model : $Y = X\beta + \varepsilon$ →

Mean structure

true model

- idea: data are generated from an underlying system, which is assumed to have the form: $y = f(x_1, \ldots, x_m) + \varepsilon$, where f is ***unknown***. ← use data to gain information
- regression *approximates* the mean structure f by a linear combination of (known) *base functions* $g_i(x_1, \ldots, x_m)$'s, $i=1, \ldots, p$, i.e.,

$$f \longleftarrow \sum_{i=1}^{p} \beta_i \cdot g_i(x_1, \ldots, x_m)$$

data determine → unknown

especially on a "local" region of the predictors

  - ➢ when the structure of f is simple and almost linear, it can be approximated by a simple structure with fewer terms, e.g.,

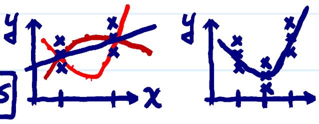
complex model ← 是否有需要?

$$E(y) = f \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

是否有能力?

  - ■ **Q**: nature is simple? → lack of fit problem → [change g_i's (transformation) / add more g_i's
  - ■ **Q**: are there sufficient data to support/fit a complex model?

  - ➢ when f is complex and non-linear ⇒ need more terms to get a good approximation on a wider region of the predictors



if YES & YES

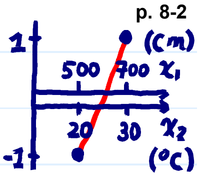
  - ▪ more parameters, need more degrees of freedom, i.e., more data
  - ▪ e.g., 2 levels, only linear effects; 3 levels, linear and quadratic effects

# of distinct x_i's

  - ■ **Q**: what other complex models? → What base functions should we consider?
- base functions for quantitative and qualitative predictors x_i's are defined in different ways

e.g. $\underset{\sim}{x} \to \underset{\sim}{x} - \bar{x}\mathbb{1}$

location and/or scale change



- $x_i \to (x_i+a)/b$ or $y \to (y+a)/b$, where a and b are given constants.
  a: change of location, b: change of scale

Q: Can we compare the magnitudes of $\hat{\beta}_i$'s to identify important effects? Ans. In general, NO.

- **Q**: why we might want to do this?
  - ➢ predictors of similar magnitude are easier to compare $\hat{\beta}_i$'s ← •ridge regression •LASSO
  - ➢ rescaling or change of units may make $\hat{\boldsymbol{\beta}}$ easier to read and may aid interpretability
  - ➢ numerical stability is enhanced when all predictors are on a similar scale

$\hat{\beta}$ ⇑ $(x'x)^{-1}$ ⇑ $|x'x| \approx 0$

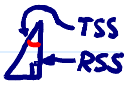
500cm 700cm

  - ➢ for experimental data, it's often that we code two levels (say, 20˚C, 30˚C) → (–1, 1); three levels (say, 20˚C, 30˚C, 40˚C) → (–1, 0, 1) → remove units
- influence caused by location/scale change on x_i (i.e., $x_i \to (x_i+a)/b$) $= x' = \frac{1}{b}\underset{\sim}{x} + \frac{a}{b}\mathbb{1}$
  - ➢ (under a model with intercept) overall F-test, t-test, R^2, $\hat{\sigma}$ all unchanged ∵ same Ω ∵ same ω

$$E(y) = \beta_0 + \cdots + \beta_i x_i = (\beta_0 - \beta_i a) + \cdots + (b\beta_i)\left(\frac{x_i+a}{b}\right) = \beta_0' + \cdots + \beta_i' x_i'$$

$$t_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{b\hat{\beta}_i}{se(b\hat{\beta}_i)} \leftarrow \boxed{\hat{\beta}_i'}$$

  - ➢ $\hat{\boldsymbol{\beta}}$ change: $\hat{\beta}_i \to b\hat{\beta}_i$, $\hat{\beta}_i'$ $\hat{\beta}_0 \to \hat{\beta}_0 - a\hat{\beta}_i$ ← $\hat{\beta}_0'$
- influence caused by location/scale change on y (i.e., $y \to (y+a)/b$) $= y' = \frac{1}{b}\underset{\sim}{y} + \frac{a}{b}\mathbb{1}$



  - ➢ (under a model with intercept) overall F-test, t-test, R^2 unchanged

$$y = \beta_0 + \sum_i \beta_i g_i(\underset{\sim}{x}) + \varepsilon \Rightarrow y' = \frac{y+a}{b} = \frac{\beta_0+a}{b} + \sum_i \left(\frac{\beta_i}{b}\right) g_i(\underset{\sim}{x}) + \frac{\varepsilon}{b} = \beta_0' + \sum_i \beta_i' g_i(\underset{\sim}{x}) + \varepsilon' \Leftarrow Var(\varepsilon') = \frac{1}{b^2} Var(\varepsilon)$$

  - ➢ $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}$ change: $\hat{\sigma} \to \hat{\sigma}/b$, $\hat{\beta}_i \to \hat{\beta}_i/b$ for each i, $\hat{\beta}_0 \to (\hat{\beta}_0 + a)/b$

$\sqrt{Var(\varepsilon')}$ $\hat{\beta}_i'$ $\hat{\beta}_0'$

❖ **Reading**: Faraway (2005, 1st ed.), 5.2 ❖ **Further reading**: D&S, 16.2, 16.3

base functions are polynomial terms →

Polynomial regression

unknown true f

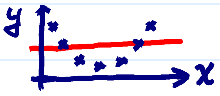
e.g. differentiable infinitely, no jump points, no broken lines

- **Q**: when to use polynomial regression? ⇒ the relationship between response and *quantitative* predictors is **smooth**, but not a straight line. ← $E(y|\underline{x})$ ← $x_1, \dots, x_m$
- **idea supports the approach** ⇒ any smooth function (mean structure of the underlying system) can be approximated by a polynomial of high enough degree ← locally ← Taylor's expansion
- one predictor case:

$$y = \beta_0 + \beta_1 \underline{x} + \beta_2 \underline{x^2} + \dots + \beta_d \underline{x^d} + \varepsilon$$

base functions

  - choice of *d*
    - start with $y=\beta_0+\beta_1 x$, keep adding polynomial terms until last term added is not significant. ⇒ danger: stop too soon

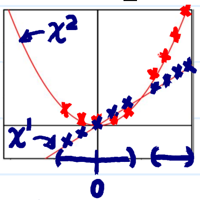


    - start with a large *d* and recursively remove insignificant largest term
    - use added variable plot/partial residual plot to gain information about *d*

consider it's for •prediction •interpretation

  - **Q**: Consider the model $y=\beta_0+\overset{\times}{\beta_1}x+\overset{\checkmark}{\beta_2}x^2$. what if β_1 not significant, but β_2 is significant? should *x* be removed from the model?

$E(y)=\beta_0+\underline{\beta_2x^2}$ has maximum minimum at $\underline{x=0}$



    - *x* and x^2 could be highly correlated
    - location shift: $x \to x+c$ ⇒ $\hat\beta_2$ unchanged, but $\hat\beta_1$ may become significant

$$E(y)=\beta_0+\overset{\approx 0}{\beta_1}x+\beta_2x^2=\beta_0'+\beta_1'(x+c)+\beta_2'(x+c)^2$$

← do not want our model sensitive to the change $x \to x+c$

    - recommendation: do not remove insignificant lower-order terms from model when the highest-order term is significant

- two predictors x_1, x_2 case: ← 1st-order model (d=1)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \quad (d=2,\ 2^{nd}\text{-order model})$$

  - the cross-product term x_1x_2 can be interpreted as an "interaction" effect, e.g., cf. ← 交互作用

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2, \text{ where } x_1,x_2\in\{-1,1\}$$

$$x_1=+1 \Longrightarrow E(y)=(\beta_0+\beta_1)+(\beta_2+\beta_3)x_2$$

$$x_1=-1 \Longrightarrow E(y)=(\beta_0-\beta_1)+(\beta_2-\beta_3)x_2$$

difference of slopes $=2\beta_3$

  - models for more predictors can be similarly extended

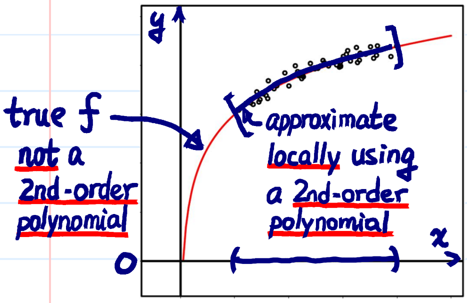
$$y=\underbrace{\beta_0+\sum_{i=1}^m\beta_{1,i}\,x_i}_{\text{1st-order model}}+\sum_{i=1}^m\beta_{2,i}\,x_i^2+\sum_{1\le i<j\le m}\beta_{3,ij}\,x_ix_j+\epsilon$$

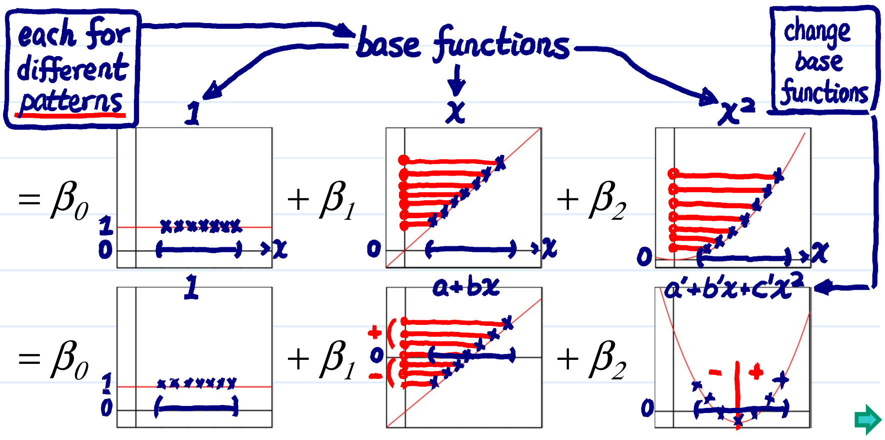
# of parameters $=1+m+m+\binom{m}{2}=(m+1)(m+2)/2$

  - increasing degree *d* ⇒ model may have too many parameters

# of x_i's	d=2	d=3
2	6	10
3	10	20
4	15	35

- orthogonal polynomials
  - polynomial terms can cause numerical instability (especially when *d* large) and collinearity ← ∵ collinearity ⇒ $|X^TX|\approx 0$ ⇒ $(X^TX)^{-1}$ unstable
  - example: 2nd-order model



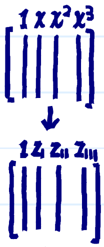


- define $z_0=1$, $z_1=a_1+b_1x_1$, $z_{11}=a_2+b_2x_1+c_2x_1^2$, $z_{111}=a_3+b_3x_1+c_3x_1^2+d_3x_1^3$, ...

  Find $a_i, b_i, c_i, \ldots$, that make $z_j^T z_k=0$ if $j \neq k$ (and $\|z_i\|=1$ sometimes)

  scale change

  Gram-Schmidt process



  - can apply regression to obtain $z_0, z_1, z_{11}, \ldots$ (note: $\underset{\Omega \ni}{\hat{y}}{}^T \underset{\in \Omega^\perp}{\hat{\varepsilon}} = 0$), e.g., regress x_1 on z_0, then the residuals is proportional to z_1; regress x_1^2 on z_0, z_1 and the residuals is proportional to z_{11}. In R, built-in function is provided to construct orthogonal polynomials.
  - cross-product terms (i.e., interactions) can be defined in a similar manner (e.g., $z_{12}= a+b\,x_1+c\,x_2+d\,x_1x_2$, regress x_1x_2 on z_0, z_1, z_2, and the residuals are proportional to z_{12})

• $x_1=\hat{\beta}_0 z_0+\hat{\varepsilon}_1$
$\Rightarrow \hat{\varepsilon}_1=x_1-\hat{\beta}_0 z_0$: a 1st-order polynomial of x_1
$\Rightarrow \hat{\varepsilon}_1 \perp z_0$
$\Rightarrow \hat{\varepsilon}_1 \propto z_1$

• $x_1^2=\hat{r}_0 z_0+\hat{r}_1 z_1+\hat{\varepsilon}_2$
$\Rightarrow \hat{\varepsilon}_2=x_1^2-\hat{r}_0 z_0-\hat{r}_1 z_1$: a 2nd-order polynomial of x_1
$\Rightarrow \hat{\varepsilon}_2 \perp z_0$
$\hat{\varepsilon}_2 \perp z_1$
$\Rightarrow \hat{\varepsilon}_2 \propto z_{11}$

- change model based on polynomial terms to model based on z's, e.g.,

$$y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_{11}\, x_1^2 + \beta_{22}\, x_2^2 + \beta_{12}\, x_1x_2 + \epsilon$$

$$\longrightarrow y = \beta_0' + \beta_1'\, z_1 + \beta_2'\, z_2 + \beta_{11}'\, z_{11} + \beta_{22}'\, z_{22} + \beta_{12}'\, z_{12} + \epsilon$$

In DOE, often use z's, rather than x's → factor X. (A.B,C)→(-1,0,1), (A.B,C)→(1,-2,1)

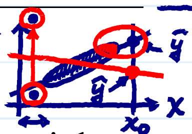
interested in β', not β

  the two models have same column space Ω (i.e., same R^2, $\hat{\sigma}$, overall F), but interpretation of β's and β''s are different (i.e., different estimates, t-tests)

- orthogonality can save works when selecting model (do not have to refit after deleting term), it's more convenient for fitting and testing ← Why? ∵ orthogonality (LNp.5-9)

• properties of polynomial model
  - offer more flexible relationship ← e.g. curvature & interaction effects
  - remember that it's an approximation, we usually do not believe it exactly represents the underlying reality ← Taylor expansion on a local area (response surface methodology)

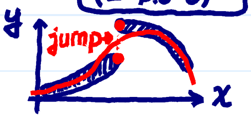
  - polynomials have the advantage of smoothness → polynomial: infinitely differentiable & $\exists k$ s.t. $\frac{d}{dx^t}E(y)=0$, $t \geq k$ | solutions of $d/dx\,E(y)=0$
  - but, have the disadvantage that each data point affects the fit **globally** (cf.)



  - For larger values of d, the fitted polynomial curves may become wiggly.
    reason: the curve may capture the random variation, rather than the overall shape of the relationship between predictors and response. ← check overfitting (LNp.6-6)

LNp.6-6 Lab06-03

  - polynomial model is hard to fit "jump function" → "local" change



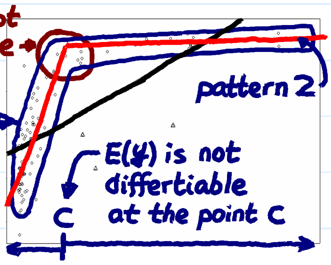
❖ **Reading**: Faraway(2005, 1st ed.), 7.2.2 ❖ **Further reading**: D&S, 12.1, 12.3, 22.2

broken stick (line) regression (segmented regression)

• Recall. polynomial regression: suitable for smooth mean structure, but cannot capture local abrupt change (example?)

  Q: how to relax the smoothness restriction?

  ⇒ one solution: broken line regression.



How about a 2nd-order polynomial ?

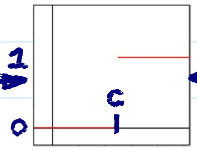
• Q: when to use broken line regression?

  ⇒ believe that different regression models apply in different regions of data, and the fit should be continuous at the broken points → but not differtiable ⇒ smoothness relaxed

• suppose the break occurs at the *known* value c, define the base function (where c is called a *knot*):

$$d_c(x) = \begin{cases} 1, & \text{if } x > c, \\ 0, & \text{if } x \leq c. \end{cases}$$

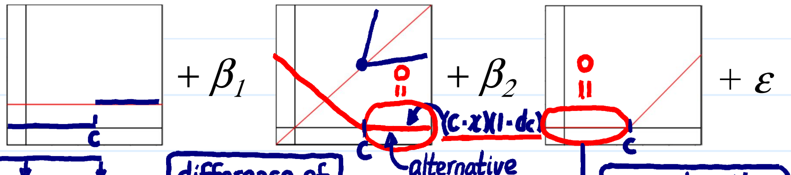
known function



cf. dummy variable (LNp.8-11) ← categorizing

- model: $y = \beta_0 + \beta_1 \underline{x} + \beta_2 \underline{(x-c)d_c(x)} + \varepsilon$ — a known function → still a linear model

intercept and slope of the line on region $x \le c$



$= \beta_0 \quad + \beta_1 \quad + \beta_2 \quad + \varepsilon$

Q: Why can the base function $(x-c)d_c(x)$ reduce "global" influence of data on the fit?

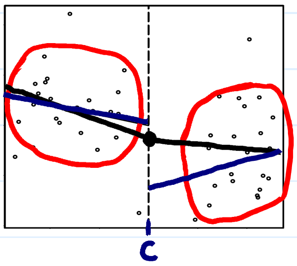
consider its influence on $\hat{\beta} = (X^T X)^{-1} X^T Y$

What is the contribution of the term $\hat{\beta}_2 (x-c) d_c(x)$ on $\hat{y}$?

difference of the 2 slopes $= \beta_2$

$$E(y) = \begin{cases} \beta_0 + \beta_1 x, & \text{if } x \le c, \\ (\beta_0 - \beta_2 c) + (\beta_1 + \beta_2) x, & \text{if } x > c, \end{cases}$$

- ➢ the two lines meet at c ⇒ continuous fit
- ➢ notice only 3 parameters in the model ⇒ one degree of freedom is saved because of the continuity restriction



change point problem

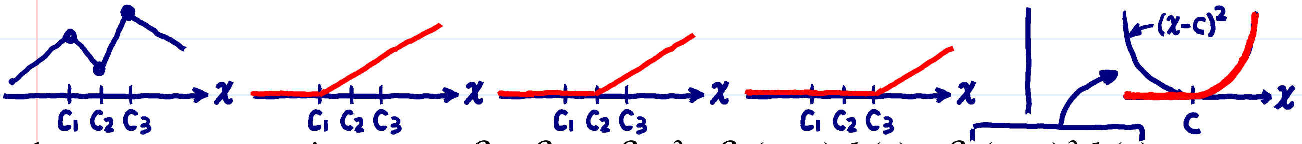
⊙ when c is unknown

4 parameter model → $E(y) = \beta_0 + \beta_1 d_c(x) + \beta_2 x + \beta_3 (x-c) d_c(x)$ — interaction

⇒ can regard c as a parameter ⇒ not a linear model any more

⇒ can be estimated by nonlinear regression

- generalization: more knot points or more predictors ⇒ define more base functions



- broken curve regression: $y = \beta_0 + \beta_1 \underline{x} + \beta_2 \underline{x^2} + \beta_3 \underline{(x-c)d_c(x)} + \beta_4 \underline{(x-c)^2 d_c(x)} + \varepsilon$

❖ **Reading**: Faraway (2005, 1st ed.), 7.2.1 ❖ **Further reading**: D&S, 14.3

regression spline and LOWESS

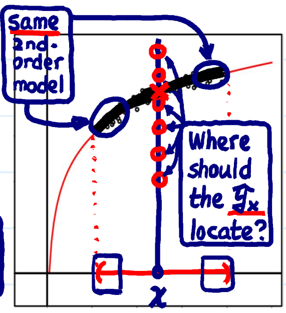
- **concept**: fitting using local points ⇔ better fit

develop $\hat{y}_x$ at any x

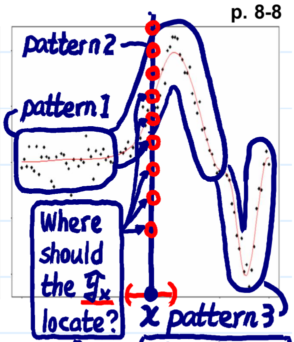
all data — $y_i \xleftrightarrow{\text{closer}} \hat{y}_i$

using global points ⇔ smoothness

Q: which should you choose for your data? using local points or global points?



2nd-order model $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$



Should we include the information in the data of pattern 1 to predict $\hat{y}_x$

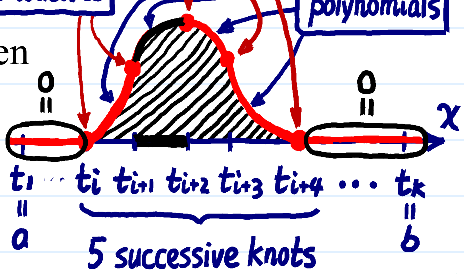
- regression spline
  - ➢ **concept**: define different *base functions* to fit data
    - power function: smoothness, but non-zero across the whole range — use global points
    - broken line: lesser smooth, but localizing the influence of data point
    - B-spline: compromise between smoothness and local influence

non-differentiable at point c

  - ➢ cubic B-spline base functions: $g_1, \ldots, g_{k-4}$ defined on an interval $[a, b]$ with knot-points $t_1 \le \ldots \le t_k$ (no need to be equally spaced) satisfying:
    1. non-zero on interval defined by 5 successive knots and zero elsewhere ⇒ local influence
    2. a cubic polynomial for each sub-interval between successive knots — infinitely differentiable
    3. continuous, and continuous in its 1st and 2nd derivatives at each knot ⇒ smoothness
    4. integrate to one over its support

continuous in 1st & 2nd derivatives

different cubic polynomials



0 ||
0 ||
x
t_1 || a
t_i t_{i+1} t_{i+2} t_{i+3} t_{i+4} ... t_k || b
5 successive knots

- ➢ base function at the ends of the interval are defined differently to ensure continuity
- ➢ regress y on these B-spline base functions, i.e.,

possesses properties 2 & 3 in LNp.8-8

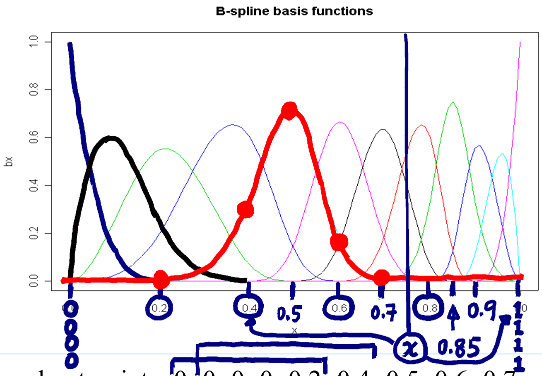
$$y = \underbrace{\Sigma_i\, \beta_i\, g_i(x)} + \varepsilon$$

knots are known

(Note: g_i's are known functions for given t's $\Rightarrow$ it's still a linear model)

non-parametric approach in statistic : dim(parameters) = ∞

non-parametric regression → distribution of ε ; mean structure $E(y_x)$

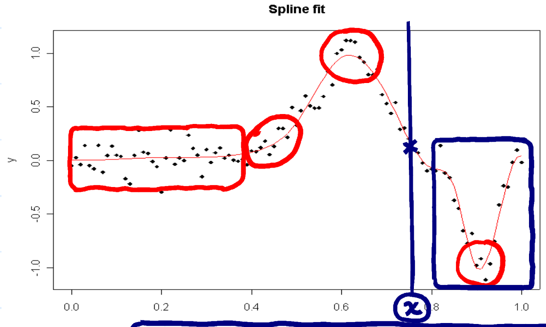


knot points: 0, 0, 0, 0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 1, 1, 1, 1

Q: why the knots are dense in some region?

- LOWESS (LOcally WEighted Scatterplot Smoothing)
  - ➢ Recall: in previous models, # of parameters are finite

Note dim($\hat{y}$) = dim(Ω) = # of βi's



  - ➢ nonparametric regressions:

model: $y = f(x_1, \ldots, x_m) + \varepsilon$

$f = \sum_{i=1}^{p} \beta_i g_i$, e.g.

- ▪ parametric regression: assume f is from a family of functions, in which # of parameters is finite cf.
- ▪ nonparametric regression: assume f is smooth only (# of parameters = ∞)

may consider the mean $E(y_x)$ as a free parameter at every x

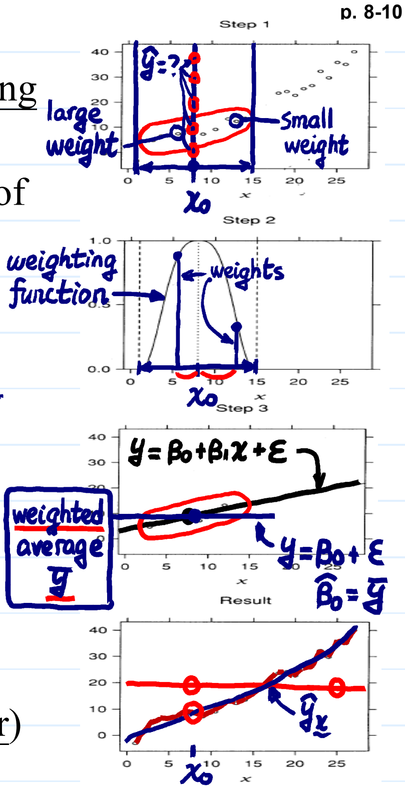
- ➢ method: (see example) → at each x, $\hat{y}_x$ = ?
  1. select a window (⇒ local information), and a weighting function (⇒ closer points, more contribution)
  2. use weighted (closer points, higher weights) average of y_i's in the window to estimate fitted value cf. Recall WLS
  3. repeat as the window moves

(i) fix window width (ii) fix # of nearest neighbors

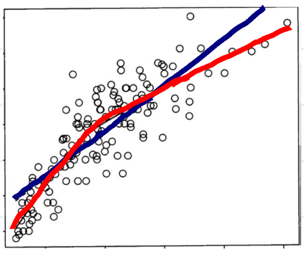
no sensible estimation of some meaningful βi's

- ▪ width of window is an issue (larger window, smoother curve)
- ▪ can plot fitted value for a variety of widths and pick best result
- ▪ different width of window to estimate f along the range of x
- ▪ sensitive to outliers: use median, not average

too large window
⇒ 規律→隨機
⇒ underfitting
too small window
⇒ 隨機→規律
⇒ overfitting



- ➢ LOESS: change step 2 to locally weighted (1st or 2nd order) polynomial regression (see example)
- ➢ difficult if extrapolation is required (same difficulty in regression spline) ← cf. → parametric model : $\hat{y}_x = \sum_i \hat{\beta}_i\, g_i(x)$
- ➢ nonparametric regressions are useful for fitting a curve for residual plots, added variable plots, partial residual plots



❖ **Reading**: Faraway (2005, 1st ed.), 7.2.3; Faraway (2006), chapter 11