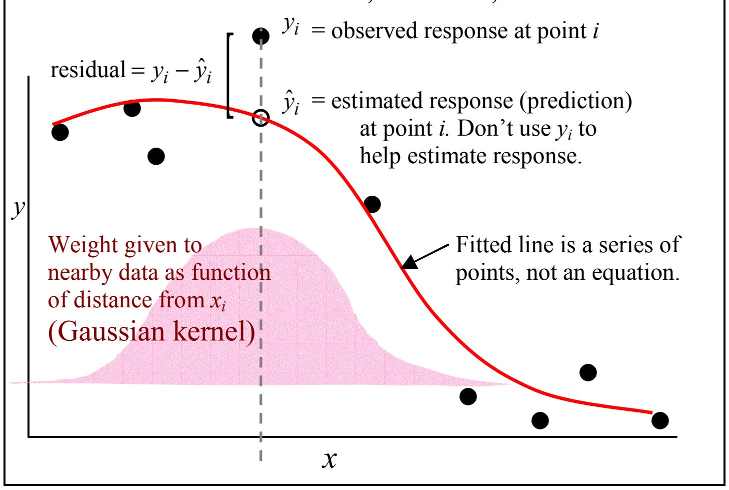
Basics of Nonparametric Regression

with leave-one-out cross validation

Observations, Estimates, and Residuals



Symbols

x = predictor

y = response

$\hat{y}$ = estimated response

$\bar{y}$ = mean response

w = weight

m = number of predictors

n = number of sample units

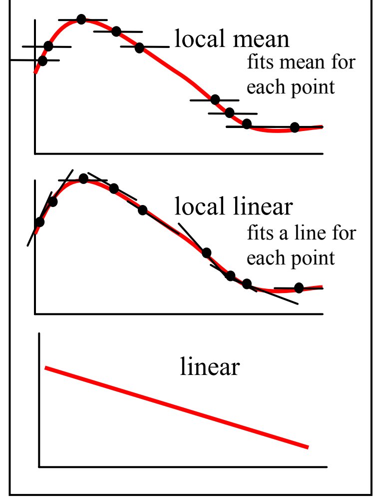
i = subscript for sample units

j = subscript for predictors

$\sum_{i=1}^{n}$ = sum from 1 to n

$\prod_{i=1}^{n}$ = multiply from 1 to n

Model Forms



Local mean NPMR (nonparametric multiplicative regression)

$$\hat{y}_v = \frac{\sum_{i=1, i \neq v}^{n} y_i \left(\prod_{j=1}^{m} w_{ij} \right)}{\sum_{i=1, i \neq v}^{n} \left(\prod_{j=1}^{m} w_{ij} \right)}$$

v is the target point

Neighborhood size

(sum of weights, amount of data used to estimate $\hat{y}_i$):

$$n_i^* = \sum_{i=1}^{n} \left(\prod_{j=1}^{m} w_{ij} \right)$$

Model Fit

Model Fit. Data point i excluded when estimating $\hat{y}_i$ (leave-one-out cross validation).

Quantitative response: xR^2

$$xR^2 = \text{cross } R^2 = 1 - \frac{\text{Residual sum squares}}{\text{Total sum of squares}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

1.0 = perfect fit 0.0 = no relationship

< 0 = model is worse than using $\hat{y}_i$ overall mean for all y_i

Binary response: log$B$ = log likelihood ratio

$$B_{12} = \frac{\text{likelihood of data with fitted model } M_1}{\text{likelihood of data with naive model } M_2} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$$

Likelihoods from binomial: $p(\mathbf{y}|M) = \prod_{i=1}^{n} \hat{y}_i^{\,y_i} (1 - \hat{y}_i)^{1 - y_i}$

log$B$ = 0 = fitted model no better than naive model

log$B$ > 0 = better than naive model; open ended

log$B$ < 0 = worse than naive model

drop in deviance = χ^2 = 4.605 log$B$

ave$B$ = B/n = $10^{(\log B)/n}$ = ave. contribution of sample unit to likelihood ratio (measure of fit independent of sample size)

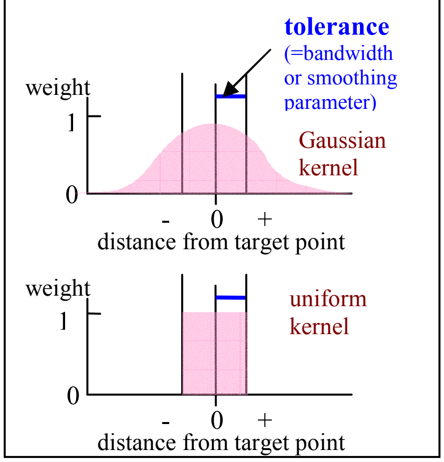
ave$B$ = 1.0 = no better than naive model

ave$B$ < 1 = worse than naive model

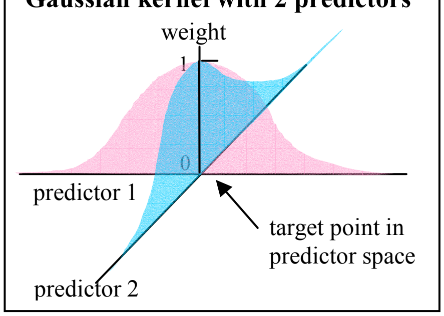
ave$B$ = 1.2 = each sample unit contributes an average of 20% improvement in likelihood of fitted model over naive model.

Weights (Kernel Types)

weights data by proximity to target



Gaussian kernel with 2 predictors



Tips

Confidence intervals and variability bands: use bootstrap resampling

For other measures of fit (AUC, chi-square, etc.): use *Evaluate Selected Model*

To achieve a smoother, more continuous response curve: increase the minimum average neighborhood size during the model fitting phase.

Help

F1 for help

FAQ and updates: HyperNiche.com

MjM Software

PO Box 129

Gleneden Beach, OR 97388 USA

mjm@centurytel.net

Sensitivity Analysis. Relative importance of particular predictors in the model. Nudges up and down the observed values for individual variables, and measures the resulting changes in estimates of the response for each data point.

$$Sensitivity = \frac{\text{mean difference in response / range in response}}{\text{difference in predictor / range in predictor}}$$

Sensitivity = 1.0 means that a 10% change in the predictor would, on average, produce a 10% change in the response.

Sensitivity = 0.0 means no change in the response.