POLITECNICO
MILANO 1863

Quality Data Analysis

Linear models - part 3

Bianca Maria Colosimo – biancamaria.colosimo@polimi.it

Trevor Hastie. Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning.

Linear models – it is the starting point for many other extensions

1) Basis function: represent the function by a linear expansion

$$Y = f(x) + \varepsilon$$

$$f(x) = \sum_{k=1}^{K} c_k \phi_k(x) = c_1\phi_1(x) + c_1\phi_1(x) + \ldots + c_K\phi_K(X)$$

- Power series: x, x^2, x^3, …
- Fourier series (for periodic functions): 1, sin (wx), cos (wx), sin (2wx), cos (2wx),….
- Spline basis (flexible, computationally convenient, non periodic)

Source: Ramsey & Siverman – ***Functional data Analysis***

Cubic spline

Define a set of knots $\xi_1 < \xi_2 < \cdots < \xi_K$.

 We want the function *f* in the model $Y = f(x) + \varepsilon$ to:

1. be a cubic polynomial between every pair of knots ξ_i, ξ_{i+1}.
2. Be continuous at each knot.
3. Have continuous first and second derivatives at each knot.

It turns out, we can write f in terms of K + 3 basis functions:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 h(x, \xi_1) + \ldots + \beta_{K+3} h(x, \xi_K)$$

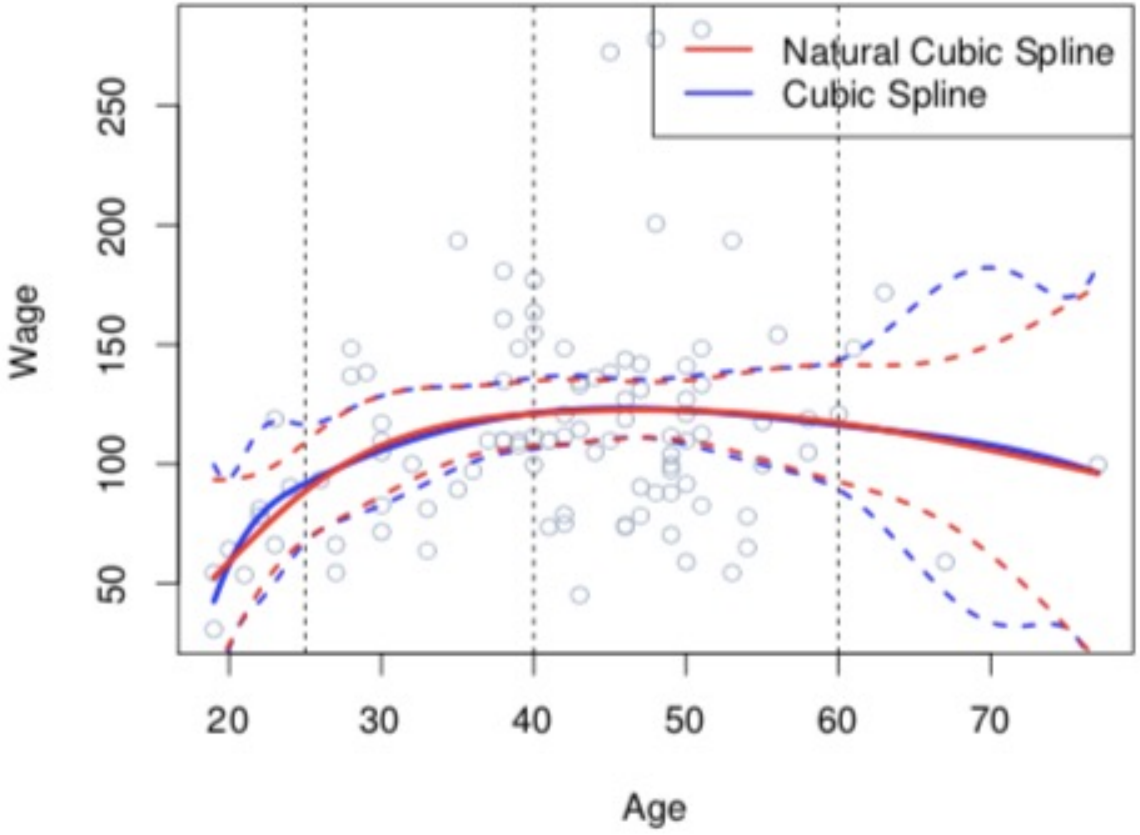
where,

$$h(x,\xi) = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \textit{otherwise} \end{cases}$$

Natural Cubic spline

Spline which is linear instead of cubic for

$X < \xi_1$, $X > \xi_K$.



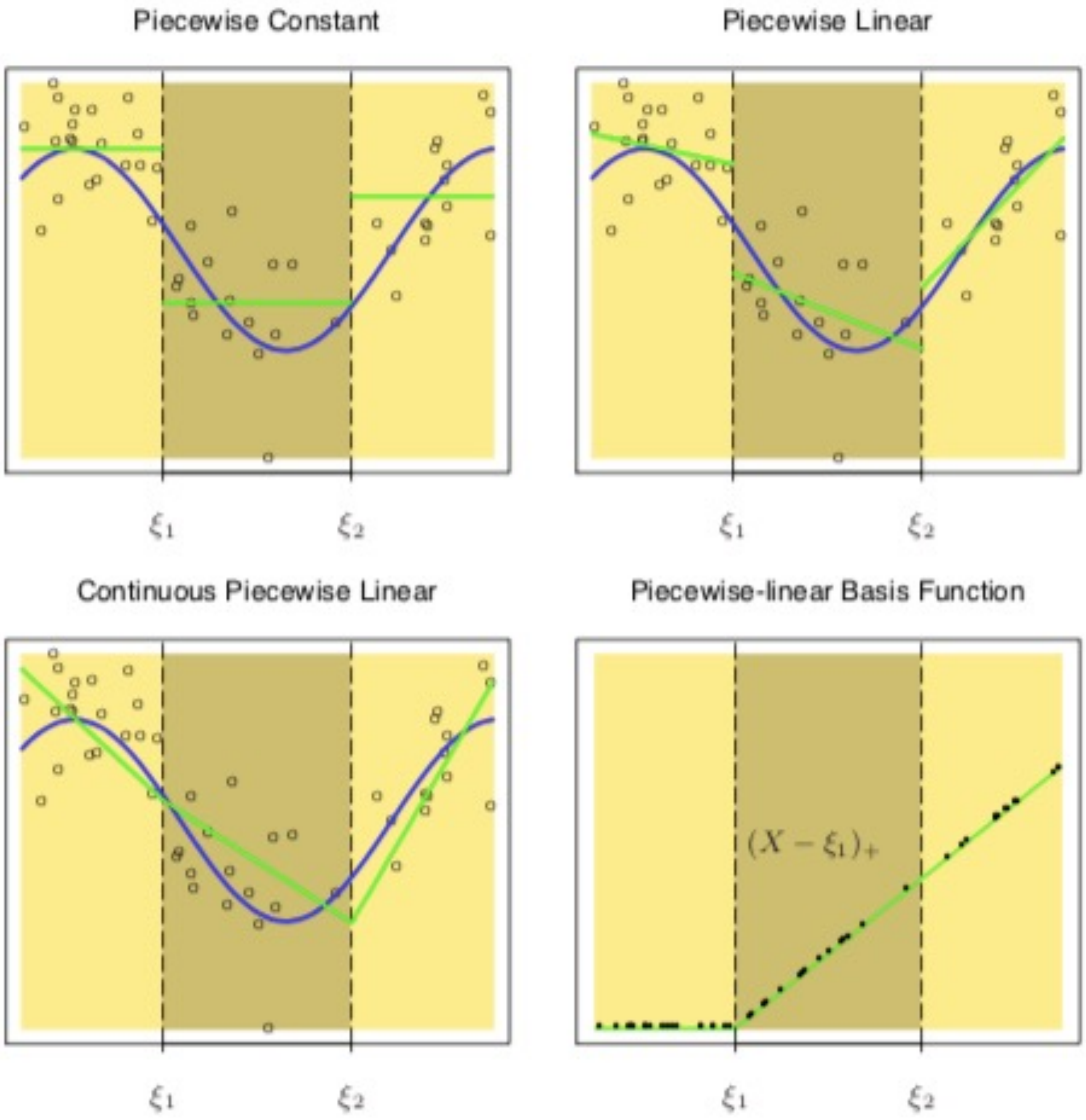
Cubic Spline

FIGURE 5.1. The top left panel shows a piecewise constant function fit to some artificial data. The broken vertical lines indicate the positions of the two knots ξ_1 and ξ_2.

The blue curve represents the true function, from which the data were generated with Gaussian noise. The remaining two panels show piecewise linear functions fit to the same data—the top right unrestricted, and the lower left restricted to be continuous at the knots.

The lower right panel shows a piecewise– linear basis function, $h_3(X) = (X - \xi_1)_+$, continuous at ξ_1.

The black points indicate the sample evaluations $h_3(x_i)$, i = 1,...,N.



Cubic spline

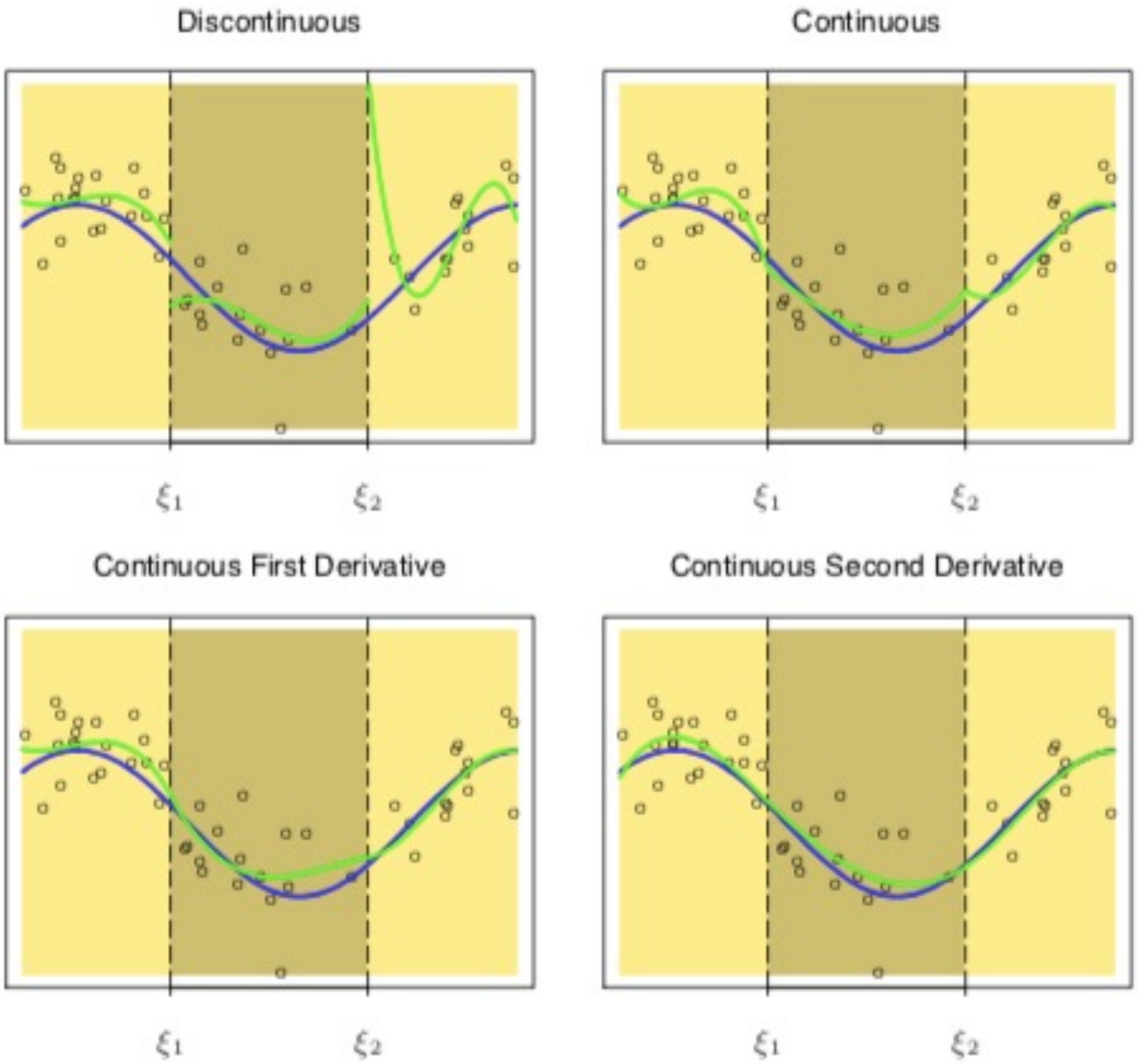


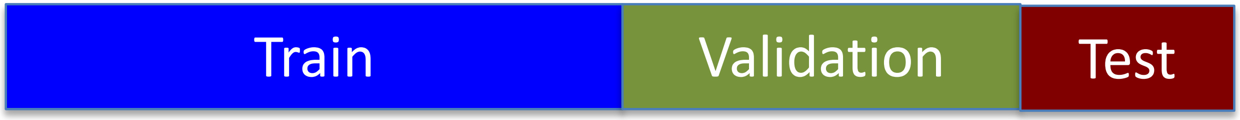
FIGURE 5.2. *A series of piecewise-cubic polynomials, with increasing orders of continuity.*

Cross-validation

Model selection: estimating the performance of different models in order to choose the best one.

Model assessment: having chosen a final model, estimating its prediction error (generalization error) on new data.

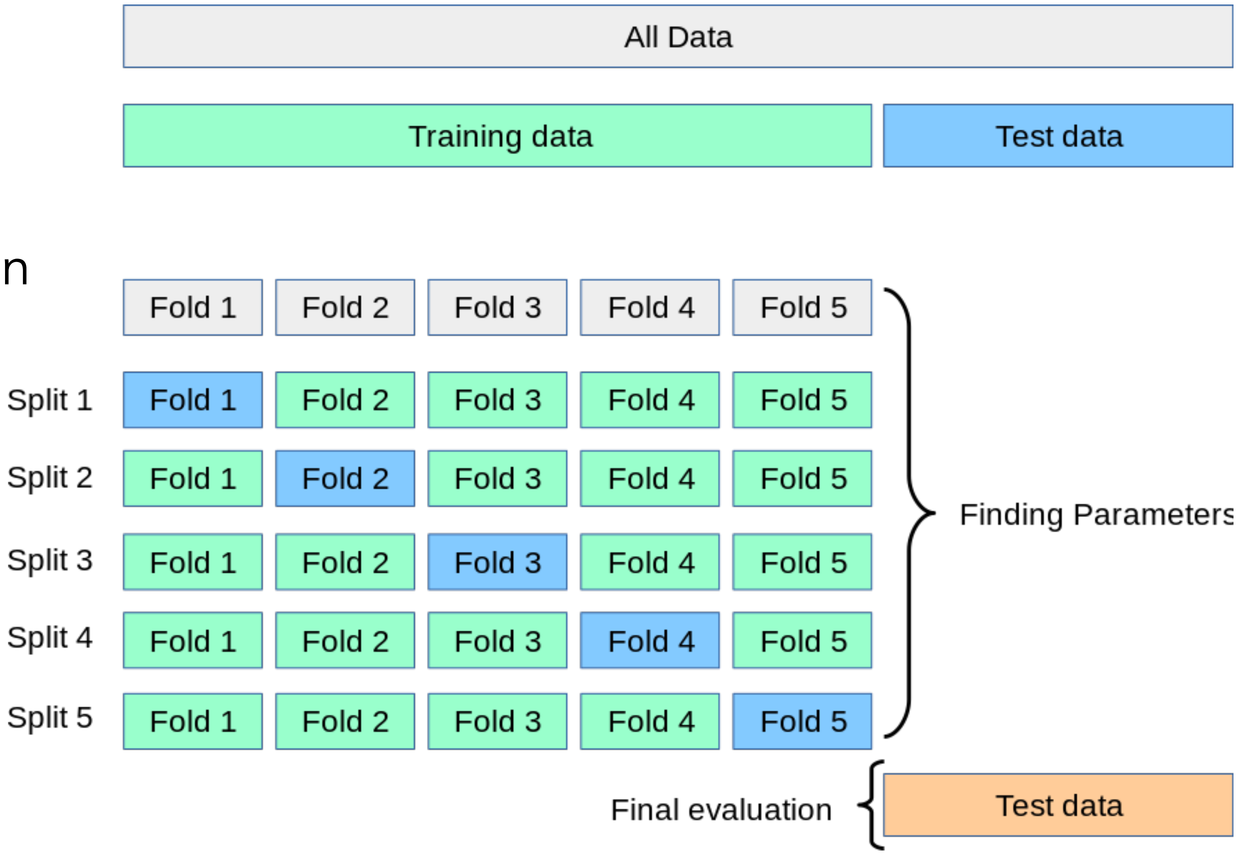
In a data-rich situation, the best approach for both problems is to randomly divide the dataset into three parts: a training set, a validation set, and a test set. The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the error of the final chosen model.



K-fold cross validation

K-fold cross- validation uses part of the available data to fit the model, and a different part to test it.

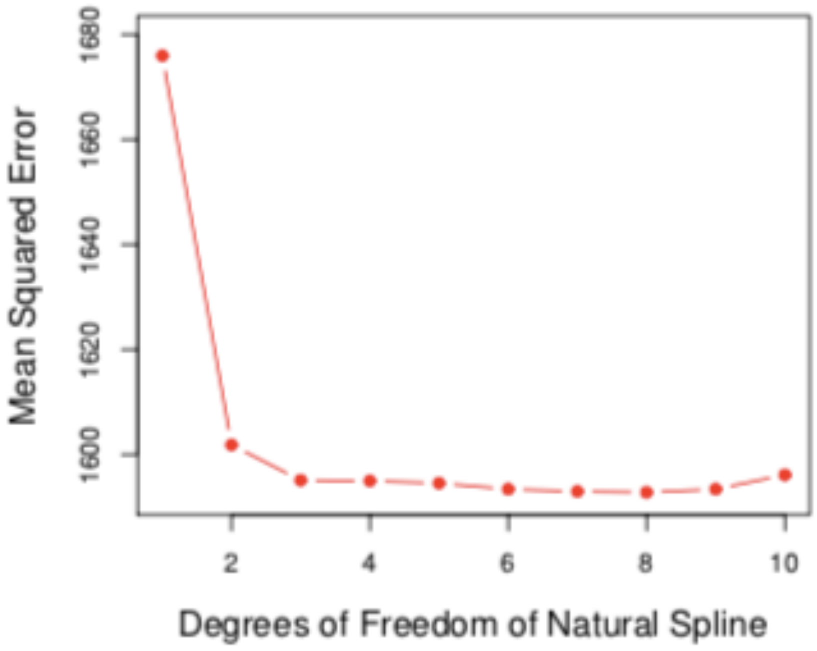
We split the data into K roughly equal-sized parts; for example, when K = 5, the scenario looks like this:

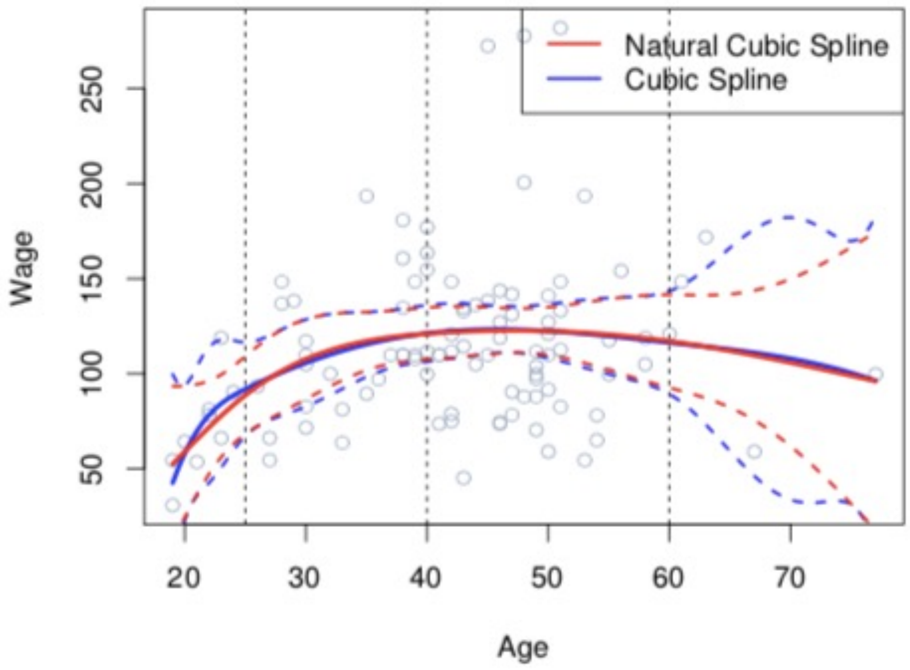


For the k -th part (k=1,…,K), we fit the model to the other K – 1 parts of the data, and calculate the prediction error of the fitted model when predicting the kth part of the data. We do this for k = 1,2,…,K and combine the K estimates of prediction error.

The case K = N is known as *leave-one-out cross-validation*.

Example: selection of the number of Knots in Spline


Mean Squared Error
Degrees of Freedom of Natural Spline


Natural Cubic Spline
Cubic Spline
Wage
Age