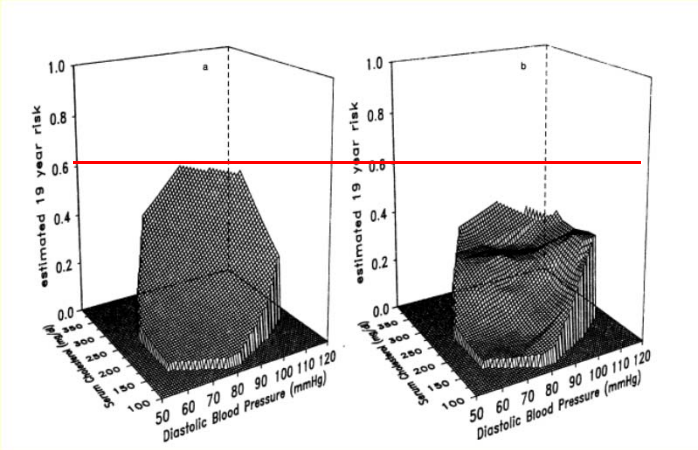
5. Introduction to non-parametric curve fitting:

Loess, kernel regression,
reproducing kernel methods,
neural networks

Distinctive aspects of non-parametric fitting

- **Objectives:** investigate patterns free of strictures imposed by parametric models
- Can produce surprising results
- Regression coefficients appear but (typically) do not have an obvious interpretation
- Often have very good predictive performance in cross-validation
- Tuning methods similar to those for parametric methods

Example: thin-plate splines



$$f(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \sum_{j=1}^{N} \alpha_j \left[\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 \right] \log\left[\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 \right]$$

Risk of heart attack after 19 years as a function of cholesterol level and blood pressure. Left: logistic regression model. Right: thin plate spline fit. Wahba (2007)

LOESS REGRESSION:

Non-parametric exploration of inbreeding depression for yield and somatic cell count in Jersey cattle

AN OVERVIEW OF LOWESS REGRESSION

1) DATA POINTS $(x_i, y_i);\ i = 1, 2, \ldots, n$

2) SPANNING PARAMETER $f;\ 0 < f < 1$

$k = fn;\ k =$ LARGEST INTEGER $\leq fn$

3) FOR EACH x_0 FIND k POINTS x_i "CLOSEST" TO x_0

$N(x_0)$= NEIGHBORHOOD OF k POINTS

4) COMPUTE $\Delta(x_0) = \max_{x_i \subset N(x_0)} |x_o - x_i|$

5) TO EACH (x_i, y_i); $x_i \subset N(x_0)$ ASSIGN WEIGHT

$$w_i(x_0) = \left\{1 - \left[\frac{|x_o - x_i|}{\Delta(x_0)}\right]^3\right\}^3$$

6) FIT BY WEIGHTED LEAST-SQUARES

$$\sum_{i=1}^{k} w_i(x_0)(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

RETURN $\hat{y}(x_0) = \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2$

7) REPEAT FOR EACH OF THE x_0

ROBUST LOWESS

- STANDARD LOWESS NOT ROBUST

  ➔ BASED ON LEAST-SQUARES WEIGHTS

- BI-SQUARE LOWESS

  ➔RE-WEIGHT POINTS ACCORDING TO RESIDUAL
  ➔IF RESIDUAL LARGE, WEIGHT IS DECREASED

1) FIT DATA USING STANDARD LOESS

2) CALCULATE LOESS RESIDUALS $y_i - \hat{y}_i$

3) COMPUTE $\hat{q}_{\frac{1}{2}} = median|y_i - \hat{y}_i|$

4) CALCULATE BI-SQUARE ROBUST WEIGHTS

$$r_i = \left\{ 1 - \left[\frac{y_i - \hat{y}_i}{6\hat{q}_{\frac{1}{2}}} \right]^2 \right\}^2$$

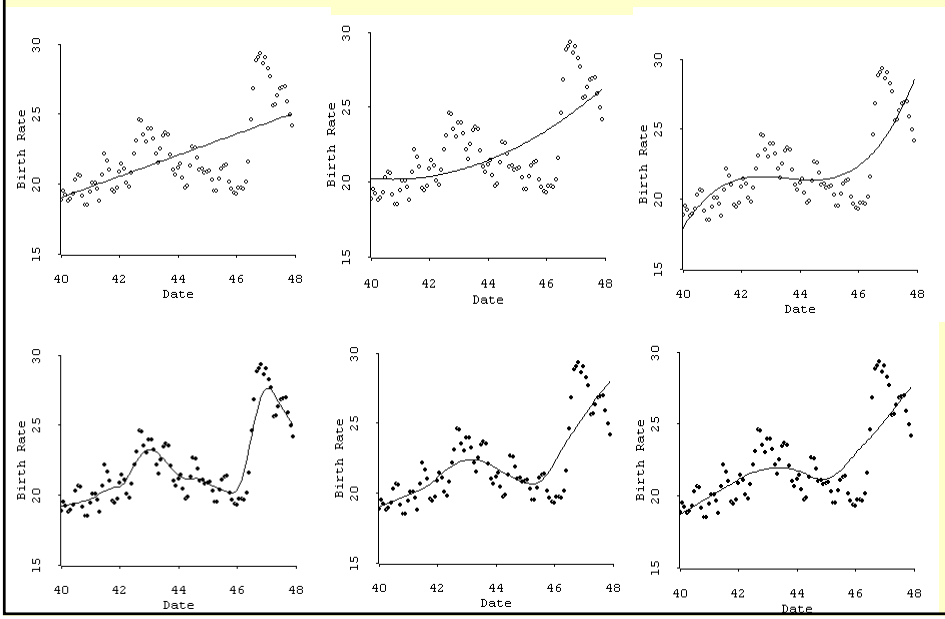
5) REPEAT LOESS WITH WEIGHTS $r_i w_i(x_0)$

6) REPEAT 2-5 UNTIL LOESS CURVE ”CONVERGES”

Example

- Birth rate in US population
  (U. S. Department of Health, Education and Welfare)
- n=96
- births per 1000 US population
- during 1940-47

Top > Ordinary Least Squares with 1st, 2nd & 3rd degree polynomial
Bottom > LOWESS fit with f = .2, f=.4 & f=.6



GALTON'S BEND
(Wachsmuth et al. 2003, Am. Stat.)

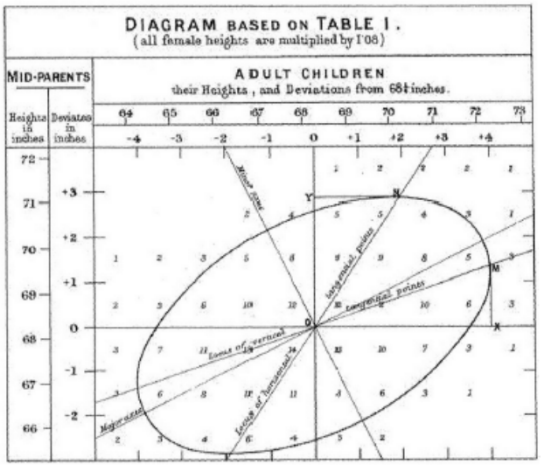


Figure 1. Galton's fitted regression model.

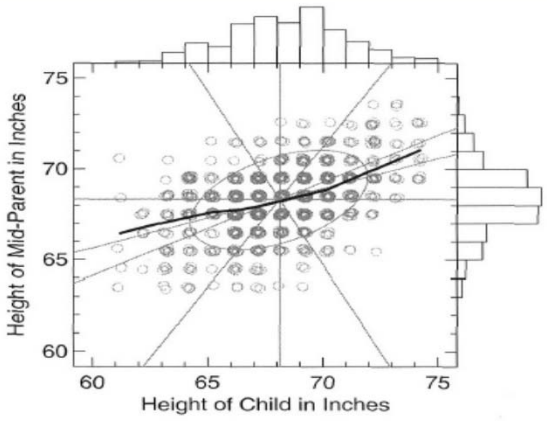


Figure 2. SYSTAT plot of Galton's Data with loess fit.

The dark curve in the center of the plot is a *loess* smoother (Cleveland and Devlin 1988). The smoother suggests that the relation between parent and child stature is not linear. There is a bend in the curve somewhere around the average height of approximately 68 inches for parents and children. A two-stage piecewise linear regression (Hinkley 1971) identifies a breakpoint at around 70 and finds it highly significant ($p < .0001$).

A possibility is that Galton ignored concealed heterogeneity

Does the bend disappear by disaggregation of the sample?
Analysis of data from Pearson and Lee (1903)

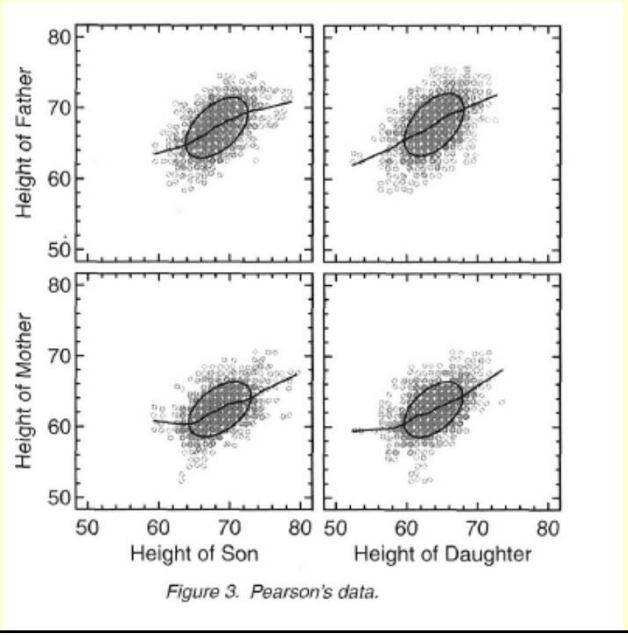


Figure 3. Pearson's data.

BEND
STILL
THERE!

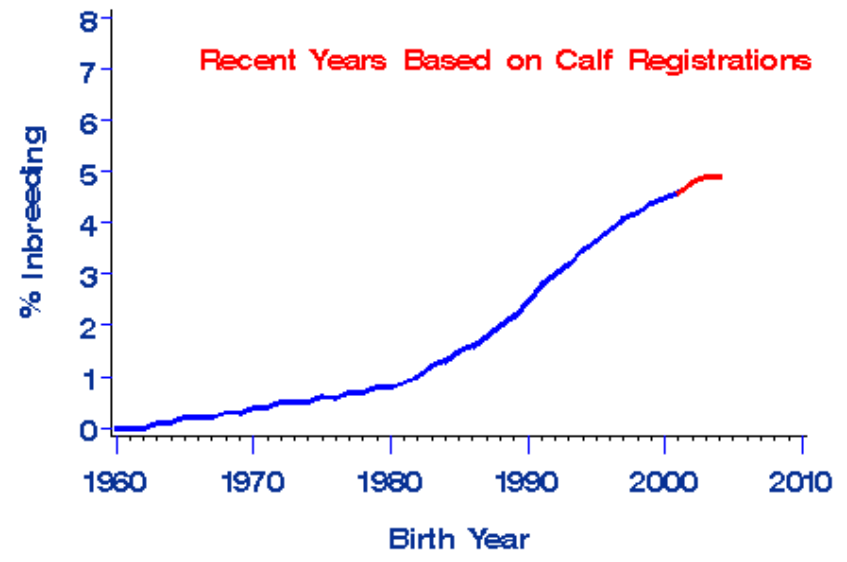
Wachsmuth et al. (2003) write:

In their search for universal hereditary laws, Galton and Pearson were driven by the linear model and the normal distribution because the associated parameters had scientific meaning for them that went beyond mere description.

INBREEDING DEPRESSION

- Examine relationships of yield (milk, protein, fat) and somatic cell score (SCS) with inbreeding coefficient (F) using field data from US Jerseys
- Use REML, BLUP and "local regression" method (LOESS) for this purpose

LEVEL OF INBREEDING IN HOLSTEINS, USA



- Relationship between mean value of a quantitative trait and inbreeding coefficient (F) expected to be linear under dominance
- Not so if epistatic interactions between dominance effects exist

(Crow & Kimura, 1970)

ONE-LOCUS MODEL

GENOTYPE (X)	A_1A_1	A_1A_2	A_2A_2
FREQUENCY	$p_1^2(1-F)+p_1F$	$2p_1p_2(1-F)$	$p_2^2(1-F)+p_2F$
PHENOTYPE	$\mu - A$	$\mu + D$	$\mu + A$

$$
\begin{aligned}
E(X) &= \mu + A(p_2 - p_1) + 2p_1p_2D - 2p_1p_2DF \\
&= \alpha + \beta F \\
&= \alpha - \beta(1 - F - 1) \\
&= (\alpha + \beta) - \beta(\%Heterozygosity)
\end{aligned}
$$

ADDITIVE MODEL WITH F (or H) AS COVARIATE➔**CONTRADICTORY**

TWO (UNLINKED) LOCI: NO EPISTASIS

Joint frequencies are product of marginal frequencies

GENOTYPE	□	A_1A_1	A_1A_2	A_2A_2
□	FREQUENCY	$p_1^2(1-F)+p_1F$	$2p_1p_2(1-F)$	$p_2^2(1-F)+p_2F$
B_1B_1	$r_1^2(1-F)+r_1F$	$\mu-A-B$	$\mu+D_A-B$	$\mu+A-B$
B_1B_2	$2r_1r_2(1-F)$	$\mu-A+D_B$	$\mu+D_A+D_B$	$\mu+A+D_B$
B_2B_2	$r_2^2(1-F)+r_2F$	$\mu-A+B$	$\mu+D_A+B$	$\mu+A+B$

$$
\begin{aligned}
E(X) &= \mu + A(p_2 - p_1) + B(r_2 - r_1) \\
&\quad + 2p_1p_2D_A + 2r_1r_2D_B \\
&\quad - 2(p_1p_2D_A + r_1r_2D_B)F \\
&= \alpha' + \beta' F
\end{aligned}
$$

TWO (UNLINKED) LOCI: EPISTASIS

GENOTYPE	□	A_1A_1	A_1A_2	A_2A_2
□	FREQUENCY	$p_1^2(1-F)+p_1F$	$2p_1p_2(1-F)$	$p_2^2(1-F)+p_2F$
B_1B_1	$r_1^2(1-F)+r_1F$	$\mu-A-B+$I	$\mu+D_A-B-$L	$\mu+A-B-$I
B_1B_2	$2r_1r_2(1-F)$	$\mu-A+D_B-$K	$\mu+D_A+D_B+$J	$\mu+A+D_B+$K
B_2B_2	$r_2^2(1-F)+r_2F$	$\mu-A+B-$I	$\mu+D_A+B+$L	$\mu+A+B+$I

- ALLELES AT ***A*** and ***B*** LOCI SAME SUBSCRIPT➔ADD **I**
  (ADDITIVE X ADDITIVE)
- HOMOZYGOUS AT ***A*** HETEROZYGOUS AT ***B***➔**SUBSTRACT AND ADD K**
  HOMOZYGOUS AT ***B*** HETEROZYGOUS AT ***A***➔**SUBSTRACT AND ADD L**
  (ADDITIVE X DOMINANCE)
- HETEROZYGOUS AT ***A*** AND **B**➔**ADD J**
  (DOMINANCE X DOMINANCE)

I,J,K,L: parameters (4 d. freedom)

Mean value under dominance x dominance epistasis

$$
\begin{aligned}
E(X) = {} & \mu + A(p_2 - p_1) + B(r_2 - r_1) + 2p_1p_2D_A + 2r_1r_2D_B \\
& + I(p_1 - p_2)(r_1 - r_2) + 2Lp_1p_2(r_1 - r_2) + 2Kr_1r_2(p_1 - p_2) \\
& + 4Jp_1p_2r_1r_2 \\
& - 2[p_1p_2D_A + r_1r_2D_B + Lp_1p_2(r_1 - r_2) + Kr_1r_2(p_1 - p_2) + 4Jp_1p_2r_1r_2]F \\
& + (4Jp_1p_2r_1r_2)F^2 \\
= {} & \alpha'' + \beta''F + \gamma F^2
\end{aligned}
$$

- Dominance, additive x dominance, and dominance x dominance intervene in **linear** regression
- Epistasis without dominance does not enter into mean-F relationship
- Dominance x dominance intervenes in second-order regression

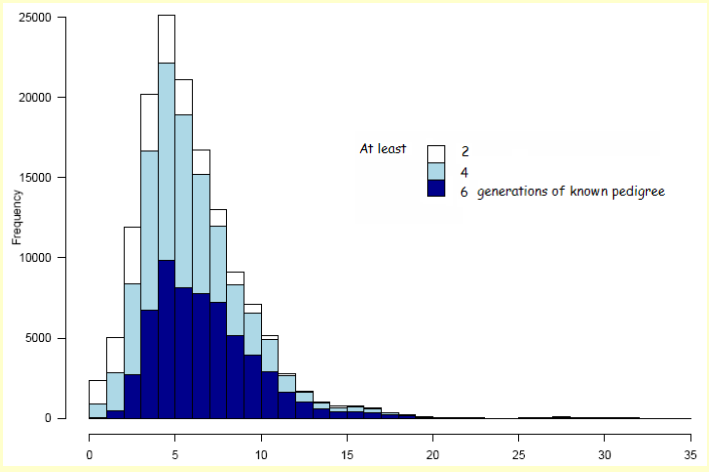
DATA

- First lactation records (herds) on 59,778 (1,142) Jersey cows
- 6 generations of known pedigree
- First calving between 1995 and 2000

Distribution of F

- F calculated from all known pedigree information
- F ranged between 0 and 34%
- Median F = 6.25%

Histogram of F values



Procedures

- Fit linear models without F as covariate
- Compute EBLUP residuals from these models
- Fit nonparametric regression to EBLUP residuals in order to obtain nonparametric lines describing relationship between performance and inbreeding level

Linear Models

Model

$$y_{ijk} = HYS_i + AGE_j + \beta_1 (D_{ijk} - \bar{D}) + a_k + e_{ijk}$$

y_{ijk} = somatic cell score (SCS), milk, protein, or fat yield;

HYS_i = fixed effect of herd-year-season (i = 1,2,….,12276 for DS2; 11158 for DS4 or 6406 for DS6, with seasons classes January–April, May–August, September–December);

AGE_j = fixed effect of age at calving class; j = 1,2,….,6 (< 617, 617- 716, 717-816, 817-916, 917-1016, or >1016 days of age);

β_1 = fixed regression coefficient of performance on days in milk;

D_{ijk} = days in milk for animal k in herd-year-season i and age of calving class j;

$\bar{D}$ = 263;

a_k = random additive genetic effect of animal k, and

e_{ijk} = random residual.

Linear Model Assumptions

- Genetic and residual effects assumed mutually independent, with $\boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2_e)$ and $\boldsymbol{a} \sim \mathrm{N}(\mathbf{0}, \mathbf{A}\sigma^2_a)$ where A is the additive relationship matrix ($1 + F_k$ in the k^{th} diagonal position, F_k is the inbreeding coefficient of animal *k*)

Nonparametric regression

- Fit LOESS regression to BLUP residuals with F as covariate
- Vary spanning parameter & degree of local polynomial
- Plot fitted values of residuals against F

LOESS

(Fitting done by **locally weighted** least squares)

- $\tilde{\varepsilon}_{ij}$ is LOESS fit using only residuals in the neighborhood of F_i, i=1,2,…n

  (i=1,2,…,n animals; j=1,…,4 traits)

- Size of neighborhood determined by $f = \frac{q}{n}$

  q = number of points in neighborhood

  n = total number of points

"Robust" LOESS

Weights assigned to $\hat{\varepsilon}_{ijk}$:

$$\Rightarrow \; w_{ijk}^{[t+1]} = w_{ijk}^{[t]} \cdot \delta_{ijk}^{[t]}$$

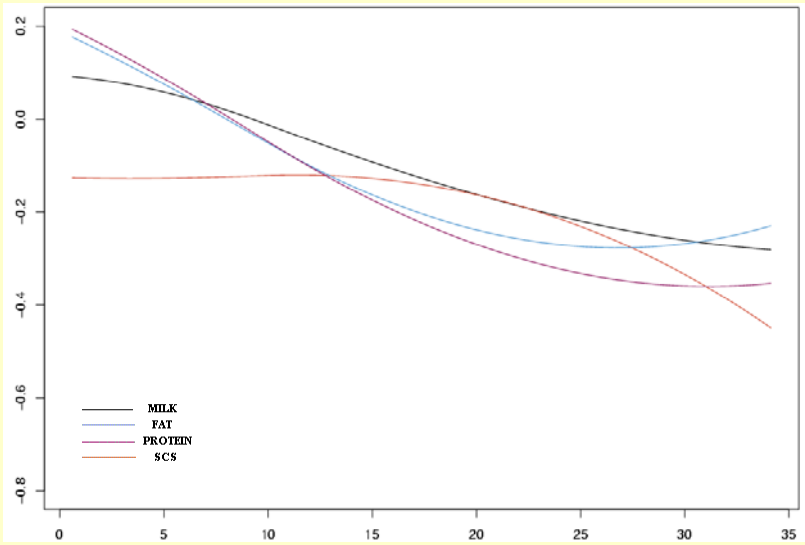
t=1,2,3,4

I) $$w^{[1]}_{ijk} = [1 - (\frac{F_k - F_i}{\max(F_l - F_i)})^3]^3 \qquad l = 1,2,...q$$

II) $$\delta^{[t]}_{ijk} = [1 - (\frac{\tilde{\varepsilon}_{ijk} - \hat{\varepsilon}_{ijk}}{6 \cdot med})^2]^2$$

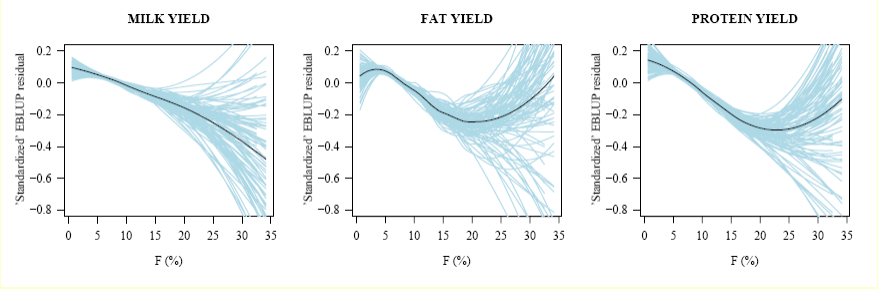
med = median of all $(\tilde{\varepsilon}_{ijk} - \hat{\varepsilon}_{ijk})$

Cows with at least 6 generations of known pedigree $f = 1$



2nd degree local polynomial

"Robust" original (black) with bootstrap (light blue) LOESS curves of yields for US Jerseys with at least 6 generations of known pedigree, based on medians of EBLUP residuals (y-axis = $\hat{e}_{ijk} / \hat{\sigma}_a$)



f=0.9 *f=0.5* *f=0.9*

2nd degree local polynomial

Conclusions

- LOESS analysis suggested local relationships.
- Effects of inbreeding seem nil, until for F values up to ~7%
- Effects of inbreeding not accounted well by additive models
- Results may be confounded by effects of selection that are unaccounted for

Kernel Regression

$$y_i = g(\mathbf{x}_i) + e_i;\ i = 1, 2, \ldots, n$$

where:

- y_i is the measurement taken on individual i
- $\mathbf{x}_i$ is a $p \times 1$ vector of observed SNP genotypes
- $g(.)$ is some unknown function relating genotypes to phenotypes.
- Set $g(\mathbf{x}_i) = E(y_i \mid \mathbf{x}_i)$ = conditional expectation function
- $e_i \sim (0, \sigma^2)$ is a random residual

Conditional expectation function

$$g(\mathbf{x}) = \int y \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} dy$$

$$= \frac{\int y\, p(\mathbf{x}, y) dy}{p(\mathbf{x})}.$$

Non-parametric estimator of density of **x**

$$\hat{p}(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$$

"Focal point"

"Kernel", possibly a probability density function with some bandwidth parameter h

We would like:

$$\int_{-\infty}^{\infty} \hat{p}(\mathbf{x})\, \mathbf{dx} = \frac{1}{nh^p} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) \mathbf{dx} = 1$$

Implying ➔

$$\int_{-\infty}^{\infty} \frac{1}{h^p} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) \mathbf{dx} = 1$$

➔ Similarly, can form non-parametric estimator of joint density

$$\hat{p}(\mathbf{x}, y) = \frac{1}{nh^{p+1}} \sum_{i=1}^{n} K\left(\frac{y_i - y}{h}\right) K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$$

➔ Recall

$$g(\mathbf{x}) = \int y \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} dy$$

$$= \frac{\int y\, p(\mathbf{x}, y) dy}{p(\mathbf{x})}.$$

← ESTIMATE NUMERATOR

← ESTIMATE DENOMINATOR

Estimate numerator

$$\int y\,\hat{p}(\mathbf{x},y)dy = \int y\,\frac{1}{nh^{p+1}}\sum_{i=1}^{n} K\left(\frac{y_i - y}{h}\right)K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)dy$$

$$= \frac{1}{nh^{p}}\sum_{i=1}^{n}\left[\frac{1}{h}\int y\,K\left(\frac{y_i - y}{h}\right)dy\right]K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right).$$

Let $z = \frac{y-y_i}{h}$, so that $dy = hdz$ and

$$\frac{1}{h}\int y\,K\left(\frac{y_i - y}{h}\right)dy = \frac{1}{h}\int (y_i + hz)K(z)hdz$$

$$= \int (y_i + hz)K(z)dz$$

$$= \int y_i K(z)dz + h\int zK(z)dz$$

$$= y_i \int K(z)dz + hE(z).$$

K(.) can be constructed such that:

$$\int K(z)dz = 1 \text{ and } E(z) = \int zK(z)dz = 0$$

Then: $\frac{1}{h}\int y\,K\left(\frac{y_i - y}{h}\right)dy = y_i$

Estimator of numerator is

$$\int y\,\hat{p}(\mathbf{x},y)dy = \frac{1}{nh^{p}}\sum_{i=1}^{n} y_i K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)$$

Forming non-parametric estimator of conditional expectation

$$\hat{E}(y \mid \mathbf{x}) = \hat{g}(\mathbf{x}) = \frac{\int y\, \hat{p}(\mathbf{x},y)dy}{\hat{p}(\mathbf{x})}$$

$$\hat{E}(y \mid \mathbf{x}) = \hat{g}(\mathbf{x}) = \frac{\frac{1}{nh^p}\sum_{i=1}^{n} y_i K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}{\frac{1}{nh^p}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}$$

$$= \frac{\sum_{i=1}^{n} y_i K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)} = \sum_{i=1}^{n} w_i(\mathbf{x}) y_i$$

Nadaraya-Watson estimator
(weighted average)

$$w_i(\mathbf{x}) = \frac{K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}$$

Relationship between
Income and age
(Chu and Marron, 1991)

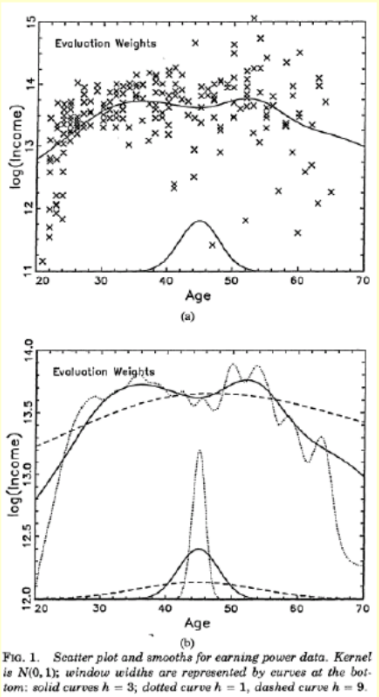


FIG. 1. *Scatter plot and smooths for earning power data. Kernel is N(0, 1); window widths are represented by curves at the bottom: solid curves h = 3; dotted curve h = 1, dashed curve h = 9.*

h=9 local features
Disappear (dashes)

h=1 lots of variation
(dots)

Bandwidth can be gauged by, e.g., cross-validation

$$CV(h) = \frac{\sum_{i=1}^{n} \left[y_i - \hat{g}_{i,-i}(\mathbf{x}_i | h) \right]^2}{n}$$

- Create a grid of *h* values
- For each value compute the CV mean squared error (above is leave-one-out, but this may not be best)
- Use the *h* value which minimizes CV(.)