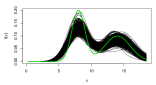
Non parametric statistical estimation

Marcello Chiodi

marcello.chiodi@unipa.it http://dssm.unipa.it/chiodi

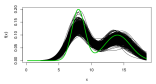
Dipartimento di Scienze Economiche Aziendali e Statistiche (SEAS)
Università di Palermo

Stuttgard, February 2019



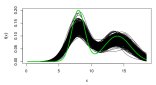
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Outline I



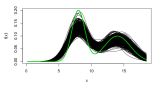
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Outline II



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Outline III



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

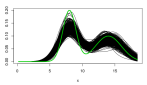
Regression function for random vectors

- The regression function is used **to express the dependence of a random variable Y** on k random variables $\{Z_1, Z_2, \ldots, Z_j, \ldots, Z_k\}$.
- It is *the expected value of the distribution of* $\mathbf{Y}$ *conditional to fixed values of the* $\mathbf{Z}$ *components:*

$$f(z_1, z_2, \ldots, z_k) = \mathrm{E}\left[\mathbf{Y} | Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k\right]$$

- This is just one of the possible ways to define the dependence of a random variable $\mathbf{Y}$ from others k random variables $\{Z_1, Z_2, \ldots, Z_j, \ldots, Z_k\}$
- For $k = 1$ we have:

$$f(z) = \mathrm{E}\left[\mathbf{Y} | Z = z\right]$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Example

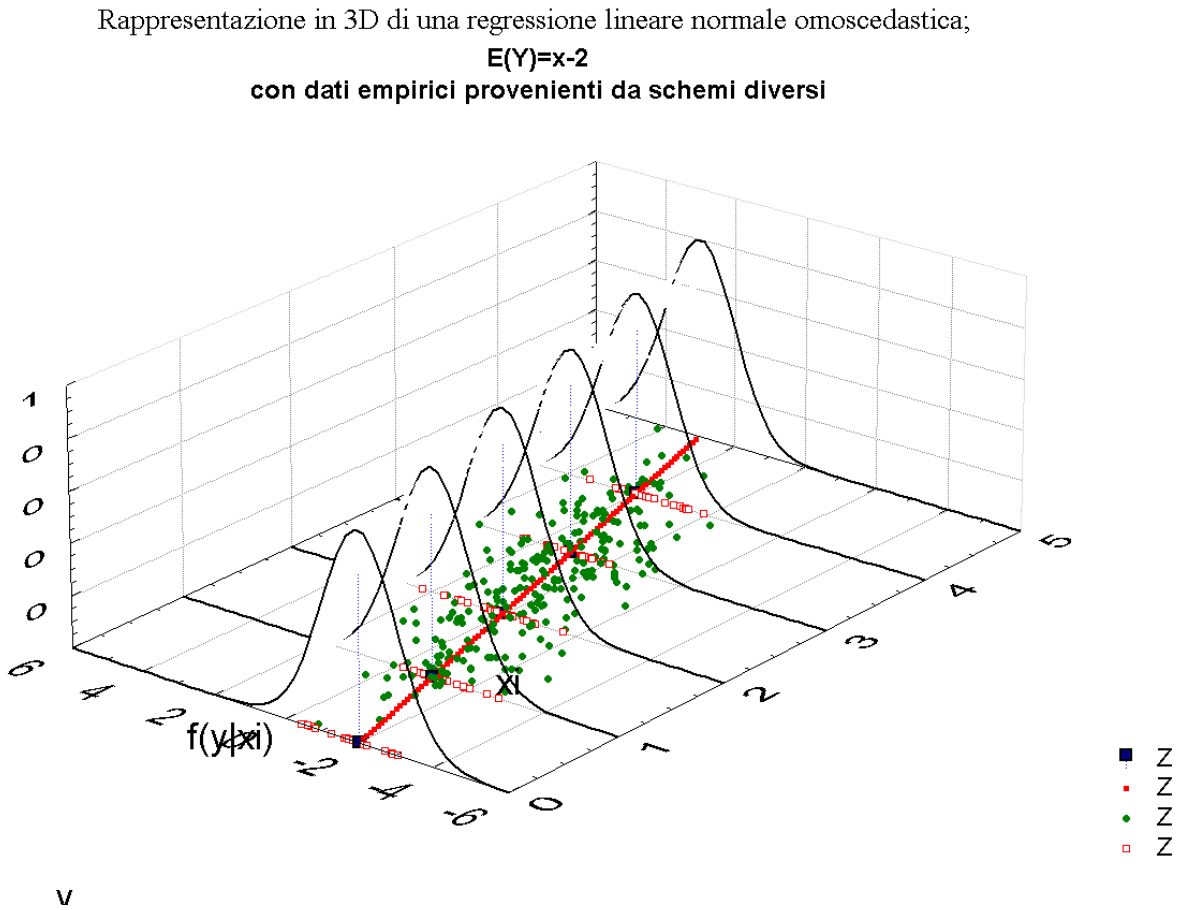
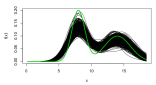


Figure: linear regression model



Example

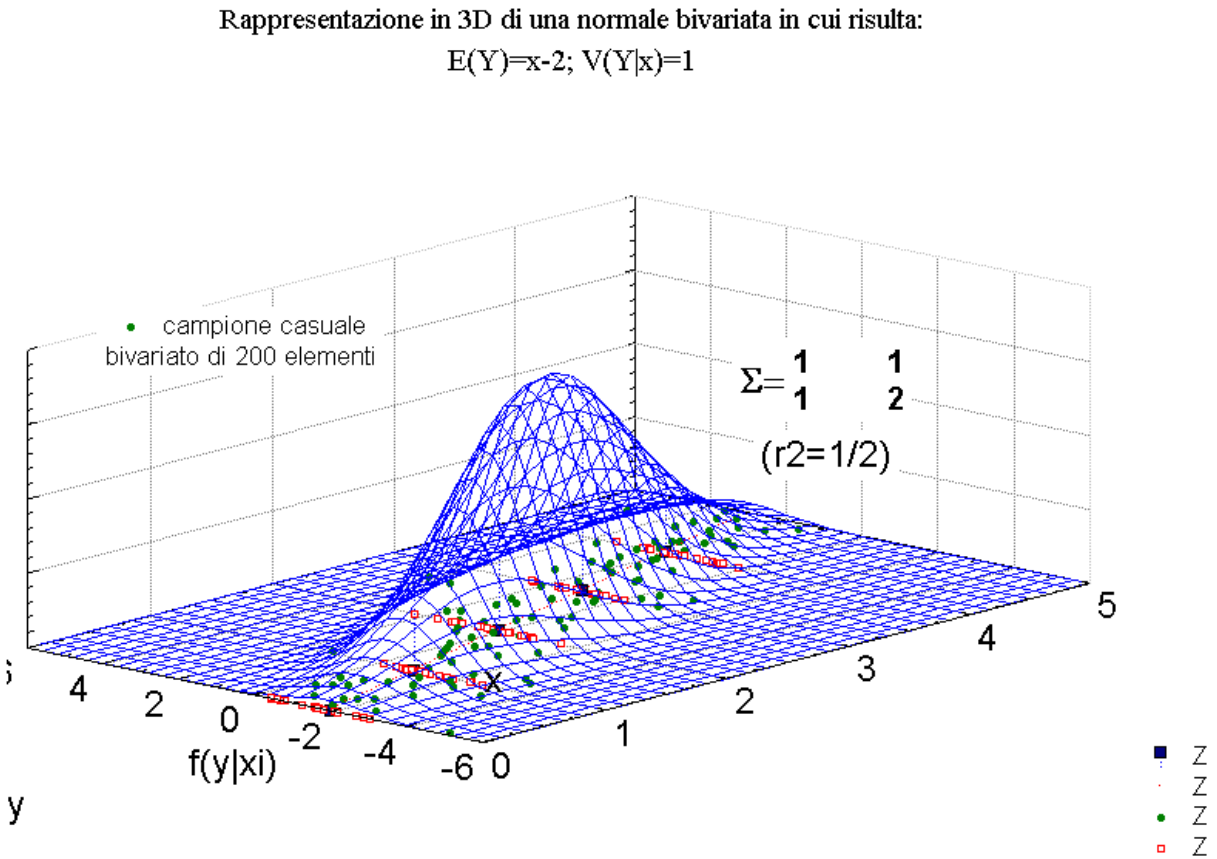


Figure: linear regression model

Parametric approaches

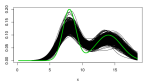
We have some information about the data generator model, that suggest usa possible regression function.

$$f(z;\boldsymbol{\beta}) = \beta_0 + \beta_1\ z$$

$$f(z;\boldsymbol{\beta}) = \beta_0 + \beta_1\ z + \beta_2\ z^2$$

$$f(z;\boldsymbol{\beta}) = \beta_0 + \beta_1\ z^{\beta_2}$$

We have to estimate parameters β

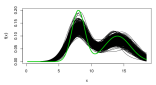


dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Review on regression: multiple linear, polynomial and weighted regression

Before starting with local polynomial regression, it is necessary a brief review on the study of linear models

- review on the board
- Multiple linear regression and linear models
- Polynomial regression
- Weighted linear regression



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Review on regression: multiple linear, polynomial and weighted regression

Multiple linear regression

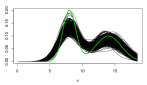
- Linear models (full rank); if:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

then the Least squares solution to minimize $[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]^{\mathrm{T}}$ [y -X $\boldsymbol{\beta}$] is given by:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$



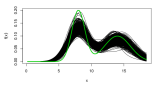
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Review on regression: multiple linear, polynomial and weighted regression

Multiple linear regression

- Polynomial regression: with $x_{ij} = z_i^{j-1}$
- Weighted linear regression:
  (if $\mathrm{V}[\epsilon] = \mathbf{\Sigma}$ then we have to minimize $\mathbf{Q} = (\mathbf{y} - \mathbf{X}\hat{\beta})^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta})$)

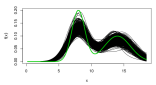
$$\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{y}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Nonparametric approaches

- When we do not have any prior information about the data generator multivariate model, a rigorously parametric approach can not be followed: in this case it is better to let data speak for themselves to suggest the possible regression function.
- If we deal with one or two variables, graphical tools could be useful; this can not be true for $k > 2$.
- Obviously the considerations on **curse of dimensionality** still hold.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Repeated observations: example

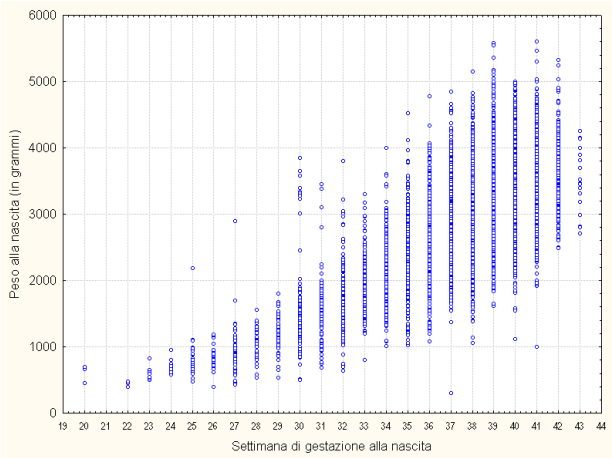
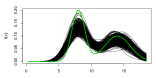


Figure: Scatterplot for two quantitative variables: x = gestational week at the birth; y = weight at the birth

Univariate case $k = 1$: in fig. weight at the birth of 26039 children are reported at the y-axes, (dependent variable); correspondent gestational week is reported at the x-axes (explicative variable). Since there are few distinct values , repeated observation of $\mathbf{Y}$ are observed in correspondence of each x_j.



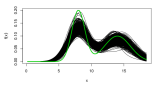
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Conditioned distribution estimation

- It is clear that the average weight increases as the gestational week ;
- for each of the h distinct values of the conditioning variable x_j we can get a distribution of values of $\mathbf{Y}|x_j$, each with mean M_j and variance S_j^2 :

$$M_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j} \qquad S_j^2 = \frac{\sum_{i=1}^{n_j} (y_{ij} - M_j)^2}{n_j} \quad j = 1, 2, \ldots, h$$

- also quantile regression could be used for this purpose....



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Box-plot example

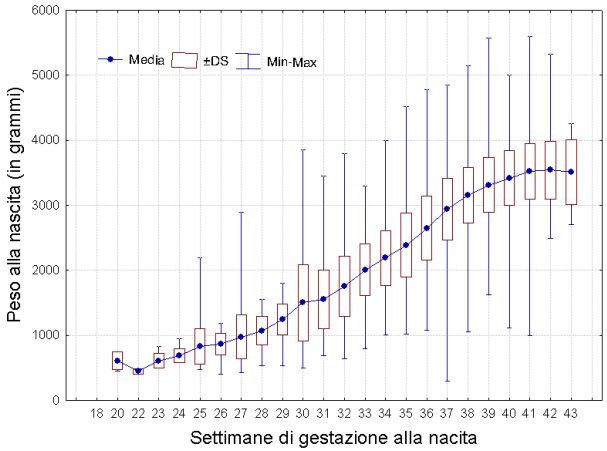
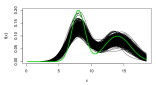


Figure: distribution of $y =$ weight at the birth in correspondence of $x =$ gestational week at the birth. The range, the interval $M_j \mp S_j$, the conditional mean M_j and the corresponding jointing regression line are represented.

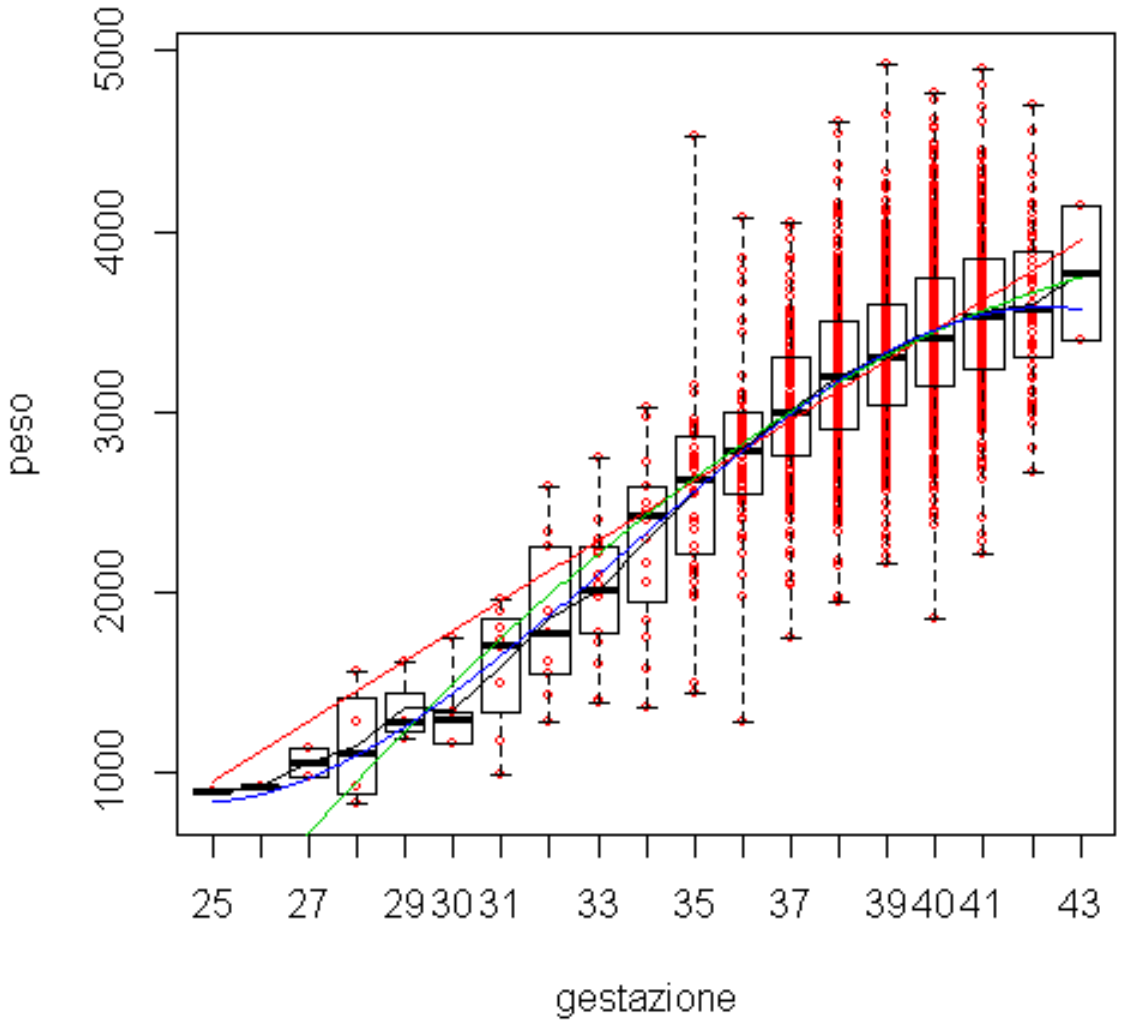
This information is summed up in a box-plot, where for each x_j the range of the conditional distribution, the interval $M_j \mp S_j$, and the conditional means are plotted

The line that joints the conditional means M_j is the regression line!



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Example

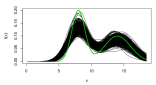

peso
1000
2000
3000
4000
5000
25
27
29
30
31
33
35
37
39
40
41
43
gestazione

Regression line

- It is the Ordinary least Squares (OLS) estimate, since it minimizes the sum of squares of the residuals from any function of the values x_j, $(j = 1, 2, \ldots, h)$, denoted by $g(x_j)$

$$Res(g) = \sum_{j=1}^{h} \sum_{i=1}^{n_j} (y_{ij} - g(x_j))^2 = \sum_{j=1}^{h} \sum_{i=1}^{n_j} [(y_{ij} - M_j) + (M_j - g(x_j))]^2$$

(sum and subtract M_j and then develop the square of the binomial)



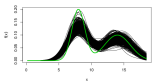
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Classical residual deviance decomposition

$$Res(g) = \sum_{j=1}^{h}\sum_{i=1}^{n_j}(y_{ij} - M_j)^2 + \sum_{j=1}^{h} n_j(M_j - g(x_j))^2$$

1. let $g(x_j) = M_j$, the second term vanishes, and we get the minimum of $Res(g)$ given by:

$$Res(g) = \sum_{j=1}^{h}\sum_{i=1}^{n_j}(y_{ij} - M_j)^2$$



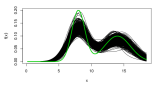
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Regression deviance

The term $\sum_{j=1}^{h}\sum_{i=1}^{n_j}(y_{ij}-M_j)^2$ is named *deviance within groups* and represents the lowest bound if we look functions of x_j that are *optimal in sense of OLS*.
The quantity:

$$\eta_{yx}=1-\frac{\sum_{j=1}^{h}\sum_{i=1}^{n_j}(y_{ij}-M_j)^2}{n\ S_Y^2}$$

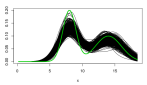
is the correlation ratio and measures the intensity of the dependence in mean of **Y** from **X** . Otherwise, how well the regression line sums up the conditional distribution of **Y** .



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Parametric and nonparametric regression

- In parametric regression we usually look not for the true regression function $f(\mathbf{z})$, but some approximations $g(\mathbf{z}, \boldsymbol{\beta})$, where $g(\cdot, \boldsymbol{\beta})$ belongs to a family of functions that depend from a vector of parameters $\boldsymbol{\beta}$.
- In nonparametric regression, instead, we avoid any assumptions about the functional form of the dependence;
- The functional form $f(\mathbf{X}, \boldsymbol{\beta})$ is not expressed: we estimate $\mathrm{E}[\mathbf{Y}_i|\mathbf{x}_i]$ (in *nonparametric way*), and then $f(.)$ is evaluated. If $k = 1, 2$ this procedure can be useful to start for the choice of the function type, or of the polynomial degree, etc.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

non parametric regression example (1)

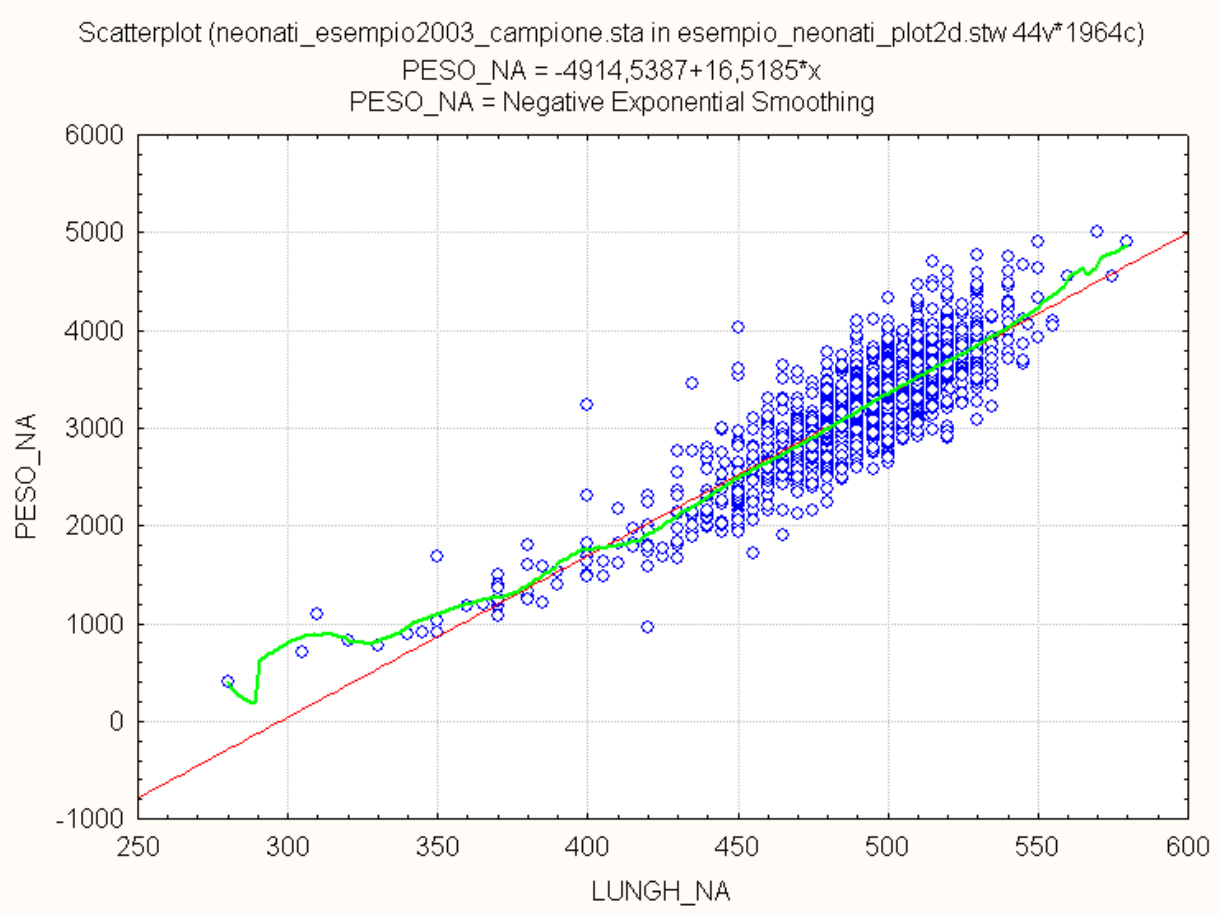
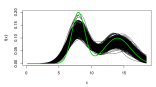


Figure: non parametric regression example for the estimation of dependence in mean of Y from X



non parametric regression example (2)

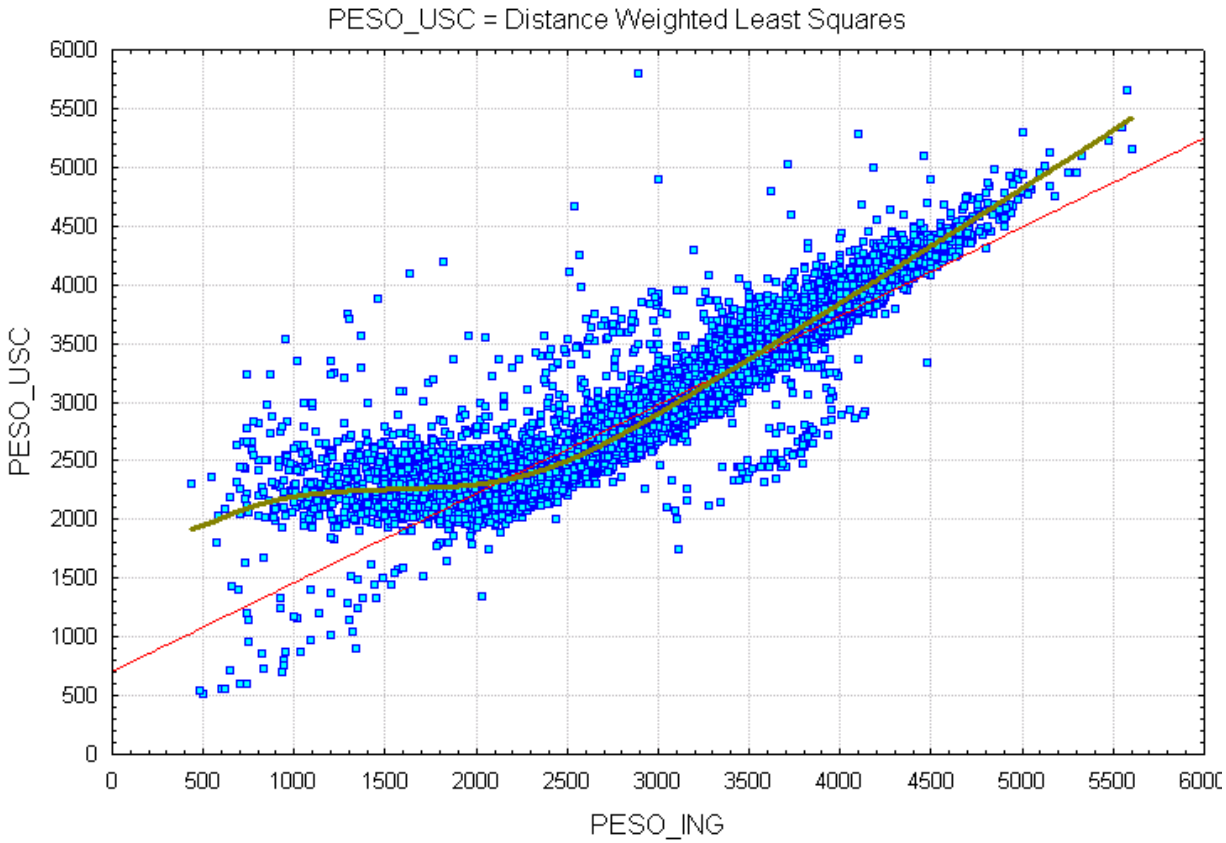


Figure: non parametric regression example

Non parametric regression with 2 explanatory variables -1

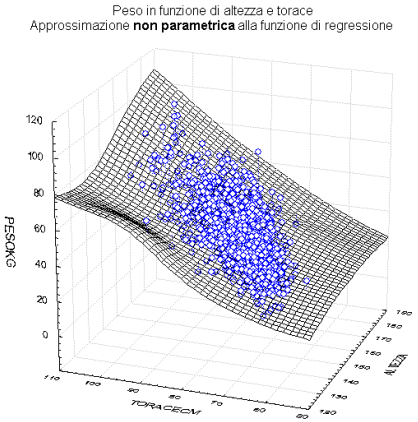


Figure: non parametric regression surface

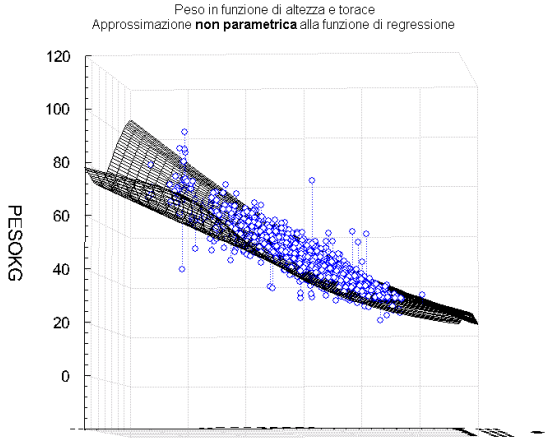
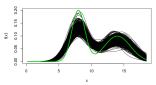


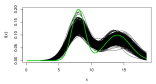
Figure: non parametric regression surface: another point of view



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Non parametric regression with 2 explanatory variables -2

(not printed)



dipartimento di scienze economiche aziendali e statistiche

department of economics business and statistics

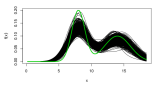
Local approximations

- These techniques are typical of an *exploratory phase* of data analysis when we don't know, at least with a good approximation, which is the shape of the relation between the explanatory variable and the dependent variable.
- We try to approximate the regression function locally, for each value of x

$$\hat{y}(x) \approx \mathrm{E}\left[y|x\right]$$

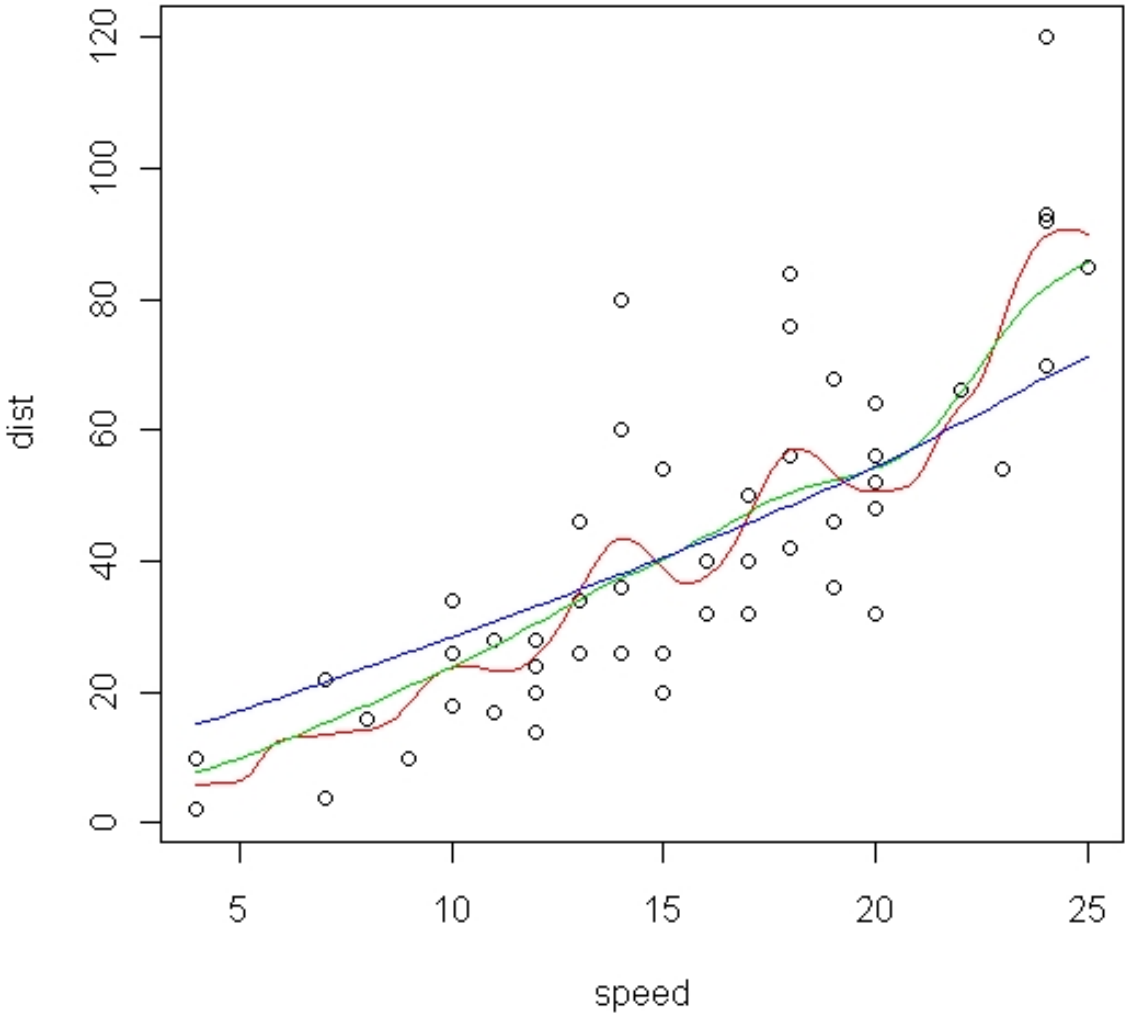
(in absence of repeated observations or if we want to obtain a more regular function of the regression line)

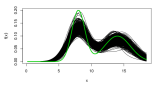
- if we have only an explanatory variable the best way to obtain information about the relation is to make a graphical analysis (*we can fit a polynomial of order 1-2-3*)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

How much should we smooth?





dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

other examples

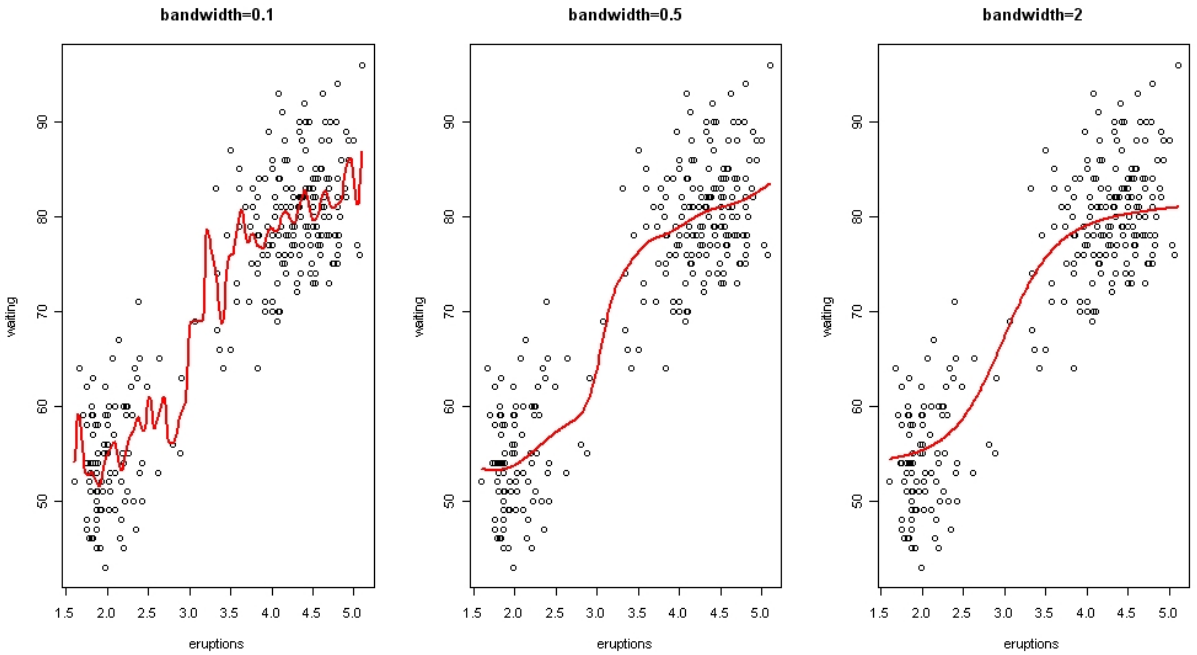
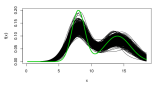
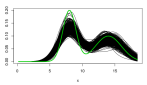


Figure: What is the best choice????



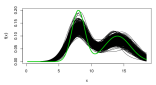
From partial averages to kernel estimators

- We return to the formula of partial averages of $j-$th class, and we make similar operations used for the approximation of incremental ratio. We already used these operations to introduce kernel estimators for densities.
- Let us rewrite the conditional average of the observations of $j-$th class C_j:
- $$M_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad \begin{cases} x_i \in C_j \Rightarrow w_i = 1 \\ x_i \notin C_j \Rightarrow w_i = 0 \end{cases}$$



From partial averages to kernel estimators

- see examples on graphs
- We can now change conditions on the weights: to obtain the partial average each observation weighs 1 if it belongs to that class, 0 otherwise;
- Instead to use $0, 1$ values for weights, if we want to estimate the conditional mean of Y for a particular value of x, we can use the same expression of a partial mean with weights varying in inverse proportion of the distance of each observation x_i from x.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Weighted mean regression

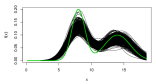
$$M_y(x) = \frac{\sum_{i=1}^{n} w_i(x - x_i)y_i}{\sum_{i=1}^{n} w_i(x - x_i)}$$

`similar to moving averages`

For instance for the weight we can choose the function $w(\cdot)$):

$$w(x - x_i) = \exp\left[-\frac{(x - x_i)^2}{2h^2}\right]$$

where h is the smoothing parameter.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Moving averages and kernel

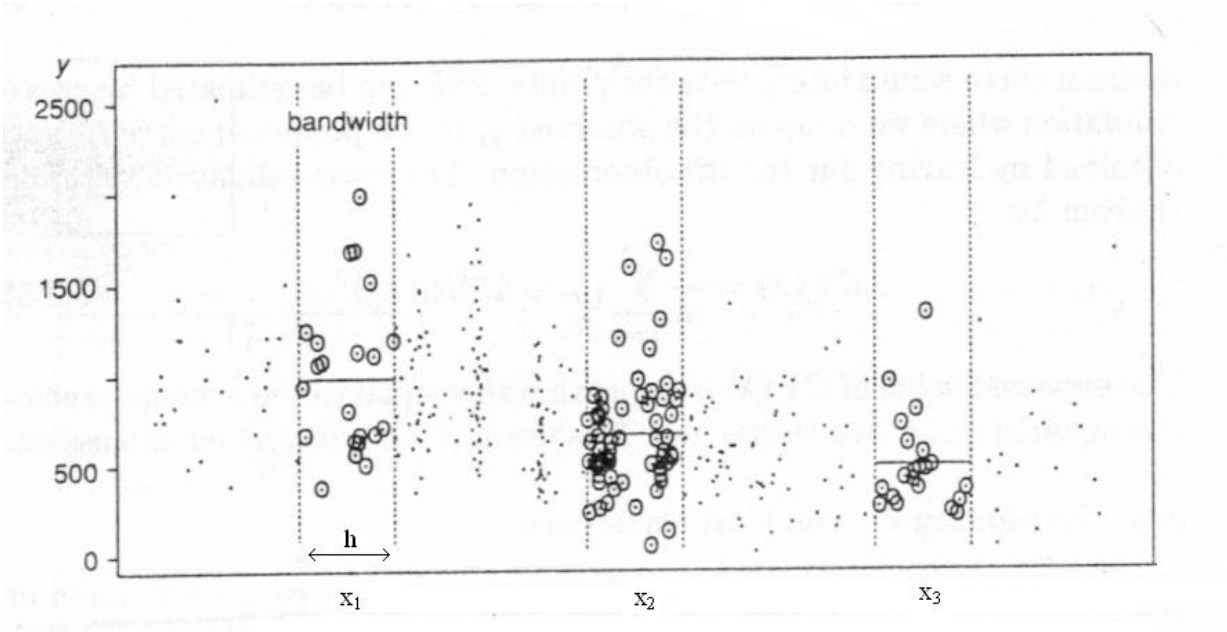
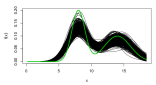


Figure: moving averages: analogies with passing from histogram to kernel

Choice of the bandwidth (smoothing parameter)

(Discussion)

- Cross-validation (generalization of the estimate of the variance of the prediction error)
- Penalized Likelihood (hints on Bayesian approach)
- Not excessive curvature



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Trade-off between two needs

In the choice of an unavoidable smoothing parameter h we will have to balance between two opposite needs:

Geometrical view

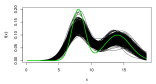
(approximation to observed points) Trade off between:

- fitting (line should be clos to points)
- smoothing (line not too rough)

Statistical point of view

(approssimation to a regression function) Trade off between:

- Low bias (Expected value of $\hat{y}(x)$ close to $\mathrm{E}[y|x]$)
- Low variance (Low variability of $\hat{y}(x)$ around $\mathrm{E}[\hat{y}(x)]$)

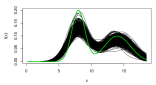


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Curvature measures

We must talk here about **curvature** and its measures

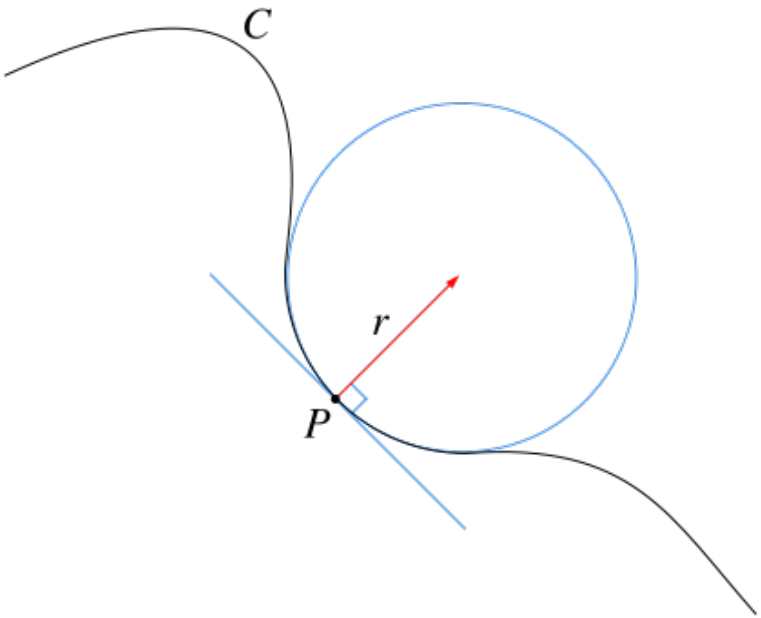
- Measures based on second derivatives
- length of a curve
- Other definitions connected with the splines



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Misure of curvature

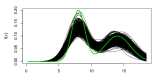
R radius of the circunference tangent to a curve $c(x, y)$ in a point P, where it has the same tangent straight line



- the inverse of the radius of the circunference is the local curvature in P
- if we have an explicit form $y = f(x)$ then:

$$\frac{1}{R} = \frac{f''(x)}{(1 + f'^2(x))^{\frac{3}{2}}}$$

- that can be approximated with $f''(x)$ if $f'(x)$ is small compared to 1.

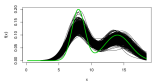


dSEAS dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

3 different approcches to non parametric regression

In simple situations, there are 3 different approaches to non parametric regression and each approach can lead to different estimators but with numerically similar results

- Direct definition starting from regression function as the expected value of the conditional distribution.
- Local polynomial regression (remind a Taylor series expansion)
- Penalized least squares (splines functions) as approximation problem more than regression problem: very flexible and analytically useful



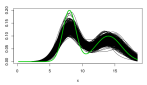
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

First approach: conditional distribution estimation

- Let us now suppose to estimate a regression function $\mathrm{E}[Y] = g(\mathbf{x})$ starting from informations drawn by a sample of n observations $y_i, \mathbf{x}_i$ with n $i = 1, 2, ..., n$.
- Let us start with a relation of simple regression

$$\mathrm{E}[Y_i] = g(x_i) \qquad \text{with} \qquad Y_i = g(x_i) + \epsilon_i$$

- We assume x is continuous and x is a realization of a random variable X
- However it's necessary to assume that the marginal and conditional distributions for each x have density.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Expected value of conditional distribution

- Let us start from the definition of regression function:

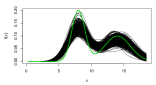
$$g(x) = \mathrm{E}\left[Y|X = x\right]$$

- where the distribution of Y conditional on $X = x$ has density:

$$f_Y(y|X = x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

- In order to evaluate the expected value of conditional distribution of Y, we can use a non parametric estimation of density of this distribution.

$$\hat{f}_Y(y|X = x) = \frac{\hat{f}_{XY}(x, y)}{\hat{f}_X(x)}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

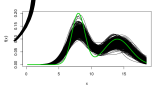
Expected value of a kernel estimator for a conditional distribution

In the last expression let us use normal kernel estimators to estimate the density functions with indipendent components, and then let us consider the expected value (with respect to y).

$$\hat{f}_Y(y|X=x) = \frac{\hat{f}_{XY}(x,y)}{\hat{f}_X(x)} \Rightarrow \hat{E}(Y|X=x) = \int y \frac{\hat{f}_{XY}(x,y)}{\hat{f}_X(x)} dy =$$

$$= \int y \frac{\frac{1}{nh_xh_y}\sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right) K\left(\frac{y-y_i}{h_y}\right)}{\frac{1}{nh_x}\sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right)} dy =$$

$$= \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right) \int y \frac{1}{h_y} K\left(\frac{y-y_i}{h_y}\right) dy}{\sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right)} = \frac{\sum_{i=1}^n y_i K\left(\frac{x-x_i}{h_x}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h_x}\right)}$$



Nadaraya-Watson estimator

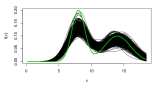
- We again obtain an estimation based on moving averages, with decreasing weights in function of the distance of x from each x_i:

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} w(x - x_i) y_i}{\sum_{i=1}^{n} w(x - x_i)}$$

  Nadaraya-Watson estimator (1964)
- where the weights $w(x - x_i)$ are decreasing function of $x - x_i$;
- in this form they are generalitation of moving averages).
- For instance for the weight we can choose the function $w(\cdot)$):

$$w(x - x_i) = \exp\left[-\frac{(x - x_i)^2}{2h^2}\right]$$

  where h is the bandwidth (smoothing parameter)



dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

Some features of Nadaraya-Watson estimator

(usually called **kernel estimator**)

Advantages:

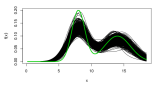
- this approach is very simple and it is an immediate extension of the kernel estimator for univariate densities (linear estimator and hat matrix)

Handicaps:

- It is not good in the extremes of the field of observation of x
- It is impossible to use for extrapolation, because

$$\min(y) \leq \hat{y}(x) \leq \max(y) \quad \forall x$$

- If h diverges, we obtain a horizontal line
- while it would be better to obtain the least squares regression line as limit (or something that can approximate a relationship of dependence!)



dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

Some examples of Nadaraya Watson estimates

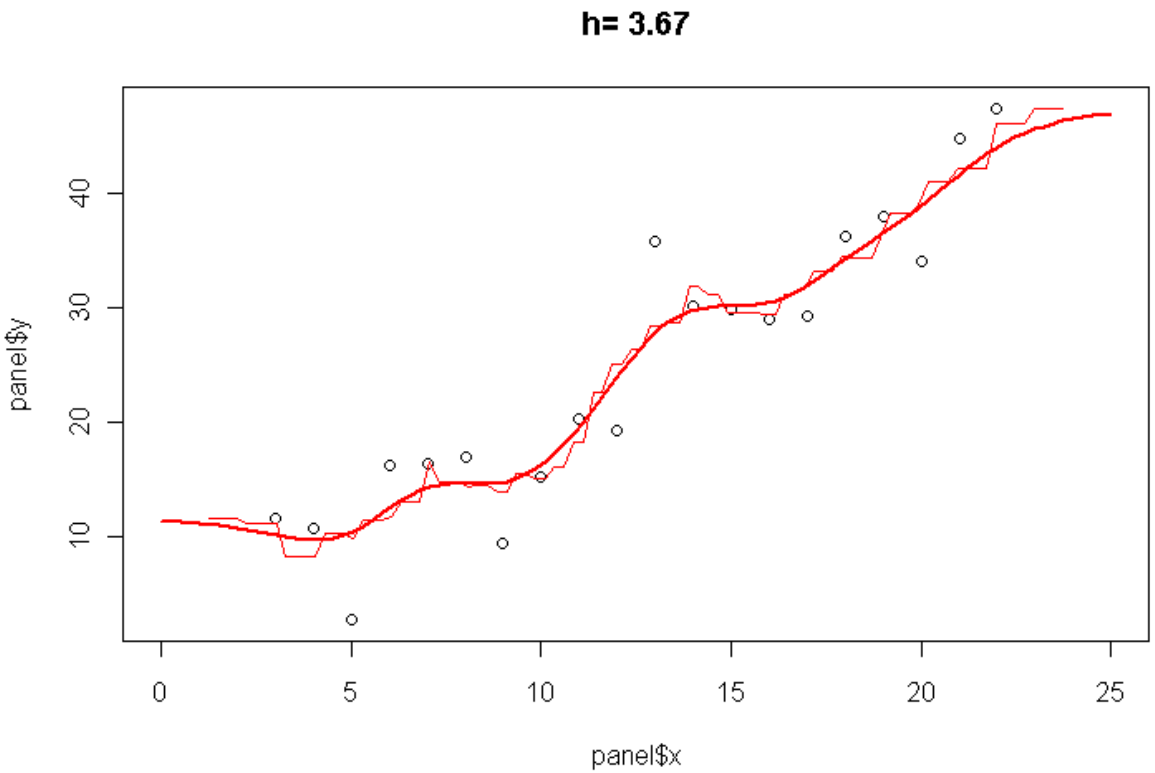
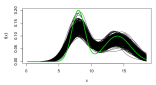


Figure: Kernel estimate of a regression: uniform and normal kernel



Some examples of Nadaraya Watson estimates

Interactive code

`code R`

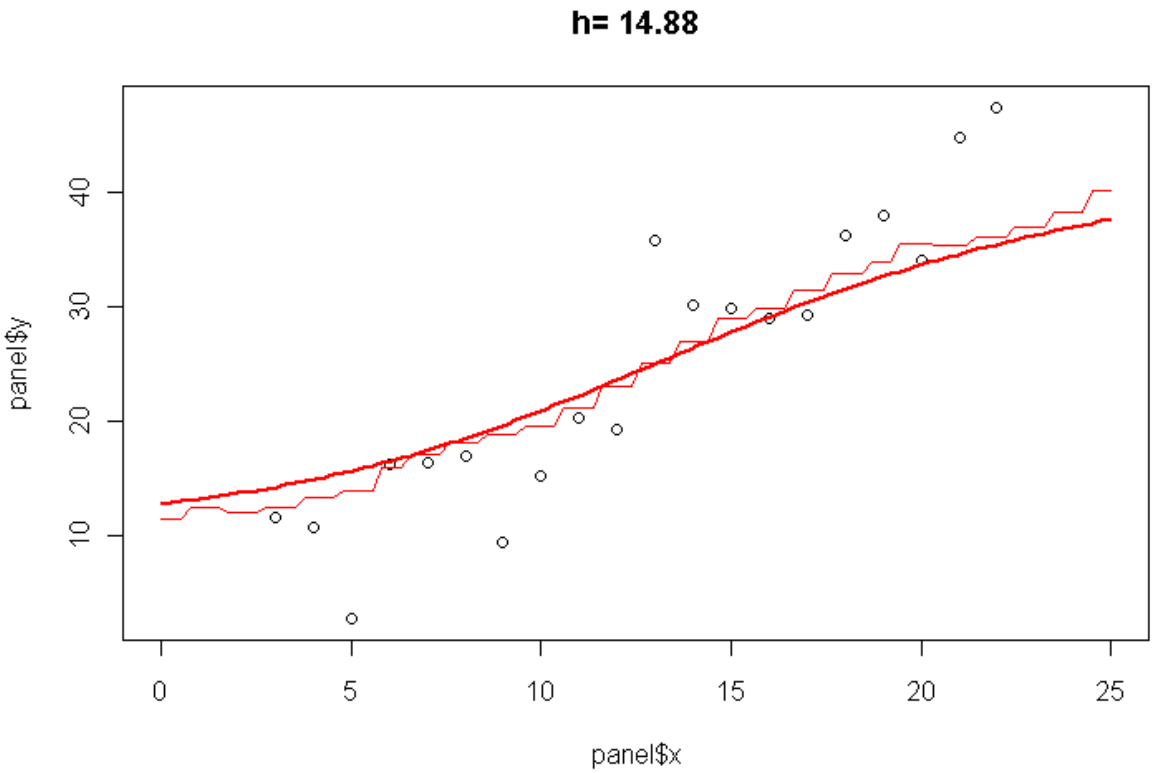
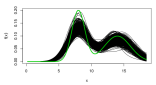


Figure: Kernel estimate of a regression: uniform and normal kernel

Review on regression: multiple linear, polynomial and weighted regression

Before starting with local polynomial regression, it is necessary a brief review on the study of linear models

- review on the board
- Multiple linear regression and linear models
- Polynomial regression
- Weighted linear regression



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Review on regression: multiple linear, polynomial and weighted regression

Multiple linear regression

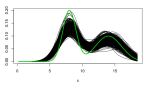
- Linear models (full rank); if:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

then the Least squares solution to minimize $[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]^{\mathrm{T}}$ [y -X $\boldsymbol{\beta}$] is given by:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$



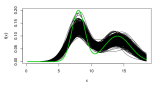
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Review on regression: multiple linear, polynomial and weighted regression

Multiple linear regression

- Polynomial regression: with $x_{ij} = z_i^{j-1}$
- Weighted linear regression:
  (if $\mathrm{V}[\epsilon] = \mathbf{\Sigma}$ then we have to minimize $\mathbf{Q} = (\mathbf{y} - \mathbf{X}\hat{\beta})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$)

$$\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{y}$$



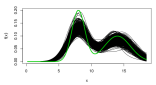
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Local polynomial regression: general solution

Let us search a local approximation in x of degree p, near $x*$:

$$m(x, x*) = a_0(x*) + \sum_{j=1}^{p} a_j(x*)(x - x*)^j$$

It corresponds (in general) to an approximation *smooth* of a Taylor series.



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Local polynomial regression: general solution

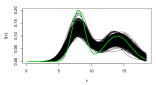
- we have to minimize:

$$\sum_{i=1}^{n} w\left(\frac{x_i - x*}{h}\right)(m(x_i, x*) - y_i)^2$$

that is to say

$$\sum_{i=1}^{n} w\left(\frac{x_i - x*}{h}\right)(a_0(x*) + \sum_{j=1}^{p} a_j(x*)(x - x*)^j - y_i)^2$$

- **our approximation in $x*$ will be given by $\hat{a}_0(x*)$**
- the coefficients are obtained solving a problem of **weighted least squares**



dSEAS dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Local polynomial approximation

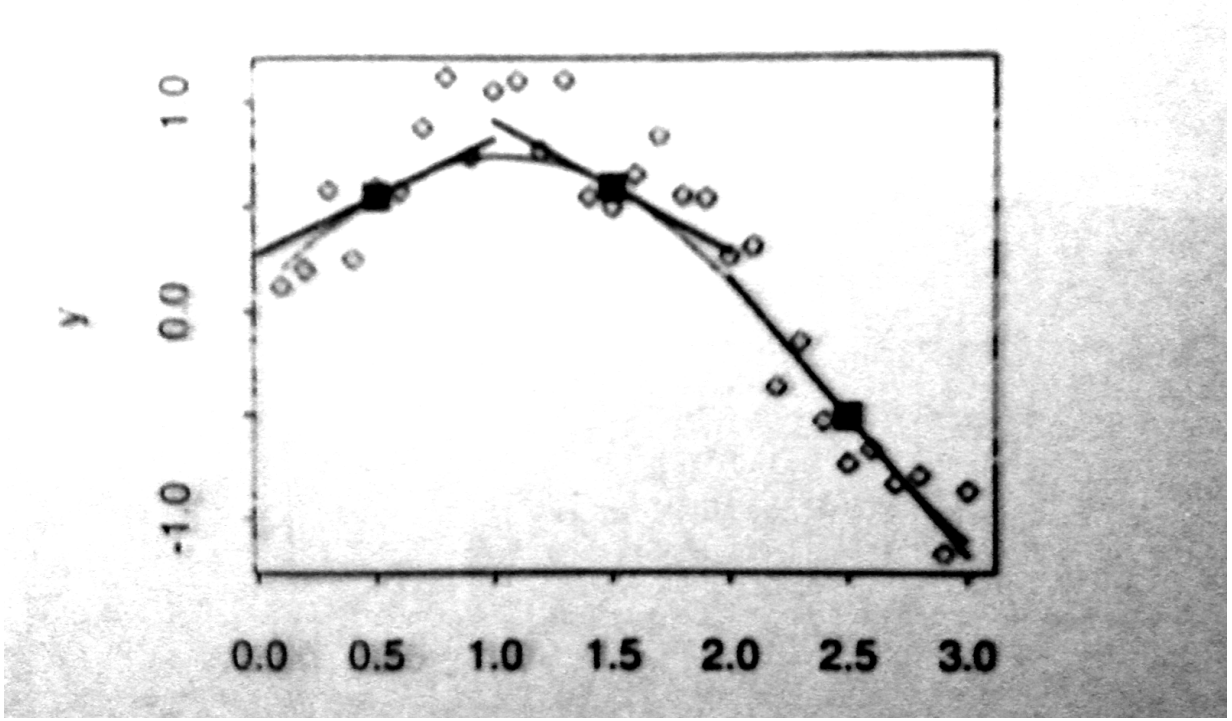
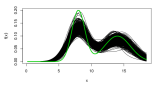


Figure: Example of local approximation through segments of straight lines

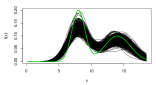


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Local polynomial regression: General features

(see slides or board)
(Draw with sloped break-lines)

- It is good in the extremes of the observed range of x
- For values of p greater than 1 and for low values of h we can have some fluctuations in the observed range
- If h diverges, we obtain least squares line (with $p = 1$)
- It corresponds to a *smooth* approximation of a Taylor series to the k-th term



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Local polynomial regression: general solution

(slides and board) For local approximation in $x*$ of degree p we set:

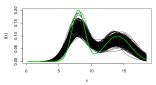
- **W** the (diagonal) matrix of the local weights possibly achieved by a kernel function

$$\{W\}_{ii} = w\left(\frac{x_i - x*}{h}\right), \qquad \text{normalized:} \qquad \sum_{i=1}^{n} \{W\}_{ii} = 1$$

- the matrix of regressors has an usual structure of polynomial regression (centered in x) with generic element:

$$\{\mathbf{X}\}_{ij} = (x_i - x*)^{j-1}$$

(the first column is a vector of 1)



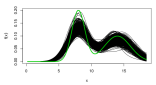
Local polynomial regression: weighted polynomial regression

We have to solve a problem of weighted least squares

$$\min_{\mathbf{a}} \left[\mathbf{y} - \mathbf{X}\mathbf{a}\right]^{\mathrm{T}} \mathbf{W} \left[\mathbf{y} - \mathbf{X}\mathbf{a}\right];$$

It is a problem of weighted least squares (for each $x*$) in which:

- $\mathbf{a}$ is a vector of parameters to estimate (we only need one parameter)
- $\mathbf{X}$ is a matrix $n \times (p+1)$ of regressors (polynomial components depend on $x*$)
- $\mathbf{W}$ is a diagonal matrix $n \times n$ of weights (dependent on $x*$)
- $\mathbf{y}$ is a vector of observed values n



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Local polynomial regression: general solution

approximant polynomial

The approximant polynomial in x has coefficients obtained from the solution of a problem of weighted least squares

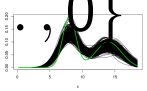
$$\min_{\mathbf{a}} \left[\mathbf{y} - \mathbf{X}\mathbf{a}\right]^{\mathrm{T}} \mathbf{W} \left[\mathbf{y} - \mathbf{X}\mathbf{a}\right];$$

$$\hat{\mathbf{a}} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}$$

we only take the first element (which gives the value of the polynomial just in x^*, according to the adopted notation)

$$\hat{m}(x) = e_1^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}$$

where e_1 is a vector of $p+1$ elements, such that: $e_1^{\mathrm{T}} = \{1, 0, \ldots, 0\}$



dSEAS dipartimento di scienze economiche aziendali e statistiche department of economics business and statistics

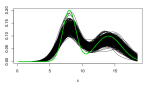
Local polynomial regression: linear estimators

(**W** e **X depend on a particolar value of** $x*$ **!!!**)

- **They are still linear estimators in y**
- with coefficients:

$$\mathbf{q}^{\mathrm{T}}(x*) = e_1^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W} \quad (\sum_{i=1}^{n} q_i(x*) = 1)$$

- such that $\hat{m}(x*) = \mathbf{q}^{\mathrm{T}}(x*)\mathbf{y}$ (weighted averages).
- Nadaraya-Watson estimator is a particular case for $p = 0$
- The most useful values are $p = 1$ e $p = 2$



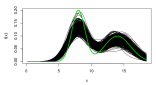
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Examples of polynomial regression for 4 values of p

`codice R`

In the following examples I used this colours:

- $p = 0$ blue
- $p = 1$ black
- $p = 2$ red
- $p = 3$ green



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Examples of polynomial regression for 4 values of p

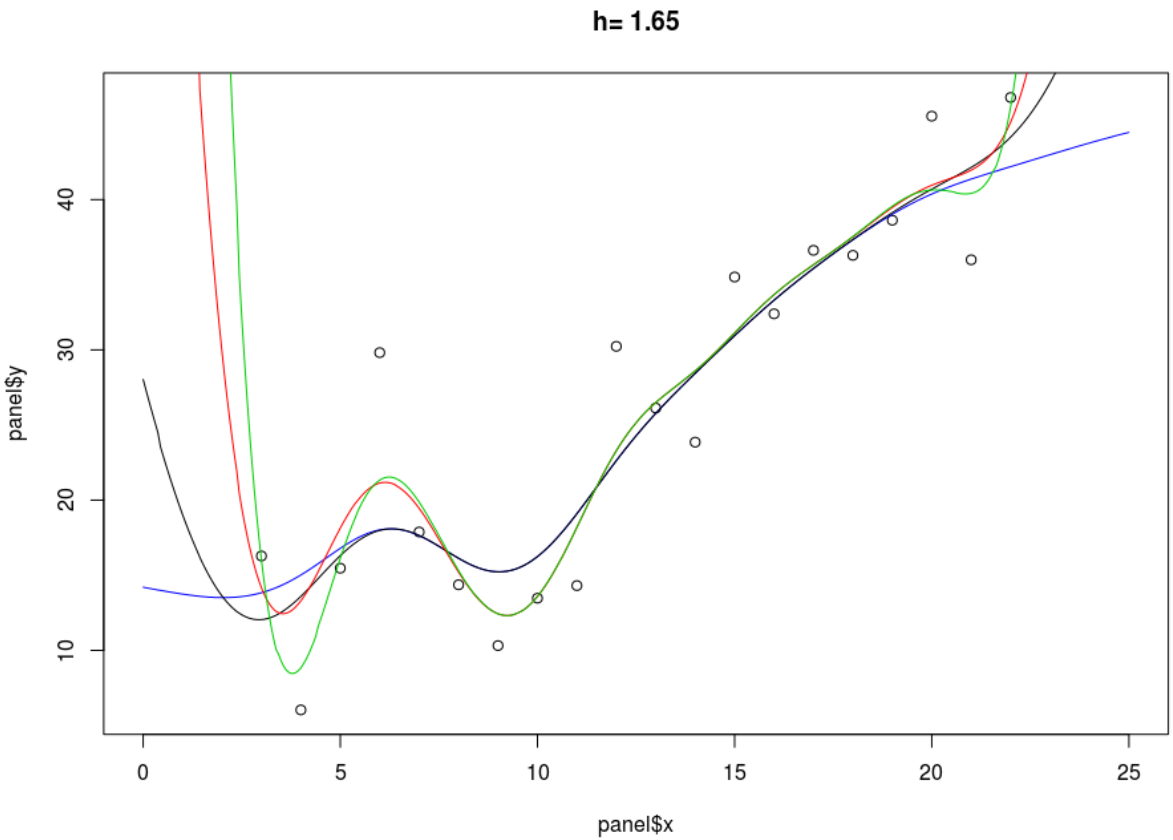


Figure: polynomial regression with $p = 0, 1, 2, 3$

Examples of polynomial regression for 4 values of p

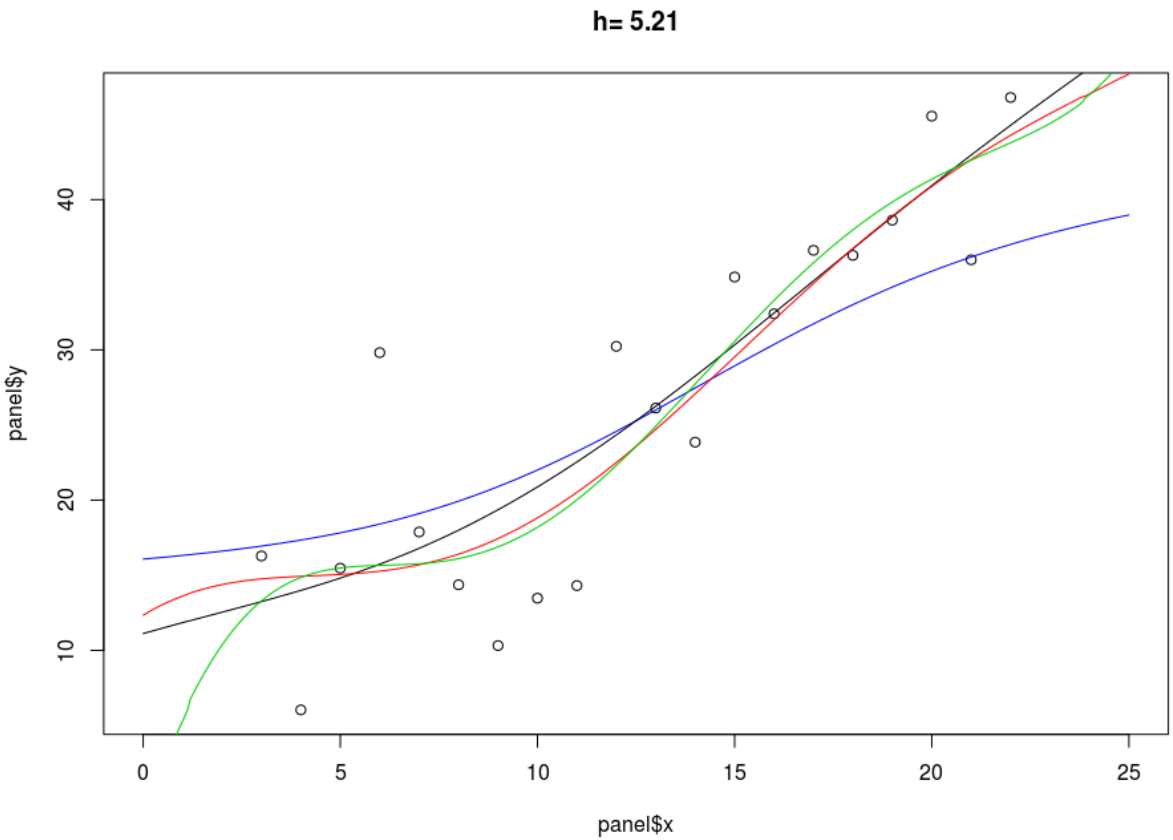


Figure: polynomial regression with $p = 0, 1, 2, 3$

Examples of polynomial regression for 4 values of p

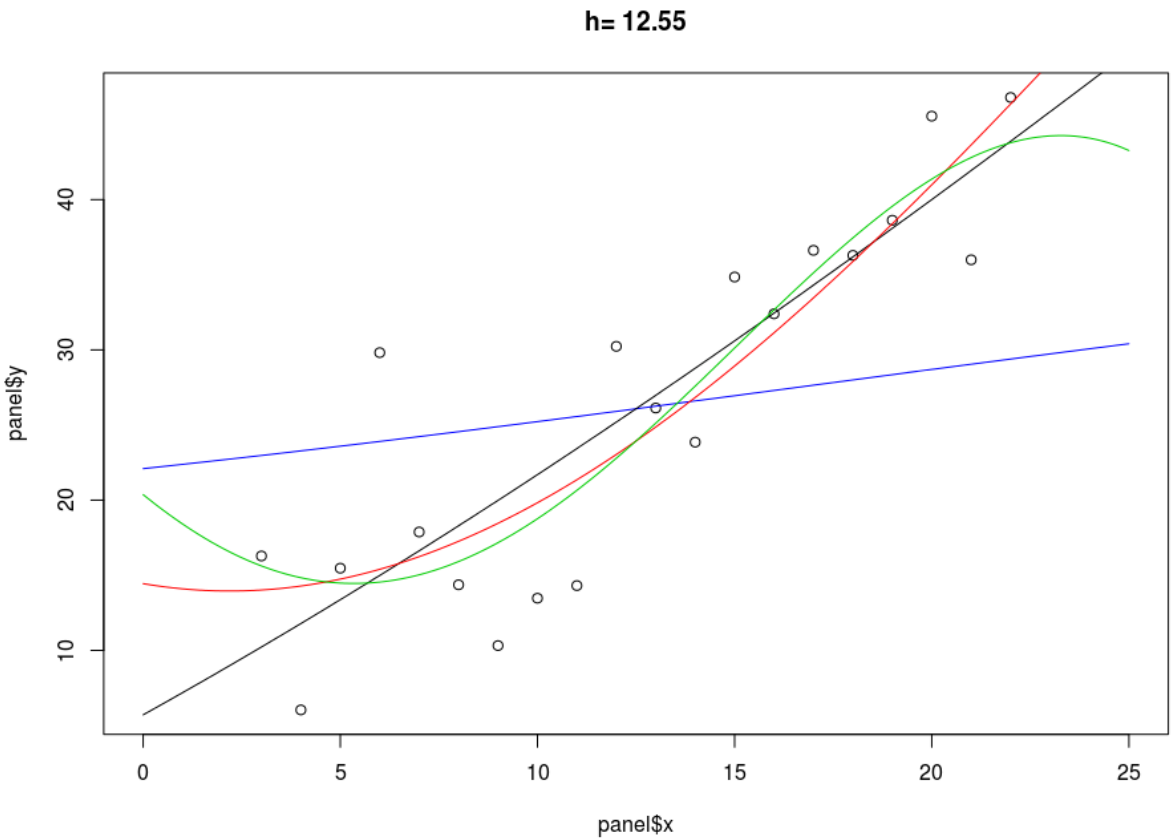
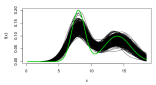


Figure: polynomial regression with $p = 0, 1, 2, 3$



Local polynomial regression: some properties

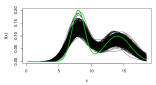
- Since: $\hat{m}(x) = \mathbf{q}^{\mathrm{T}}(x)\mathbf{y}$
- Assume that $Y = m(x) + \varepsilon$, with $\mathrm{E}[\varepsilon] = 0$, $V(\varepsilon) = \sigma^2$e $\mathrm{E}[\varepsilon_i \varepsilon_j] = 0$

$$V(\hat{m}(x)) = (\sigma^2) \sum_{i=1}^{n} q_i^2(x)$$

- The variance is less than σ^2 and it is a decreasing function of h
- Particular cases ($h \to 0$ e $h \to \infty$: what happens to $V(\hat{m}(x))$)
- σ^2 can be estimated from residual deviance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y - \hat{m}(x_i))^2}{n - p - 1}$$

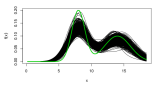
but it is not so useful



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Curvature or roughness measures

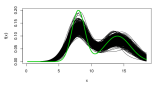
- Measures based on the second derivative
- Length of a curve
- Other definitions linked with the features of splines



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Third approach to non parametric regression

1. Direct definition starting from regression function as expected value of the conditional distribution
2. Local polynomial regression (local appromixations)
3. Penalized least squares (splines)
   More a problem of approximation than a regression problem



Two goals

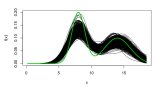
Geometric Aspect (approximation of observed points) Generally we want to find a function $g(x)$ providing

- a good fitting to the points, with a low distance (for now the Euclidean one) from observed points

$$\sum_{i=1}^{n} [y_i - g(x_i)]^2$$

- smoothing (very regular curve) the slope changes slowly so we have a small second derivative in the interval, because $g''()$ measures the velocity of slope changing

$$\int \left[g(x)''\right]^2 \, d\, x$$

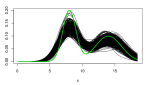


dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

A step behind: linear and polynomial interpolations

Let us talk only about interpolations:
reviews and graphical examples on the board

- features of linear interpolations
- features of polynomial interpolations
- a polynomial of degree p interpolates exactly $p + 1$ points
- Lagrange and Newton Formula
- linear break-lines
- polynomial break-lines



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Polynomial features

Polynomial features

Advantages:

- continuity
- differentiabilty of each order (it is a closed family under the operator derivative)
- computational complexity proportional to p and n
- computability (only additions and multiplications)

Disadvantages:

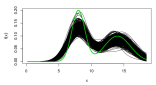
A polynomial of degree p has:

- up to p passages from the origin
- up to $p-1$ points of local minimum or maximum
- up to $p-2$ inflection points

Splines functions

Splines are segments of polynomials connected in order to respect some continuity constraints of some derivatives

- cubic splines
- natural splines
- interpolating splines
- smoothing splines

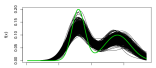


Polynomial Approximations and splines

If we search for sufficeintly regular approximations, then splines functions are a very useful tool, both in univariate and in bivariate case, and can be used in GAM, generalized additive models

Splines functions are particular functions obtained from the composition of segments of polynomials such that the curve is sufficiently smooth and regular without points of discontinuity in changes of segment.

To obtain the parameters of the segments of polynomial, the technique is to impose some constraints to the functions and to their derivatives in the knots



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

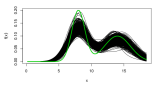
Interpolating splines

Interpolating splines (of degree r)

- n pairs of points (x_i, y_i), $i = 1, 2, \ldots, n$
- $n-1$ polynomial $p_i(x)$ $i = 1, 2, \ldots, n-1$, of degree r
- $(n-1)(r+1)$ parameters

Constraints:

- n constraints $f(x_i) = y_i$ $i = 1, 2, \ldots, n$ *passage by points*
- $(n-2)r$ constraints of continuity $p_i^{(j)}(x_{i+1}) = p_{i+1}^{(j)}(x_{i+1})$ $i = 1, 2, \ldots, n-2$, $j = 0, 1, \ldots, r-1$



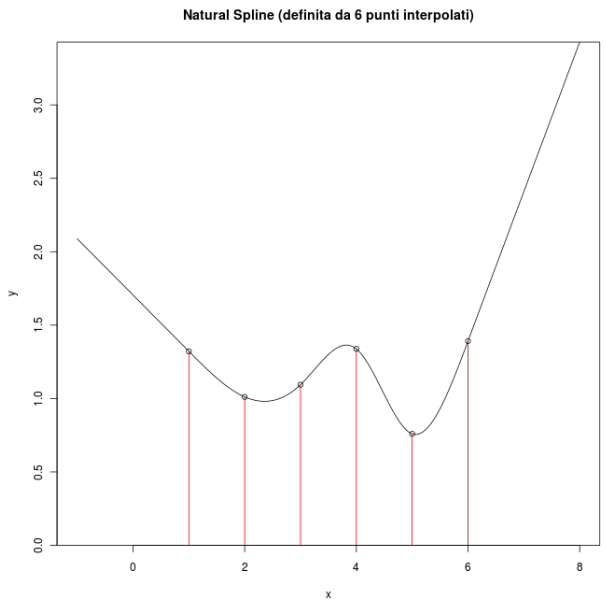
Natural Splines

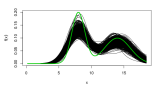
We have:

- $nr - r + n - 1$ parameters
- $nr + n - 2r$ constraints

$r - 1$ *further conditions are necessary* to determine parameters. If $r = 3$

$$\min_b \int_{x_1}^{x_n} (f''(x))^2 dx \Leftrightarrow f''(x_1) = f''(x_n) = 0$$





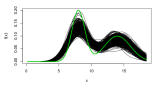
dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Splines

esempi di spline

`R code`

- If the two constraints are the previous one, then we have **natural splines**
- The support can be extended over the whole real axis, maintaining unchanged the **property of minimum curvature**
- (we have to think that the extensione out of the observed range does not change the curvature because we have two straight segments)
- For particular applications we can think about other pairs of constraints



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Natural Splines

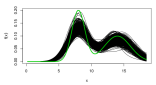
Parameterization on a closed interval:

$$s(x) = a_0 + a_1 x + \frac{1}{12} \sum_{i=1}^{n} b_i |x - x_i|^3$$

$n + 2$ parameters are obtained through $n + 2$ constraints

- interpolation of n pairs of points (x_i, y_i), $i = 1, 2, \ldots, n$;
  n constraints $s(x_i) = y_i \;\; i = 1, 2, \ldots, n$
- $\sum_{i=1}^{n} b_i = 0 \qquad \sum_{i=1}^{n} b_i x_i = 0$;
  2 furthers constraints deriving from the condition of minimum curvature:

$$\min_b \int_{x_1}^{x_n} (f''(x))^2 dx \Leftrightarrow f''(x_1) = f''(x_n) = 0$$



Solution

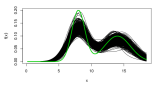
The solution is formally obtained, assuming:

$$\mathbf{R} : \{\mathbf{R}\}_{ij} = \frac{|x_i - x_j|^3}{12}$$

$$\mathbf{Q} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \qquad \mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Then the constraints are satisfied solving the system in $n + 2$ unknown factors $\mathbf{a}, \mathbf{b}$:

$$\begin{pmatrix} \mathbf{R} & \mathbf{Q}^{\mathrm{T}} \\ \mathbf{Q} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ \mathbf{a} \end{pmatrix} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$$



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

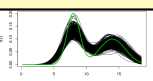
Penalized Least Squares:trade-off between two needs

Searching a functions $g(x)$ such that:

$$g(x): \min_{g(x)} \sum_{i=1}^{n} [y_i - g(x_i)]^2 + \lambda \int [g(x)'']^2 \, d\, x$$

$$(Q(g(\cdot)) = R(g(\cdot)) + P(g(\cdot)))$$

The solution is a natural spline



(with knots in x_i)

Penalized Least Squares:trade-off between two needs

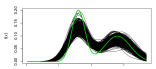
The solution is a natural spline Proof:
Let us suppose that the solutione $\hat{g}(x)$ is not a natural spline; then we can obtain a natural spline $\tilde{g}(x)$ interpolating the points $\hat{g}(x_i)$ for the first part of the contribution to PLS, we have:

$$R(\tilde{g}(\cdot)) = R(\hat{g}(\cdot))$$

But $\tilde{g}(x)$ is a natural spline, so it is the interpolating of the points $\hat{g}(x_i)$ that minimizes the curvature $P(g(\cdot))$, so

$$P(\tilde{g}(\cdot)) \leq P(\hat{g}(\cdot)) \qquad \text{and then:} \qquad Q(\tilde{g}(\cdot)) < Q(\hat{g}(\cdot))$$

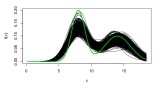
This is in contradiction with the assumption that $\hat{g}(x)$ is the best solution, so $\hat{g}(x)$ it has to be a natural spline



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

Particular cases

- $lambda = 0 \Rightarrow \hat{g}(x)$ is a natural spline interpolating data
- $lambda \to \infty \Rightarrow \hat{g}(x)$ tends to Least Squares straight line



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

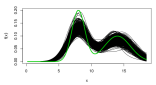
Solution of the PLS problem

Solution of the PLS problem can be obtained solving a linear problem of penalized least squares replacing the penalization term
we know that the problem is now to find $n+2$ coefficients of a natural spline
It can be proved the penalization term can be written (adopting the same symbols used before for the coefficients of a natural spline $\hat{g}(x)$):

The penalization is expressible in compact form

$$\lambda \int \left[g(x)''\right]^2 \; d\, x = \lambda \mathbf{b}^{\mathrm{T}} \mathbf{R} \mathbf{b}$$

(through a boring proof by integration by parts)



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

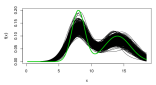
How much smoothing?

How much smoothing?

λ can be chosen by means of cross-validation.
But graphical interactive choices are of practical utility.

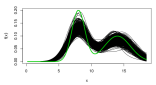
slides

cercare slides in italiano



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics

GAM: Generalized additive models



dSEAS
dipartimento di scienze economiche aziendali e statistiche
department of economics business and statistics