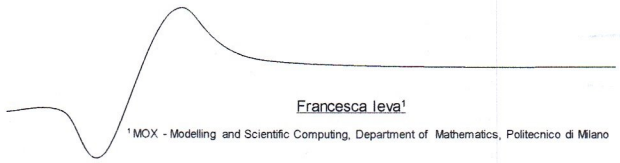
(till now it was testing and CIs, from now it'll be modelling, fitting and prediction)

Nonparametric Regression

We are now PREDICTION-ORIENTED

Nonparametric Statistics
AA 2020-2021



Francesca Ieva[1]

[1] MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano

Outline

- Why Nonparametric Regression
- Dataset & Examples
- Nonparametric Simple Regression
  - Binning and Local Averaging
  - Kernel estimation and Weighted Local Averaging
  - Local Polynomials
  - Splines
    - Regression splines
    - Smoothing splines
- Nonparametric Multiple regression: GAMs
- References

Generalized Additive Model

Outline

- Why Nonparametric Regression
- Dataset & Examples
- Nonparametric Simple Regression
  - Binning and Local Averaging
  - Kernel estimation and Weighted Local Averaging
  - Local Polynomials
  - Splines
    - Regression splines
    - Smoothing splines
- Nonparametric Multiple regression: GAMs
- References

Why Nonparametric Regression

- Regression analysis traces the average value of a response variable Y as a function of several predictors $\boldsymbol{X} = (X_1,\dots,X_p)$
- The object of regression analysis is to estimate

$$\mu|\mathbf{x} = f(\mathbf{x}) \quad \text{where} \quad \mu = E[Y]$$

- Typical (very strong) assumptions:
  - Linear relationship between μ and $\mathbf{x}$ => $f(\mathbf{x})$ is of the form $ax+b$
  - Omoschedasticity
  - Normal distribution of the errors (then of the responses)
  - Independent observations (errors $\perp\!\!\!\perp$ $\Rightarrow$ observations $\perp\!\!\!\perp$)
- There are many ways in which such assumptions may go wrong
  - In time series errors may not be independent
  - Conditional variance of Y may not be constant (eteroSchedasticity)
  - Conditional distribution of Y may not be Normal

Y = output of interest
X = input

Remember: when we talk about expected values, we always intend conditional to data
($E[Y|\mathbf{x}]$)

Why Nonparametric Regression

- Nonparametric regression **relaxes the assumption of linearity**, substituting the much **weaker assumption of smooth regression function *f*** such that

$$Y = f(X) + \varepsilon \qquad \mathrm{E}[\varepsilon] = 0$$

- The goal of nonparametric regression is to construct an **estimate of *f*** from i.i.d. samples $(y_1, x_1),\ldots,(y_n, x_n)$ having the same joint distribution, **not assuming any parametric form for *f***.

Note: we usually assume X as being fixed, i.e. we are conditioning on a given realization of **X=x**, but then errors may result i.i.d. only if we assume independence between errors ε and inputs **X**.

(and so we make this assumption)

- Gain --> more accurate estimation => better performance/prediction of the model
- **Cost** --> greater computational effort and more difficult-to-understand results (we completely lose, for instance, the interpretability of the betas. Now we won't interpret them)

Outline

- Why Nonparametric Regression
- Dataset & Examples
- Nonparametric Simple Regression
  - Binning and Local Averaging
  - Kernel estimation and Weighted Local Averaging
  - Local Polynomials
  - Splines
    - Regression splines
    - Smoothing splines
- Nonparametric Multiple regression: GAMs
- References

---

1. Wage dataset (from JHWT)

- This application (which is referred to as the Wage dataset throughout the JHWT book) examines a number of factors that relate to wages for a group of males from the Atlantic region of the United States.
- Goal: to understand the association between an employee's age and education, as well as the calendar year, on his wage.

Wage dataset (from JHWT)

this is the typical case in which a standard (linear) regression would poorly behave.

The other two cases (Year and Education) show that maybe a linear regression would be ok.

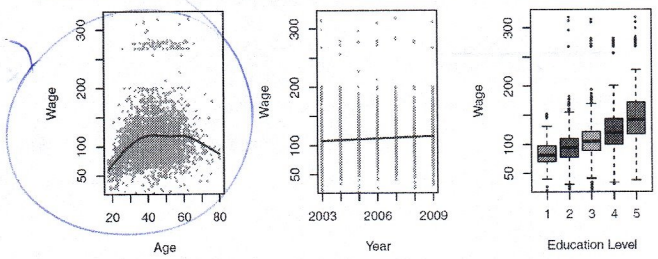


FIGURE 1.1. Wage *data, which contains income survey information for males from the central Atlantic region of the United States. Left:* wage *as a function of* age. *On average,* wage *increases with* age *until about 60 years of age, at which point it begins to decline. Center:* wage *as a function of* year. *There is a slow but steady increase of approximately $10,000 in the average* wage *between 2003 and 2009. Right: Boxplots displaying* wage *as a function of* education, *with 1 indicating the lowest level (no high school diploma) and 5 the highest level (an advanced graduate degree). On average,* wage *increases with the level of education.*

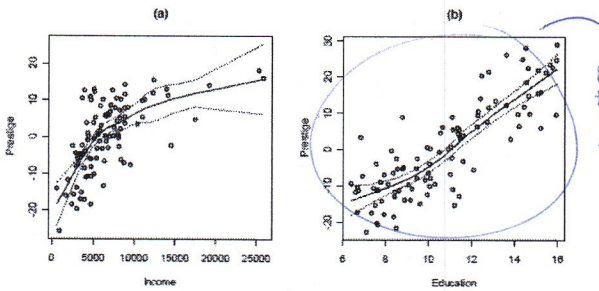
2. Occupational prestige

- Blishen and McRoberts (1976) reported a linear multiple regression of the rated prestige of 102 Canadian occupations on the income and education levels of these occupations in the 1971 Canadian census.
- Goal: to produce substitute predicated prestige scores for many other occupations for which income and education levels were known, but for which direct prestige ratings were unavailable.
  - Figure 3 shows the results of fitting an *additive nonparametric regression* to Blishen's data:

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon$$

  - The graphs in Figure 3 show the estimated partial regression functions for income (f_1) and education (f_2).

Occupational prestige



In here we may even say that the relation is linear. However, for the income there is no way to be linear.

Figure 3. Plots of the estimated partial-regression functions for the additive regression of prestige on the income and education levels of 102 occupations.

---

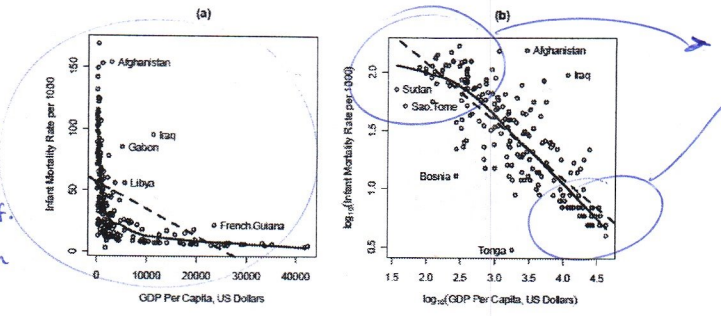
3. Infant mortality

- Goal: modeling the relationship between infant-mortality rates (infant deaths per 1,000 live births) and GDP per capita (in U.S. dollars) for 193 nations of the world (Figure 1 (a)).
  - Although infant mortality declines with GDP, the relationship between the two variables is highly nonlinear: As GDP increases, infant mortality initially drops steeply, before leveling out at higher levels of GDP.

    Note: The nonparametric regression line on the graph was produced by a method called *lowess* (or *loess*), an implementation of local polynomial regression, and the most commonly available method of nonparametric regression.
- Because both infant mortality and GDP are highly skewed, most of the data congregate in the lower-left corner of the plot, making it difficult to discern the relationship between the two variables.
- The linear least-squares fit to the data does a poor job of describing this relationship.

  Note: in Figure 1 (b), both infant mortality and GDP are transformed by taking logs. Now the relationship between the two variables is nearly linear.

Infant mortality



Strongly nonlinear. The problem is: even if we perform some transf. (e.g. the logarithm) we end up in a situation that is still nonlinear (record picture)

Note: always do a closer look on what's happening on the boundaries: is it a sign of something? can we omit it?

Figure 1. Infant-mortality rate per 1000 and GDP per capita (US dollars) for 193 nations.

Outline

- Why Nonparametric Regression
- Dataset & Examples
- Nonparametric Simple Regression
  - Binning and Local Averaging
  - Kernel estimation and Weighted Local Averaging
  - Local Polynomials
  - Splines
    - Regression splines
    - Smoothing splines
- Nonparametric Multiple regression: GAMs
- References

Nonparametric Simple Regression

- Also known as **Scatterplot Smoothing**

simple ⇒ $dim(x) = 1$, that's why "scatterplot" (the dimension is 1, we can always plot)

- Given a random pair $(Y,\mathbf{X}) \in R \times R^p$, the regression function may be expressed as $f(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}]$.
  Let consider the case **p=1 (Simple regression).**
- The goal of nonparametric regression is to build an estimate of *f* from $(Y_1,X_1),\ldots,(Y_n,X_n)$ having the same joint distribution of (Y,X), not assuming any parametric form for *f*
- We will assume X as a fixed/given input, i.e. saying that $Y_i|X=x$ are i.i.d. implicitly means that we are assuming that input (X) and errors (ε) are independent in the regression model.

We can build estimates for f essentially in two ways:

GLOBAL ESTIMATES → (f) ← LOCAL ESTIMATES

the entire range of X is considered

only part of the range of X is considered

Global estimates: POLYNOMIAL REGRESSION

- **Polynomial Regression** allows for very flexible curves, describing extremely nonlinear behaviors.

grade of the polynomial (max grade)

$$y_i = \beta_0 + \sum_{j=1}^{d} +\beta_j x^j{}_i = \beta_0 + \beta_1 x_i + \ldots + \beta_d x_i^d + \varepsilon_i \qquad i = 1,\ldots n$$

- Coefficients can still be estimated through LS.
- Interpretability of coefficients is no more a point.
- ⚠ Being a global approach, individual observation can exert an influence on remote parts of the curve

(this is a disadvantage: all the points influence all the parts in the curve)

Global estimates: POLYNOMIAL REGRESSION

(1.)

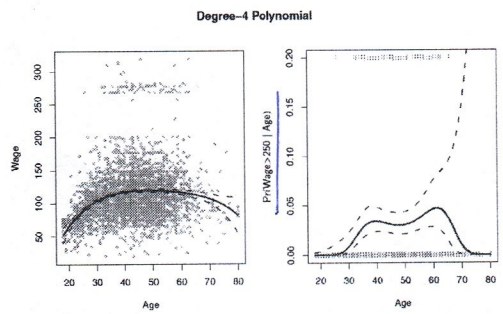


FIGURE 7.1. *The* Wage *data. Left: The solid blue curve is a degree-4 polynomial of* wage *(in thousands of dollars) as a function of* age*, fit by least squares. The dotted curves indicate an estimated 95 % confidence interval. Right: We model the binary event* wage>250 *using logistic regression, again with a degree-4 polynomial. The fitted posterior probability of* wage *exceeding $250,000 is shown in blue, along with an estimated 95 % confidence interval.*

Global estimates: STEP FUNCTIONS

- Step functions are built selecting K cutpoints, creating K+1 dummy variables indicating the value of the function in each sub interval.

$$y_i = \beta_0 + \beta_1 c_1(x_i) + \cdots + \beta_d c_k(x_i) + \varepsilon_i \qquad i = 1, \ldots n$$

$$\begin{aligned} c_0(x) &= 1(x < c_1) \\ c_1(x) &= 1(c_1 \le x < c_2) \\ &\cdots \\ c_k(x) &= 1(c_k \le x) \end{aligned}$$

} indicators

- Coefficients β_j represent the mean in the response for $x \in [c_j, c_{j+1}]$

Global estimates: STEP FUNCTIONS

(1.)

Basically we divide in K portions and in each portion the value is constant (it's the mean of the points belonging to that portion)

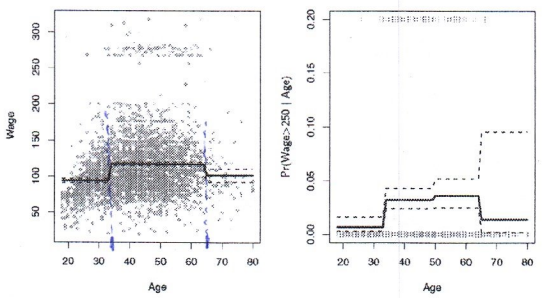


FIGURE 7.2. *The* Wage *data. Left: The solid curve displays the fitted value from a least squares regression of* wage *(in thousands of dollars) using step functions of* age*. The dotted curves indicate an estimated 95 % confidence interval. Right: We model the binary event* wage>250 *using logistic regression, again using step functions of* age*. The fitted posterior probability of* wage *exceeding* $250,000 *is shown, along with an estimated 95 % confidence interval.*

Towards local estimates: PIECEWISE POLYNOMIALS

- Merging the idea of fitting a polynomial regression and fitting a model on sub intervals of the domain, we get the piecewise polynomial approach.
- In this case, flexibility is guaranteed by fitting low degree polynomials over different regions of X, instead of fitting a high-degree polynomial over the entire range => $\boldsymbol{\beta_j}$ **are region specific**.
- The points where coefficients change are called knots.

Ex: piecewise cubic polynomial with 1 knot ⟹ only 2 expressions of Y (before and after the knot)

$$\begin{aligned} y_i &= \beta_{01} + \beta_{11} x_i + \beta_{21} x^2{}_i + \beta_{31} x^3{}_i + \varepsilon_i \qquad x_i \le c \\ y_i &= \beta_{02} + \beta_{12} x_i + \beta_{22} x^2{}_i + \beta_{32} x^3{}_i + \varepsilon_i \qquad x_i > c \end{aligned}$$

Note: we don't have to choose the same degree of polynomial in every portion (here we did, but we don't have to)

- Coefficients can still be estimated through LS.

Towards local estimates: PIECEWISE POLYNOMIALS

Disadvantages:

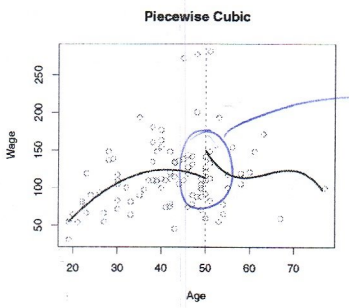
- ⚠ No guarantees about continuity of the fitting.
- ⚠ **High number of parameters** to be estimated as long as the number of knots increases.
- Constraints are needed at knots, e.g. continuity on derivatives up to the d-1 degree.

--> spoiler: splines

Here we will ask for the continuity of the derivatives up to the **d-1** degree

- Note: human eye cannot distinguish discontinuities over second derivatives

That's why in the literature we often hear about cubic splines



if we don't impose anything we might end up with something like this, which actually makes no sense! (We want our (global) curve to be at least continuous)

Local estimates: BINNING

- Suppose that the predictor variable X is continuous.
- If the sample is large enough, we may divide the range of X in a number of intervals (bins) in order to get the estimate for μ|x.

$$\mu|x \;\text{<--}\; \hat{f}(x) = \bar{y}|x$$

we divide the region in bins and we compute the average in each bin

$\bar{y}_i$ = mean of the responses in the i-th bin (for the inputs belonging in the i-th bin)

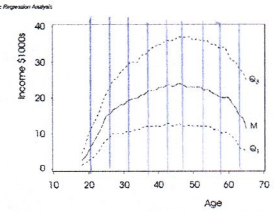


Figure 4. Simple nonparametric regression of income on age, with data from the 1990 U. S. Census one-percent sample.

Ex: each bin might be the age rounded to the nearest year

Each bin contains data => local estimates can be computed as

$$\bar{y}_i = \bar{y}|x_i \;\; for\; x\; in\; i-th\; bin$$

Proper binning provide a good estimate of μ|x anywhere

Local estimates: BINNING

- Given sufficient data, binning is essentially no-cost.
- If **few observations** are present, averages become unstable.
- A solution may be to allow bins being not uniform in size (as in Figure 5)

  --> not always possible to act on bin size.

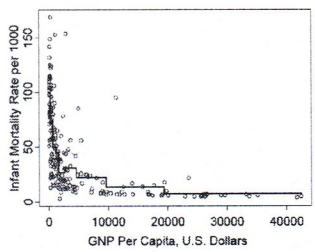


Figure 5. The binning estimator applied to the relationship between infant mortality and GDP per capita.

Local estimates: Bias-Variance trade-off

robustness vs. accuracy (being close to the data)

Bias-Variance trade-off

$$MSE\left[\hat{f}(x)\right] = V[\hat{f}(x)] + Bias^2[\hat{f}(x)]$$

- Minimizing MSE work at cross purpose:
  - **Wide bins** produce small variance and large bias in estimates
  - **Small bins** produce large variance and small bias in estimates

  Only with many observations we can care of both jointly (many observations that are well distributed)
- All nonparametric methods bump up against this problem.
- Even if the binning estimator is biased, it is consistent as long as the regression function is smooth.

--> *"All we need is to shrink the bins toward 0 as the sample size grows up, but shrinking them sufficiently slow so that the number of observations in each bin grows as well".*

we would like to have minuscolous bins but with a lot of points in each bin

One step ahead of just making the binning:

Local estimates: LOCAL AVERAGES

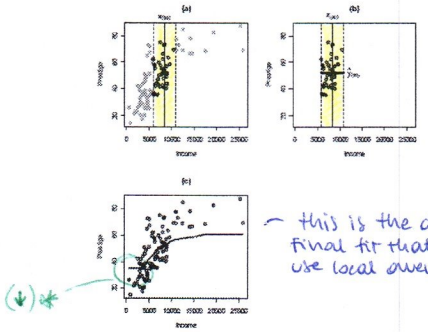
- The basic idea of local averaging is that, as long as the regression line μ is smooth, **observations** associated to x values **near to focal** x_0 **are more informative** about $f(x_0)$.
- Local averaging is different from binning since data are not dissected into a non-overlapping bins, but we move a **window** continuously over the data, averaging the observations falling in that window.

  We then calculate $\hat{f}(x)$ at a number of focal values, usually equally spread.
- Options:
  - Fixed width (**w**) windows, centered at focal values
  - Width-adjusted window, **containing k points** => k Nearest Neighbours (kNN)

different behavior in bias-variance trade-off

Local estimates: LOCAL AVERAGES

(2.)



Local averages (FIXED WINDOW) (we've continuously moving the window)
$\hat{y}$ at focal point x_0 (the center line of the window) is defined as the mean of all the points contained in the window

this is the data and the final fit that we get if we use local averages

Figure 6. Nonparametric regression of prestige on income using local averages.

Local estimates: kNN

we adjust the window every time, forcing it to contain each time **k** points

- The **kNN estimate** of $f(x)$ is

  $\hat{f}(x) = \frac{1}{k}\sum_{i \in \aleph_k(x)} y_i$ where $\aleph_k(x)$ contains the k nearest points to x

- **Problems** with local averaging :
  - Boundary Bias (artificial flattening of the curve at extremes) (↑)
  - It might produce rough estimates ($\hat{f}(x)$ tends to take small jumps as observations enter/exit the windows
  - Subject to distortion whenever outliers fall into the window
- Varying the number of neighbors k we may tune the flexibility of the estimate.
  - The lower the number, the more flexible the estimate (low bias, high variance)
  - The higher the number, the less flexible the estimate (high bias, low variance)
- The proportion k/n is referred to as **span** of the estimator.

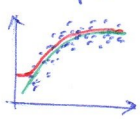
proportion of data we allow each window to contain

since we want **k** points, the points that are at the beginning or at the end are influenced only by one-direction:

these points are influenced only by the right-side

these points are influenced only by the left side

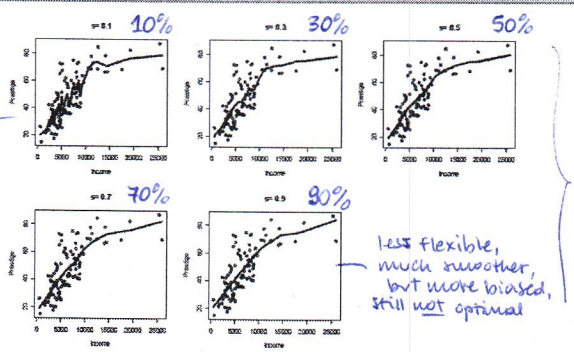
Example:



— what we want
— what we get (the flattening at the end is not a problem. The flattening at the beginning is)

Local estimates: kNN

(2.)



high variability, less bias (not optimal, the initial part is very "nervous")

made with different span for each fitting (the choice of the optimal is something subjective)

less flexible, much smoother, but more biased, still not optimal

Figure 10. Nearest-neighbor local linear regression of prestige on income, for several values of the span s.

Local estimates: WEIGHTED LOCAL AVERAGING

- If we rewrite

  $\hat{f}(x) = \frac{1}{k}\sum_{i \in \aleph_k(x)} y_i$ where $\aleph_k(x)$ contains the k nearest points to x

  as $\hat{f}(x) = \sum_{i=1}^{n} w(x_i)\, y_i$ where $w(x_i) = \frac{1}{k}$ if x_i is one of the kNN and 0 otherwise

  then we may read the estimate as a **weighted local averaging.**
- Note that $w(x_i)$ in the case above **is a discontinuous function of x** => the same is true for $\hat{f}(x)$
- This representation reveals that kNN estimates belong to a class called LINEAR SMOOTHERS.
  Being $y = (y_1, \ldots, y_n)$ and $\hat{y} = (\hat{f}(x_1), \ldots, \hat{f}(x_n))$, then $\hat{y} = Sy$
- The estimates presented up to now have a really poor dependence on d: as d increases, exponentially more samples are required to achieve a reasonable bound for the MSE
  => ***Curse of dimensionality!***

the weights provide a discontinuous function of **x** (and so, again, we have to decide the priority: smoothness? accuracy?)

where **S** is a linear operator (**S** is the matrix containing all the weights

Local estimates: KERNEL ESTIMATION

- **Kernel estimation** is an **extension of local averaging**
  => greater weight on observations close to the focal, according to a kernel density
- Let z_i denote the **scaled signed distance** between the i-th observation and the focal

$$z_i = \frac{x_i - x_0}{h}$$

h → **Bandwidth**: scale factor playing the role of window's width in the binning

- A **Kernel K(z)** is a function attaching greater weights to observations close to x_0 and that falls off symmetrically and smoothly as |z| grows.
- Denoting $w_i = K(\frac{x_i - x_0}{h})$, the kernel estimation at focal point is

$$\hat{f}(x) = \hat{y}|x_0 = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

It's still a local method: in the local neighborhood we weight datapoints differently

Local estimates: KERNEL ESTIMATION

- Popular kernel (h fixed) are:
  - Gaussian,
  - Tricube,
  - Rectangular,
  - Epanechnikov

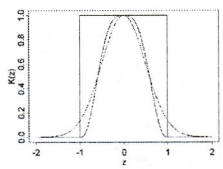
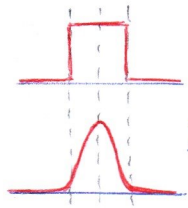


Figure 7. Tricube (light solid line), normal (broken line, rescaled) and rectangular (heavy solid line) kernel functions.



all the points in the bandwidth weight the same

the closer to the focal, the more weight a point has

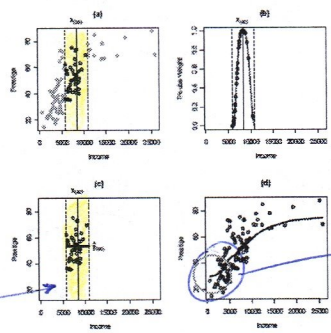
- Also the kernel bandwidth may be made adaptive to contain k elements:

$$\frac{k}{n} = \text{SPAN of the kernel}$$

- Varying the bandwidth of the kernel controls the bias-variance trade-off: **larger bandwidth => smoother results**

Local estimates: KERNEL ESTIMATION

(2.)



the boundary bias is still present (especially for those points of the curve where we have high derivatives)

in this window we apply a kernel (in this case it's a bell-shape kernel) and we get an estimate for the focal point that weights the surrounding k-points according to the kernel

Figure 8. The kernel estimator applied to the Canadian occupational prestige data.

Local estimates: LOCAL POLYNOMIALS

- **Local polynomials** extend kernel estimation to a **polynomial fit at the focal point using local kernel weights**.
- The resulting Weighted Least Squares (WLS) regression equation is

$$y_i = \beta_0 + \beta_1(x_i - x_0) + \beta_2(x_i - x_0)^2 + \ldots + \beta_d(x_i - x_0)^d + \varepsilon_i$$

where the estimates can be obtained minimizing the **Weighted Residual Sum of Squares**

$$WRSS = \sum_{i=1}^{n} w_i e_i^2 = \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)(y_i - \beta_0 - \beta_1(x_i - x_0) - \cdots)^2$$

- Truncating at d=1 we get a local linear regression.
- Truncations at d>1 provide more flexibility (i.e., greater variability) =>Useful to mitigate the boundary bias issue.

Now we have:

- different weights on the observations in the window
- we assume a model for the fitting of the points in each bin
- we average the estimates through a kernel

→ The result of all of this is a weighted least square regression equation where the estimation of the parameters is weighted through a kernel.

Local estimates: LOCAL POLYNOMIALS

- The bandwidht *h* may be fixed or may vary as a function of the focal point, in order to include a given amount of points in the neighbor.
- When the bandwidth define a window of neighbor (as in the tricube case), the degree of smoothness may be targeted specifying the proportion o observations to be included into the window

$$m = \#\text{ of obs included} = \lfloor sn \rfloor$$

↑ SPAN of the local regression smoother

- How can we choose the "right" span?
  - **Trial and error** => start with 0.5 and increae/decrease in order to choose the smallest s providing a smooth fit
  - **Cross validation** => omit the i-th observation from local regression at the i-th focal value, then evaluate the **CV(s)** function at different s, selecting the s that act as a minimizer.

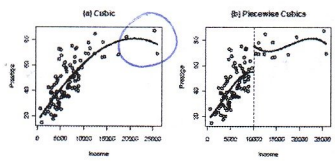
$$CV(s) = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_{-i}(s) - y_i|^2 \quad \text{being } \hat{y}_{-i}(s) = \hat{y}_{-i}|x_i \text{ for span s}$$

- Also the local polynomial regression can be seen as a case of linear smoothing.

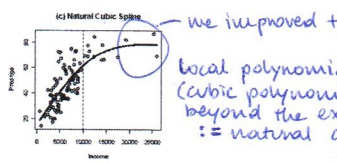
Local estimates: LOCAL POLYNOMIALS

(2.)

global cubic fit →



2 independent cubic fits (2 bins without constraints)



we improved this (w.r.t. the global estimate)

local polynomial fit (cubic polynomial, linear beyond the extreme := natural cubic spline)

Figure 15. Polynomial fits to the Canadian occupational prestige data: (a) a global cubic fit; (b) independent cubic fits in two bins, divided at Income $= 10,000$; (c) a natural cubic spline, with one knot at Income $= 10,000$.

Global

Local estimates: SPLINES

we select suitable basis functions that allow us to fit in each bin a polynomial

- **Regression and smoothing splines** are motivated by a different perspective than kernel and local polynomials.
  - Step back to global estimate, starting from a set of selected basis functions
  - Start from local averaging and move towards higher order local models
- **Splines** are piecewise polynomial functions which are constrained to join smoothly at knots.
- The traditional application for splines is interpolation, but they may be used also for nonparametric regression purposes.
- Among others, **cubic splines** are the most popular.
  (we can impose the smoothness at knots up to the 2nd derivative)

This is an upgrade, with the previous method of local polyn. we had the problem of continuity (smoothness in general) at the knots.
Modeling through basis functions guarantees the smoothness at the knots.

Local estimates: REGRESSION SPLINES

- **Regression splines** guarantee a local fit of polynomials with smooth connections at knots.
  - **Cubic regression splines** => 3rd order polynomial in each bin with 1st and 2nd derivtives continuous at knots
  - **Natural Cubic Splines (NCS)** => add knots at boundary and impose linear fit beyond
- Regression spline "bricks":
  - Outputs: Inputs: $y_1,\ldots,y_n$
  - Inputs: $x_1,\ldots,x_n$
  - **# of knots:** m
  - **Degree of the spline:** k
- The way for ensuring the constraint of smoothness at knots for the function and the k-1 derivatives is to build a regression through suitable basis functions

so we'll use **k**-order splines and they'll be fitted to **m** selected knots

We saw:
- general global polynomial fits: able to capture high varying terms; however they have a high variability, especially at the borders
- local polynomial: we fit data into bins and then we fit polynomial regressors. This approach has the problem of being (possibly) discontinuous

Solution to the problems:
REGRESSION SPLINES;
(which averages between the two) it guarantee the local fitting and the smoothness at the knots.

Local estimates: REGRESSION SPLINES

- **General global polynomial** fits are able to capture widely varying forms, but are **highly nonlocal** and subject to considerable sampling variation.
- As an alternative, we can fit data into bins, and fit polynomial regression in each bin. --> **PB**: discontinuity at knots
- Regression splines guarantee a local fit polynomials with smooth connections at knots.

we're in the univariate case: n observations of the output and the inputs

Given inputs $x_1,\dots x_n$ and outputs $y_1,\dots y_n$, a k order spline with knots $t_1,\dots,t_m$ may be expressed in the basis form

$$f(x) = \beta_0 + \sum_j \beta_j g_j(x) + \varepsilon$$

$g_j(\cdot)$: basis functions

being the basis composed by a k-degree polynomial and by a truncated power basis function per each knot => k+m+1 parameters

$$\Rightarrow x, x^2, \dots, x^k, h(x,t_1), \dots, h(x,t_m)$$

where $h(x,t_j) = (x-t)_+^{\,j} = (x-t)^j \text{ if } x > t, 0 \text{ otherwise}$

Local estimates: REGRESSION SPLINES

- Estimation of the m+k+1 parameters via LS :

$$\hat{\beta} = \arg\min_\beta \|y - G\beta\|_2^2 = \arg\min \sum (y_i - (G\beta)_i)^2$$

$$\hat{\beta}^T = (G^T G)^{-1} G^T y$$

$$\Rightarrow \hat{y} = \hat{f}(x) = G\hat{\beta} = G(G^T G)^{-1} G^T y = Sy$$

also the regression splines are linear smoothers

We define $G \in \mathbb{R}^{n \times (m+k+1)}$:

$$G = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^k & h(x_1,t_1) & \dots & h(x_1,t_m) \\ \vdots & & & & & & & & \\ 1 & x_n & x_n^2 & x_n^3 & \dots & x_n^k & h(x_n,t_1) & \dots & h(x_n,t_m) \end{bmatrix}$$

- Also in this case, we can say that regression splines are linear smoothers.

Ex: cubic splines

3rd degree polynomial

$$f(x) = \beta_0 + (\beta_1 x + \beta_2 x^2 + \beta_3 x^3) + (\beta_4 (x-t_1)^3 + \dots + \beta_{3+m}(x-t_{3+m})^3) + \varepsilon$$

Local estimates: REGRESSION SPLINES

- How can we choose the location and number of knots?
  - **Location**:
    regression splines are more flexible in regions containing many knots, since coefficients may change rapidly
    => put more knots where you know the function varies most rapidly, less where is more stable.
  - **Number**:
    1) Trial and error 2) CV

("ignorant choice":) otherwise, if we know nothing about the problem we can choose to put the points over a grid (equally distanced points)

- **Note:** compared to polynomial regression, regression splines introduce more flexibility into the fit increasing the number of knots, but mantaining fixed (and possibly low) the degree of the polynomial

  => more stable estimates

we capture the same variability that can be captured by high-degree polynomial regression but with fewer parameters (since we do not modify the degree of polynomials, we simply add knots)

Local estimates: REGRESSION SPLINES

⚠ **Problem**: regression splines tend to have high variance at the boundary of the domains

---> Differently from kernel smoothing, which has poor bias at the boundary

- Solution: force piecewise polynomial function to have a lower degree to the left of the leftmost knot and to the right of the rightmost knot

=> **Natural Splines** (!! Only defined for odd order k !!)

- Polynomial of degree k in each interval $[t_1,t_2],\dots,[t_{m-1},t_m]$
- Polynomial of degree $\frac{k-1}{2}$ on $(-\infty, t_1]$ and $[t_m, \infty)$
- Continuous and with continuous derivatives o order 1,...,k-1 at knots $t_1,\dots,t_m$

Cubic splines: in the boundaries the fitting is linear while the fitting in the internal intervals is of degree 3

Local estimates: SMOOTHING SPLINES

- The idea underlying regression splines relies on specification of a set of knots, producing sequence of basis functions and then using LS for estimating coefficients.
- Now we introduce a different approach, also producing splines-based estimates.
- **Smoothing splines** perform a penalized regression over the natural spline basis (orthonormal), placing knots at all the inputs $x_1, \dots, x_n$

  => No problem of node selection

  => Control for overfitting by shrinking the coefficients of the estimated function. (more shrinkage is assigned to eigenvectors of the basis that correspond to larger eigenvalues)
- **Smoothing splines** can be **motivated also from a functional minimization perspective.**

Local estimates: SMOOTHING SPLINES

- In fitting a smooth curve to a set of data, we look for a function $g(x)$ minimizing

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

- **Note** that if we don't constrain $g(x)$, we can always get RSS=0 simply interpolating the data (**overfit**).

=> We want a *smooth* function $g(x)$ *making RSS to be small.*

- A good fit to the data is not only aiming at curve fitting, but also displaying a not "too rapid fluctuation".

=> **Reformulate** the problem as a **roughness penalty approach**, quantifying the notion of "rapidly fluctuating curve" and posing the estimation problem in a way that makes it explicit the necessary **trade-off between conflicting aims**.

Local estimates: SMOOTHING SPLINES

How can we quantify roughness?

("rough" function ≃ funzioni "nervose")

- Given a curve g defined on $[a, b]$, there are many different ways of measuring how rough g is. One may be

$$\int_a^b (g''(t))^2 dt$$

see also [3, p. 11] for a more general setting

- Why should be this reasonable and/or effective?
  - For not being affected by addiction of constant or linear functions
  - $|g''|$ is the maximum number of inflection points of the curve g, but $\int_a^b (g''(t))^2 dt$ is a global measure of roughness
  - Computational advantages

Local estimates: SMOOTHING SPLINES

Loss+Penalty formulation:

- Given a twice-differentiable function g defined in [a,b] and a parameter $\gamma > 0$, we may define the cost function

$$S(g) = \underbrace{\sum_{i=1}^{n} [y_i - g(x_i)]^2}_{RSS} + \gamma \underbrace{\int_a^b (g''(t))^2 \, dt}_{Roughness\ Penalty}$$

tell us about how good the fitness is

$\gamma \cdot L^2$ penalization over the domain of the second derivative

γ has the role of **tuning param.**

- The penalized LS estimator $\hat{g}$ is the minimizer of $S(g)$ over the class of twice-differentiable functions.
- **Natural cubic splines** can be demonstrated to result the **interpolant with minimal roughness** (see [1], p.18).
  This comes as a mathematical consequence of choosing $\int_a^b (g''(t))^2$ as a roughness penalty.

Local estimates: SMOOTHING SPLINES

- The penalty term γ controls the **bias-variance trade-off** in the **smoothing spline**

  => the higher γ, the smoother g <=

- The penalty term ensures that the **cost $S(g)$ of a particular curve is determined by the balance** of the loss in fitting and penalty on roughness
  - **Large γ** => RSS is the main component, the minimizer will show small curvature

    $\gamma = \infty$ means high penalty on roughness, making g be the linear regression fit (high bias, low variance/fluctuating curve)
  - **Small γ** => Penalty matters, the curve will follow the data closely

    $\gamma = 0$ means no penalty on roughness (low/no bias, high variance/very fluctuating curve interpolating data)

Local estimates: SMOOTHING SPLINES

- How can we choose $\gamma \in [0, \infty)$?

  Cross validation approach based on **LOOCV (Leave One Out Cross Validation) ERROR** --> reasonable computational costs

- Knowing that $\hat{g}$ is a natural cubic spline (NCS) is an enormous advantage, since we need to **minimize $S(g)$ over a finite-dimensional class of functions**, i.e. the (NCS) with knots at x_i, instead of considering the infinite dimensional set of smooth functions on $[a, b]$.

  Optimizing algorithms in [1], pp 18-24

- None of the optimizing algorithm depend on the interval $[a, b]$.

Local estimates: SMOOTHING SPLINES

(2.)

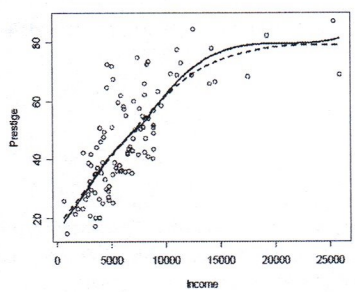


Figure 16. Nonparametric regression of occupational prestige on income, using local linear regression (solid line) and a smoothing spline (broken line), both with 4.3 equivalent parameters.

Notice that: higher the density of the points in a particular zone and smaller the difference between local linear regression and smoothing (or in general regr.) Splines

! **Remember**: splines are the most used. (all along the nonlinear regression) Local average or something else is used only if there are not enough data.

Outline

- Why Nonparametric Regression
- Dataset & Examples
- Nonparametric Simple Regression
  - Binning and Local Averaging
  - Kernel estimation and Weighted Local Averaging
  - Local Polynomials
  - Splines
    - Regression splines
    - Smoothing splines
- Nonparametric Multiple regression: GAMs
- References

The multiple regression case: GAMs

- So far we looked mostly at the **univariate case** p=1.
- Several methods have a **multivariate counterpart** (p>1):
  - Kernel smoothing --> multiple dimensional kernel
  - Splines --> thin-plate splines or tensor-product splines [6]
  - Mercer Kernels --> defined directly in multiple dimensions
- In high dimensions (large p) things are problematic, and suffer from poor variance

⇒ Curse of dimensionality, i.e., the estimation gets esponentially harder as the number of dimensions increase.

- To mitigate this problem, a solution is to adopt an additive approach

⇒ **GAMs – Generalized Additive Models**

Provide a general framework for extending a standard linear model by allowing nonlinear functions o each predictor while mantaining additivity

The multiple regression case: GAMs

- Instead of considering a full p-dimensional function

$$f(\boldsymbol{x}) = f(x_1, \ldots, x_p),$$

we restrict our attention to functions of the form

$$f(\boldsymbol{x}) = f(x_1) + \cdots + f(x_p)$$

- This choice need not to be regarded as an assumption we make about the true function, just like we do not assume that the true model is linear in a traditional regression.
- Additive models may represent a **good approximation to the truth**, that is able **to scale well with the number of dimensions p.**
- Stone (1985) shows that *while it may be difficult to estimate an arbitrary regression function in high dimension, we can still estimate its best additive approximation well.* In other words, each time we cannot hope to recover the true underlying function arbitrarily, we can recover its major structure along the coordinate axes.

The multiple regression case: GAMs

- **GAMs** rely on a sort of **building blocks model**, where all the models adopted for each predictor may be packed into one big regression matrix.
- Estimation with additive models is actually simple: we can just choose our favourite univariate smoother (i.e., nonparametric procedure) and cycle through estimating each function individually (**backfitting**) [3].

One at the time:
once we're dealing with one variable the other remain fixed. We update the variable we're dealing with, we fix it and we pass to another variable.

  + GAMs allow for fitting nonlinear models to each predictor
    => we can model relationships that linear models would miss
  + Due to additivity, we can still analyze and interpret the effect of each covariate on the response, holding all the other covariates fixed
  - GAMs are restricted to be additive

The multiple regression case: GAMs

(1.)

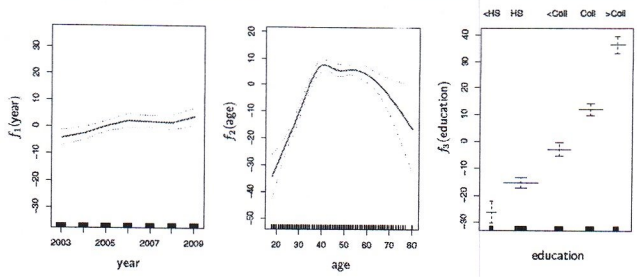


FIGURE 7.11. *For the* Wage *data, plots of the relationship between each feature and the response,* wage, *in the fitted model (7.16). Each plot displays the fitted function and pointwise standard errors. The first two functions are natural splines in* year *and* age, *with four and five degrees of freedom, respectively. The third function is a step function, fit to the qualitative variable* education.

The models are different from variable to variable (even between the two continuous): we decide the best fitting separately.

Note: this is different from 3 simple (single) regressions (even if the other covariates are fixed we're considering them)

References

Books & journal papers

[1] Green, P.J., Silvermann, B.W. (1994) Nonparametric Regression and Generalized Linear Models. A roughness penalty approach. Chapman and Hall

[2] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R. Springer

[3] Tibshirani, R., Wasserman, S. (2015) Statistical Machine Learning course.

[4] Fox, Nonparametric Simple Regression, Sage (2000)

[5] Fox, Multiple and Generalized Nonparametric Regression, Sage (2000)

[6] Wood, S.N. (2003) Thin plate regression splines. Journal of the Royal Statistical Society – series B, 65(1): 95-114

Links

- https://statisticallearning.org/nonparametric-regression.html
- https://www.semanticscholar.org/paper/Nonparametric-Regression-Statistical-Machine-%2C-2019-Tibshirani-Wasserman/afb4376dd9983371c648852e8c3ede0cd7666305
- http://www.stat.cmu.edu/~ryantibs/statml/
- https://www.publichealth.columbia.edu/research/population-health-methods/thin-plate-spline-regression

Packages

splines - https://www.rdocumentation.org/packages/splines/versions/3.6.2

gam and mgcv - https://www.rdocumentation.org/packages/mgcv/versions/1.8-33/topics/gam

+ Examples:

https://pjbartlein.github.io/GeogDataAnalysis/lec14.html