# Hints on non-parametric approaches to investigate the mean-relationship between Y and X

Consider (popul. level)

$$Y = \underbrace{f(X)}_{E(Y|X)} + \varepsilon \qquad \varepsilon \perp\!\!\!\perp X, \; E(\varepsilon)=0, \; var(\varepsilon)=\sigma^2$$

and possibly $\varepsilon \sim N(0,\sigma^2)$

The simple linear regression model we are considering postulates a simple parametric form for $E(Y|X)$, in particular, 1st ord polynomial

$$f(X) = \beta_0 + \beta_1 X$$

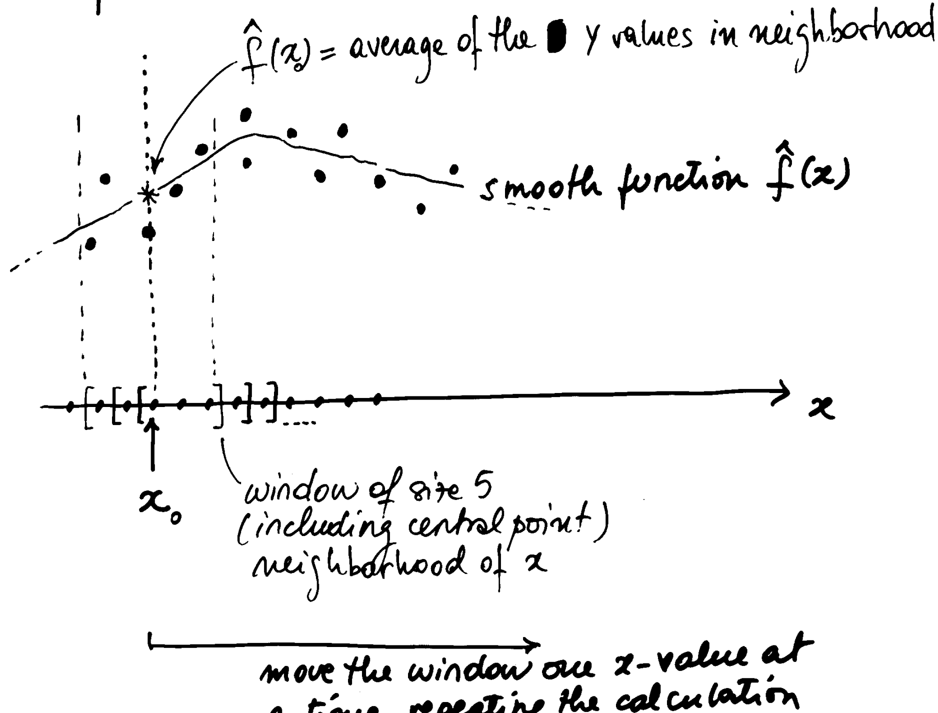
↑ ↑ parameters estimated on the data by L.S.

different approach: "reconstruct" $f(\cdot)$ from the data without postulating a parametric form. Big advantage in many circumstances, although non-parametric approaches DO NOT PRODUCE AN EXPLICIT EQUATION for $f(\cdot)$.

Note: could be used as preliminary exploratory tools, to suggest a parametric form based on the data...

## Moving averages (or medians)

when observed X values are equispaced (e.g. time series) - Basic "parameter" determining the degree of data-smoothing: SIZE of the MOVING WINDOW

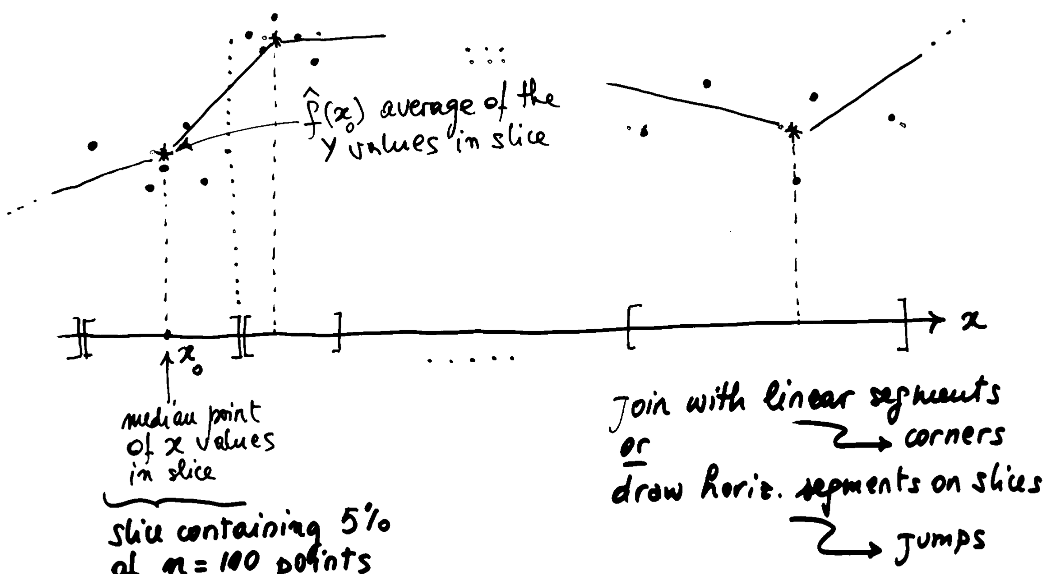



If we suspect the data are contaminated by erroneously or anomalously high/low values of Y, robustify using Moving medians to capture the "center"-relationship of Y to X

## Using Slices: (non-overlapping)

applicable also when observed X values are not equispaced. Divide the observed X range in intervals → of same length
→ **better**, containing an equal share of data points (e.g. 10 slices, each with 10% of the points)

Basic "parameter" determining the degree of data smoothing: # of slices / percent. of points per slice



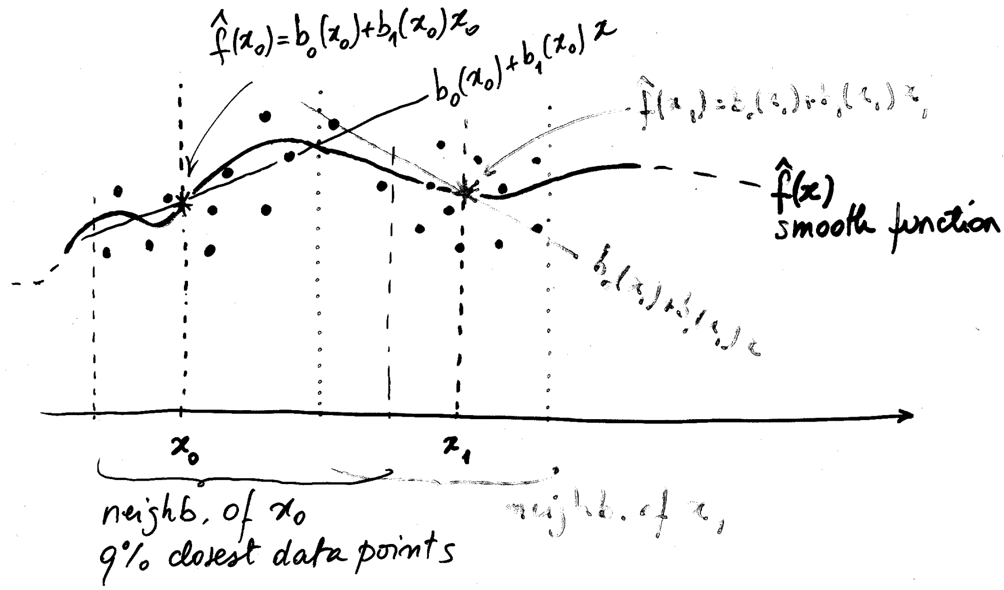

Also here, could use medians instead of means

## A more sophisticated solution:

Fitting successive parametric regressions (e.g. lines) in local neighborhoods

Basic idea

- pick a percentage of data points $q$ (smoothing "param")
- at each $x_0$ (observed or not) form a surrounding neighborhood with the closest $q\%$ data points
- on this subset of points, fit $E(Y|X) = \beta_0 + \beta_1 x$ to obtain $b_0(x_0)$ $b_1(x_0)$
- set $\hat{f}(x_0) = b_0(x_0) + b_1(x_0) x_0$



More detail:

a. the parametric model used in local fits may also be of 2nd order (parabola)

b. The fit is not accomplished by L.S., but by <u>weighted</u> <u>least squares</u> (an important variant) Within the neighb. of an $x_0$ defined by $q$, points are further assigned a weight inversely proportional to their distance from $x_0$. These weights are then used in forming the sum of squares to be minimized, e.g.

$$\sum_{i \in \text{neighb } x_0} \omega_i * (Y_i - (\beta_0 + \beta_1 x_i))^2$$

c. Weights are computed with special formulae

d. To robustify the procedure towards outlying Y values the fit of the parametric model in each neighborhood maybe iterated several times, each time revising the weight $\omega_i$ as to reflect not only distance of $x_i$ from $x_0$, but also residual $Y_i - \hat{y}_i$ from the previous fit (larger residual, smaller weight)

With these specifications, we have

<u>LOWESS</u> locally weighted regression scatter plot smoothing

Cleveland 1979, Cleveland & Devlin 1988

Note 1 : most common weight function, TRICUBE

$d_q$ = euclidean distance from $x_0$ of the furthest $X_i$ in its q% neighborhood

$$\omega_i = \begin{cases} \left(1 - \frac{d(X_i; x_0)}{d_q}\right)^3 & ; \; d(X_i; x_0) \leq d_q \\ 0 & ; \; d(X_i; x_0) > d_q \end{cases}$$

→ there are points outside the neighb.

Note 2 : Could the lowess technique be extended to the study of the mean relationship between Y and more than one predictor, say $X_1$, $X_2$ ?

Yes, use for example $E(Y|X_1 X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ in each local fit ...

define neighb.'s and weights using 2D Euclid. dist...

But soon a problem of SPARSENESS comes in

↑ as the # of predictors increases