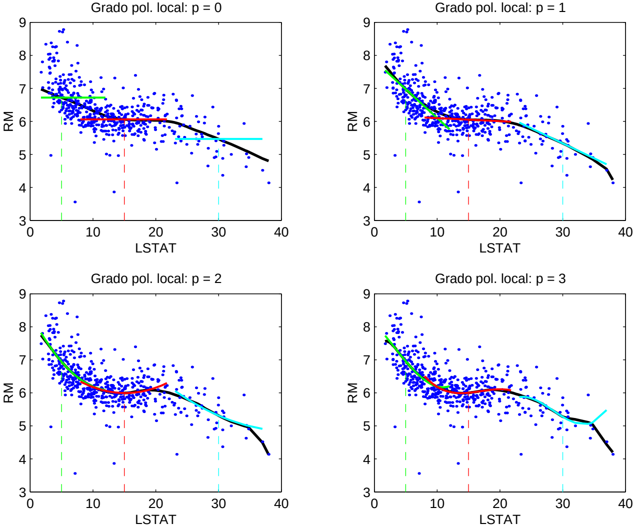
$$\hat{m}_K(t) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - t}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - t}{h}\right)} = \sum_{i=1}^{n} w(t, x_i) y_i.$$

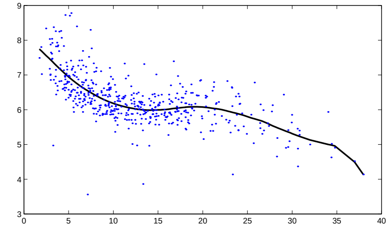
- Nadaraya-Watson was proposed before local polynomial estimators.
- Observe that $\hat{m}_K(t)$ is a moving weighted mean.
- It can be proved that every local polynomial estimator is itself a weighted mead,

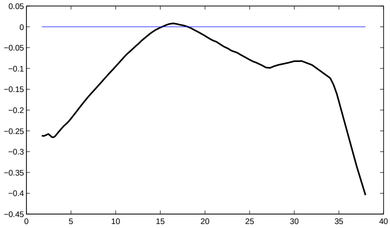
$$\hat{m}_q(t) = \sum_{i=1}^{n} w_q^*(t, x_i) y_i.$$

but weights $w_q^*(t, x_i)$ are not necessarily non-negative.

Example: Boston housing data




9
8
7
6
5
4
3
0
5
10
15
20
25
30
35
40


0.05
0
−0.05
−0.1
−0.15
−0.2
−0.25
−0.3
−0.35
−0.4
−0.45
0
5
10
15
20
25
30
35
40

Matrix formulation of the local polynomial estimator

Let

$$X_t = \begin{pmatrix} 1 & (x_1 - t) & \dots & (x_1 - t)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - t) & \dots & (x_n - t)^q \end{pmatrix}$$

be the regressors matrix.

Define $Y = (y_1, \dots, y_n)^{\mathsf{T}}$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^{\mathsf{T}}$, $\beta = (\beta_0, \dots, \beta_q)^{\mathsf{T}}$.

Let $W_t = \mathrm{Diag}(w(x_1, t), \dots, w(x_n, t))$ be the weight matrix.

We fit the multiple linear regression model $Y = X_t\beta + \varepsilon$ using generalized least squares (GLS):

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{q+1}} (Y - X_t\beta)^{\mathsf{T}} W_t (Y - X_t\beta).$$

The solution is

$$\hat{\beta} = \left(X_t^{\mathsf{T}} W_t X_t\right)^{-1} X_t^{\mathsf{T}} W_t Y.$$

- Solution: $\hat{\beta} = \left(X_t^\mathsf{T} W_t X_t\right)^{-1} X_t^\mathsf{T} W_t Y$.
- For $j = 0, \ldots, q$, let e_j be the $(q+1)$-dimensional vector having all its coordinates 0 except the $(j+1)$-th one, that is equal to 1.
- Then

$$\hat{m}_q(t) = \hat{\beta}_0 = e_0^\mathsf{T} \hat{\beta} = e_0^\mathsf{T} \left(X_t^\mathsf{T} W_t X_t\right)^{-1} X_t^\mathsf{T} W_t Y = S_t Y = \sum_{i=1}^{n} w_q^*(t, x_i) y_i,$$

  where $S_t = e_0^\mathsf{T} \left(X_t^\mathsf{T} W_t X_t\right)^{-1} X_t^\mathsf{T} W_t$ is a n-dimensional row vector.
- We say that the local polynomial regression estimator is a linear estimator because, for a fix t, $\hat{m}_q(t)$ is a linear function of $y_1, \ldots, y_n$.
- The local polynomial estimator of the s-th derivative of m at point t is

$$\hat{m}_q^{(s)}(t) = s! \hat{\beta}_s(t) = s! e_s^\mathsf{T} \hat{\beta},$$

  that is also linear in $y_1, \ldots, y_n$.

Practice:

- Local polynomial regression in R with functions `lpr_visual` and `locpolreg`.
- Local polynomial estimation in R: standard libraries and functions.

Bowman, A. W. and A. Azzalini (1997).
Applied Smoothing Techniques for Data Analysis.
Oxford: Oxford University Press.

Fan, J. and I. Gijbels (1996).
Local polynomial modelling and its applications.
London: Chapman & Hall.

Hastie, T., R. Tibshirani, and J. Friedman (2001).
The Elements of Statistical Learning. Data Mining, Inference, and Prediction.
Springer.

Loader, C. (1999).
Local regression and likelihood.
New York: Springer.

Simonoff, J. S. (1996).
Smoothing methods in statistics.
New York: Springer.

Wand, M. P. and M. C. Jones (1995).
Kernel smoothing.
London: Chapman and Hall.

Wasserman, L. (2006).
All of Nonparametric Statistics.
New York: Springer.

Wood, S. (2006).
Generalized Additive Models: An Introduction with R.
Chapman and Hall/CRC.