Outline: **Smoothing**
(F. Chiaromonte)

Non-parametric regression with the LOWESS

Non-parametric regression

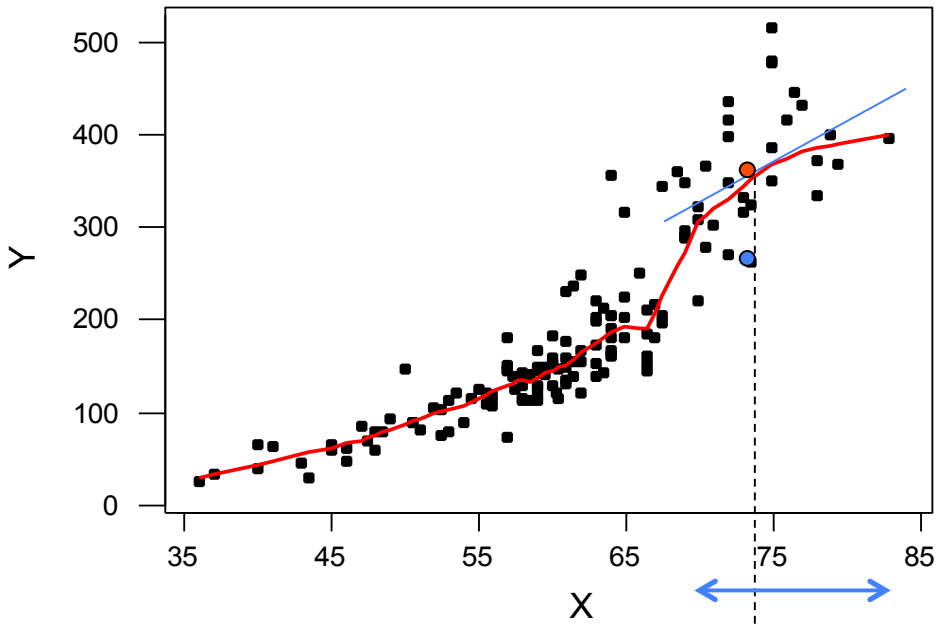
Do without the specification of a parametric model (an explicit equation), for the regression function. If p is very small (p=1,2) and we have a fairly large n, we can instead let the data suggest the shape of the systematic relationship linking the response to the features – using a smoothing technique.

For instance, **LOWESS**: *Locally Weighted Scatter Plot Smoothing.* Implemented in most statistical software packages, including R. See
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lowess.html

The procedure uses local Least Square fits of linear or quadratic forms, with weights and iterations (to dampen effect of outliers).

To implement it, one needs to specify:
- the fraction of data in each local neighborhood (smoothing parameter, q in (0,1))
- Degree of locally fitted polynomial (1=linear most common, but also 2=quadratic)
- Weight function for the least square fit
- Number of iterative weighted least square fits

For instance:
Smoothing param = 0.2
Degree = 1
# iterations = 2



Weighted least squares linear fit, reduce weight of outliers, iterate twice

Data point, and 20% of points closest to it in X

Perform moving along the data points…

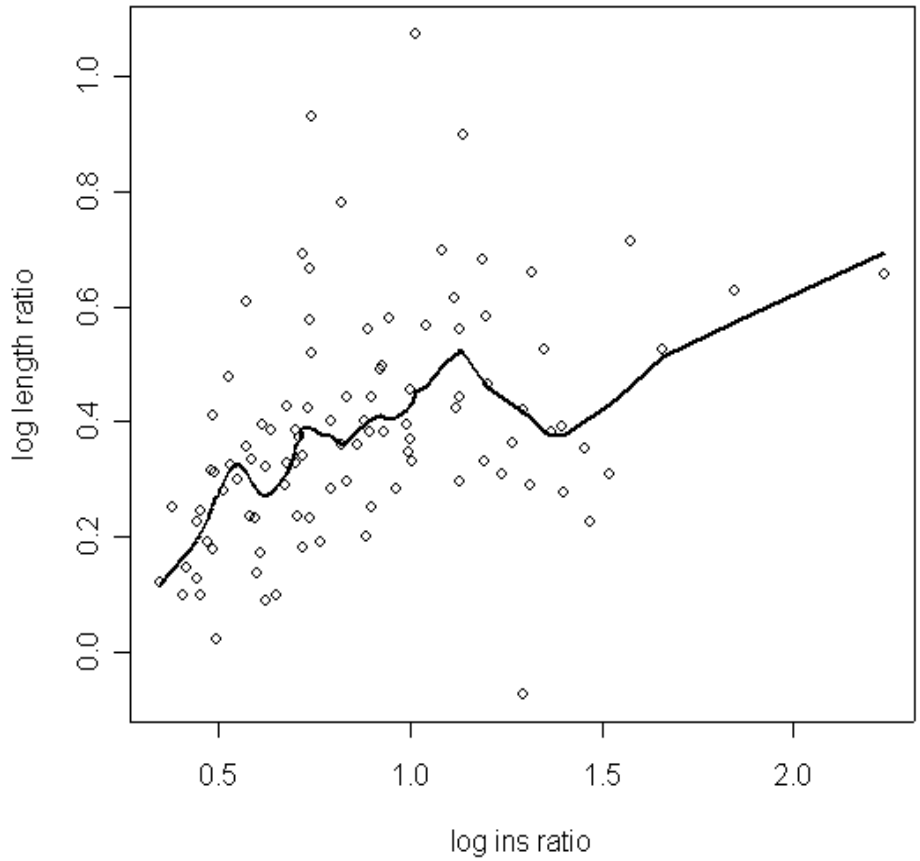
Weighted least squares linear fit

$$\min_{\beta_0,\beta_1} \sum_{j \in N_i} w_j (y_j - (\beta_0 + \beta_1 x_j))^2$$

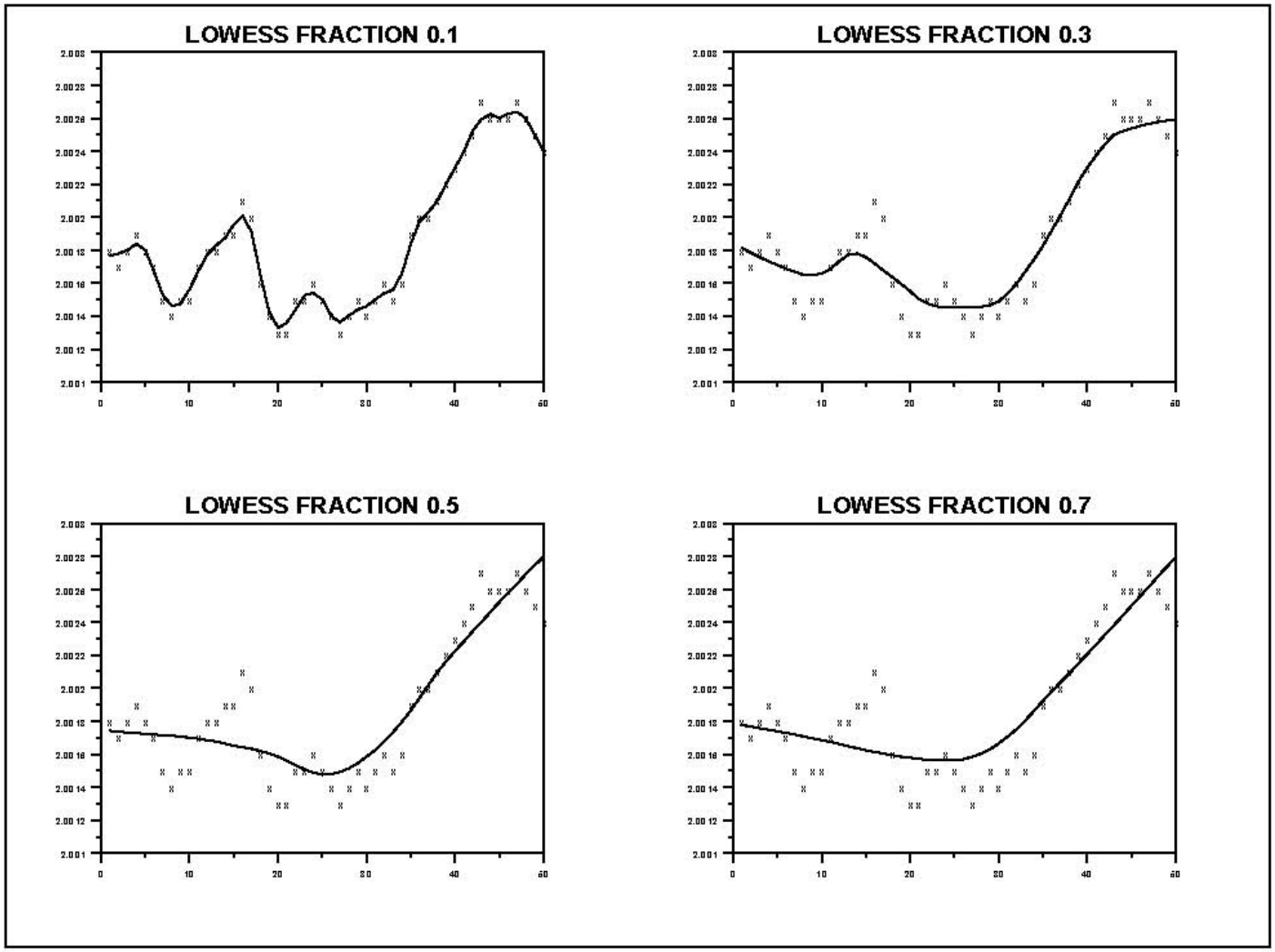
Weight function

$$w_i = \left(1 - \left(\frac{d(x_j, x_i)}{\max_{1 \in N_i} d(x_1, x_i)}\right)^3\right)^3$$

An example: first degree, q=0.2, # iterations = 2



Fraction controls degree of smoothing:



Uses of the LOWESS together with *parametric* regression:

• On the plot of Y vs X, can suggest an appropriate regression function (the form of an explicit equation may be suggested by the smoother, and one can then fit the corresponding parametric model).
• On the plot of residuals vs X or vs the fitted values; helps diagnosing departures from whatever regression funxtion was postulated (e.g. a line) – visualizing systematic mean patterns in the residuals.

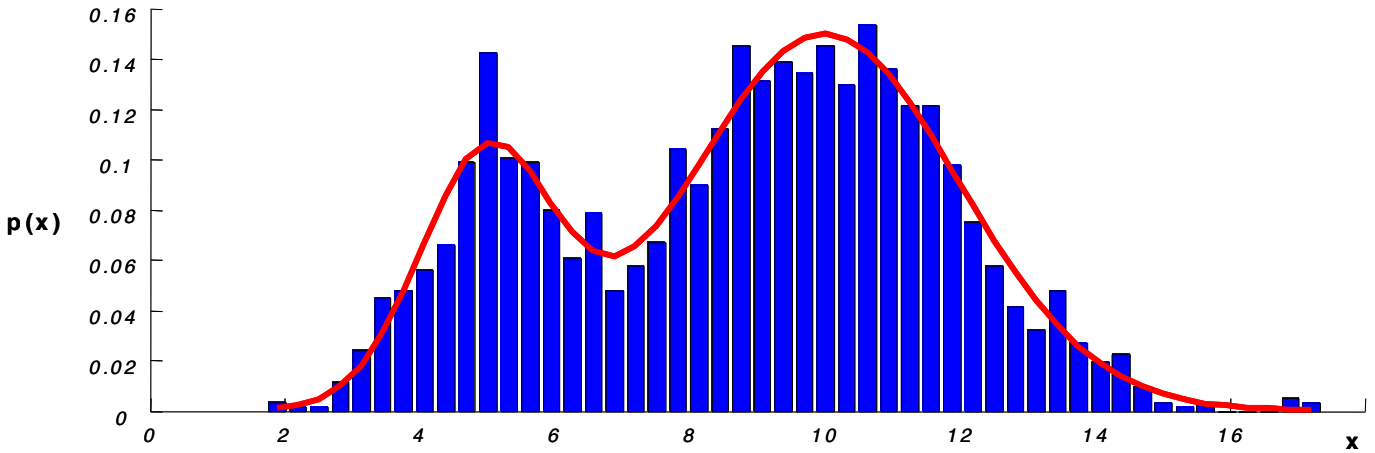
Drawbacks of non-parametric regression approaches:

• No explicit equation is obtained for the systematic relationship between response and predictors (but the form of an explicit equation may be suggested)
• The smoothing parameter (a tuning parameter of the procedure) is critical; how does one chose an appropriate level of smoothing? What's systematic signal and what is noise?

Note: also Smoothing Splines and Kernel Smoothing are used in regression. These non-parametric smoothing technique are also used to estimate density functions (less so lowess).

Density estimation with Kernels

DENSITY ESTIMATION



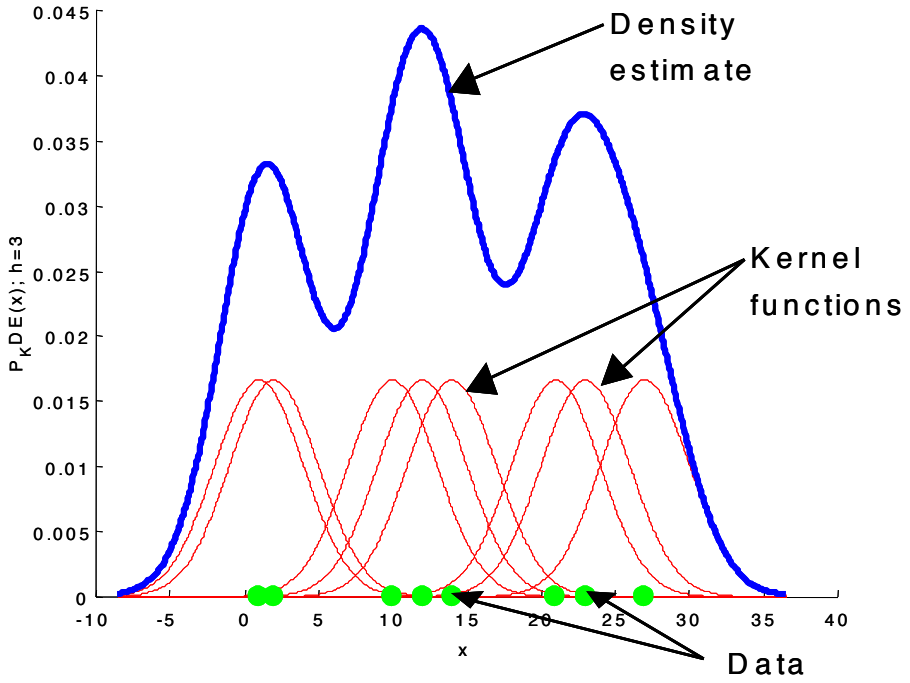
Instead of the histogram, produce a continuous, smooth estimate of the underlying density based on the empirical distribution (the observed data)

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big)$$

the smooth kernel estimate is a sum of “bumps”

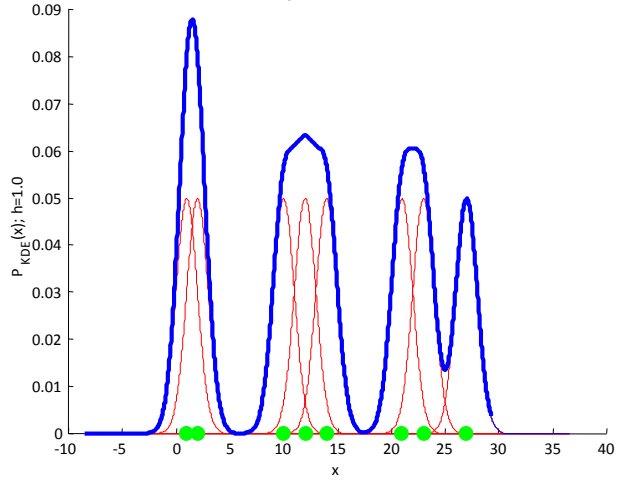
The kernel function determines the shape of the bumps

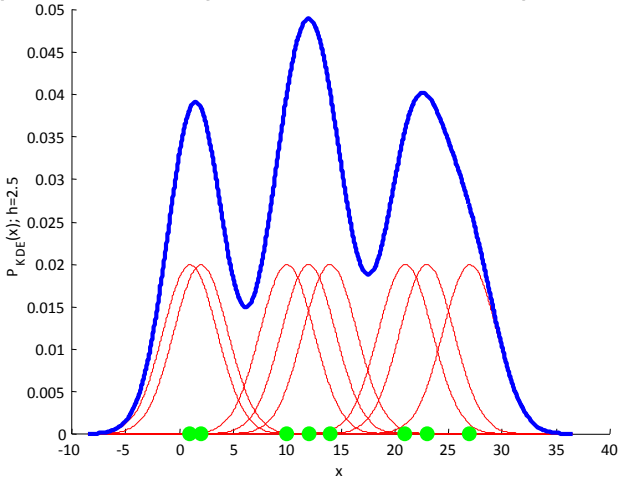
The parameter h, also called the smoothing parameter or bandwidth, determines their width

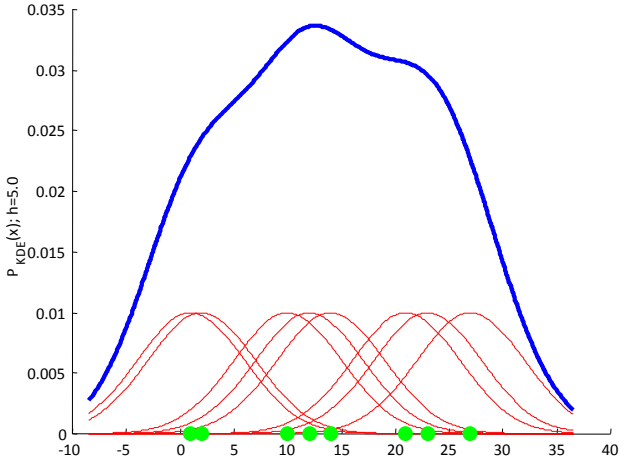


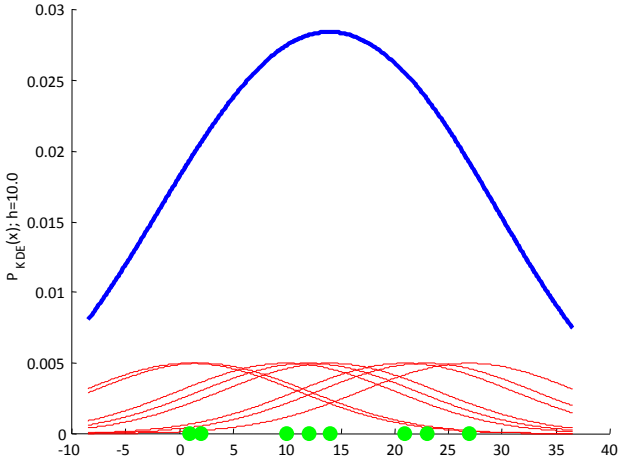
The problem of choosing h is crucial in density estimation

- A large h will over-smooth the DE and mask the structure of the data
- A small h will yield a DE that is spiky and very hard to interpret









We would like to find a value of h that minimizes the error between the estimated density and the true density

- A natural measure is the MSE at the estimation point x, defined by

$$E[(p_{KDE}(x) - p(x))^2] = \underbrace{E[p_{KDE}(x) - p(x)]^2}_{bias} + \underbrace{var(p_{KDE}(x))}_{variance}$$

bias-variance tradeoff

- A large bandwidth will reduce the differences among the estimates of $p_{KDE}(x)$ for different data sets (the variance), but it will increase the bias of $p_{KDE}(x)$ with respect to the true density $p(x)$
- A small bandwidth will reduce the bias of $p_{KDE}(x)$, at the expense of a larger variance in the estimates $p_{KDE}(x)$