Advanced Statistical Modeling

Part 2. Nonparametric Modeling

Session 6:

Spline smoothing

Pedro Delicado

Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

Penalized least squares nonparametric regression

- Consider the nonparametric regression model

$$y_i = m(x_i) + \varepsilon_i,\ i = 1, \dots, n,$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. with

$$E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2 \text{ for all } i,$$

and values $x_1, \dots, x_n$ are known.
- We have studied a family of nonparametric estimators of $m(x)$, namely the local polynomial estimators, and we have seen that they have good properties.
- Now we deal with the problem of estimation $m(x)$ with another approach: We express the estimation problem as an optimization problem that leads to a new family of nonparametric regression estimators.

- Let us start with a least squares (LS) problem:

$$\min_{\tilde{m}:\mathbb{R}\longrightarrow\mathbb{R}} \sum_{i=1}^{n} (y_i - \tilde{m}(x_i))^2.$$

- Any function $\tilde{m}$ interpolating the observed data (x_i, y_i), $i = 1, \ldots, n$, (that is, verifying that $\tilde{m}(x_i) = y_i$ for all i) is an optimal solution of this LS problem.
- But, in general, an interpolating function $\tilde{m}$ is not smooth enough as a function of x.
- If we want an optimal solution being a smooth function, we must to include in the LS problem a penalization by the lack of smoothness.
- This way we obtain the penalized least squares problem:

$$\min_{\tilde{m}\in\mathcal{M}} \left\{ \sum_{i=1}^{n} (y_i - \tilde{m}(x_i))^2 + \phi(\tilde{m}) \right\},$$

where $\mathcal{M}$ is a class of smooth functions (for instance having p continuous derivatives) and $\phi(\tilde{m})$ is a functional ($\phi : \mathcal{M} \longrightarrow \mathbb{R}$) penalizing the lack of smoothness of $\tilde{m}$.

- When data x_i are in the interval $[a, b] \subseteq \mathbb{R}$ a common choice for $\mathcal{M}$ is the second order Sobolev space in $[a, b]$:

$$\mathcal{M} = W_2^2[a,b] = \left\{ m : [a,b] \longrightarrow \mathbb{R} : \int_a^b (m(x))^2 dx < \infty, \right.$$

$$\left. \text{there exists } m''(x) \text{ and } \int_a^b (m''(x))^2 dx < \infty \right\},$$

- Then the penalty function is $\phi(m) = \lambda \int_a^b (m''(x))^2 dx,\ \lambda > 0$.
- The penalized least squares problem can be stated as

$$\min_{\tilde{m} \in W_2^2[a,b]} \left\{ \sum_{i=1}^n (y_i - \tilde{m}(x_i))^2 + \lambda \int_a^b (\tilde{m}''(x))^2 dx \right\}.$$

- This problem has a unique solution: a cubic spline with knots at $x_1, \ldots, x_n$, the observed values of the explanatory variable.

Splines, cubic splines and interpolation

- **Definition.** The function $s : [a, b] \longrightarrow \mathbb{R}$ is a **spline function** of degree p with knots $t_1, \ldots, t_k$ when
  1. $a < t_1 < \cdots < t_k < b$ (we define $t_0 = a$, $t_{k+1} = b$).
  2. at each interval $[t_j, t_{j+1}]$, $j = 0, \ldots, k$, $s(x)$ is a degree p (or lower than p) polynomial.
  3. $s(x)$ has $(p-1)$ continuous derivatives in $[a, b]$ (the polynomials defining $s(x)$ in $[t_{j-1}, t_j]$ and $[t_j, t_{j+1}]$ have a smooth link at t_j).
- **Exemple:** Cubic splines. The most commonly used spline functions are those with degree 3 (cubic splines).
  - They are piecewise third degree polynomials, that are continuous and have first and second continuous derivatives at the knots.
  - It is said that the human eye is not able to detect discontinuities in the derivatives of degree 3 or higher.
  - So cubic splines fit well the concept of smooth function.

Periodic and natural splines

- A degree p spline $s(x)$ is said to be periodic when $s^{(j)}(a) = s^{(j)}(b)$ for $j = 0, \ldots, p-1$.
- Let p be an odd number, $p = 2l - 1$, with $l \geq 2$. A degree p spline $s(x)$ is said to be natural when

$$s^{(l+j)}(a) = s^{(l+j)}(b) = 0, \ j = 0, 1, \ldots, l-1.$$

  So $s(x)$ must verify $p + 1 = 2l$ restrictions.
- **Exemple.** Natural cubic splines.

  When $p = 3$, then $l = 2$ and the $2l = 4$ restrictions that a natural cubic spline must verify are:

$$s''(a) = s''(b) = 0, \ s'''(a) = s'''(b) = 0.$$

  Therefore a natural cubic spline $s(x)$ is linear in $[a, t_1]$ and $[t_k, b]$. Moreover, $s''(t_1) = s''(t_k) = 0$.

Proposition

Let $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ be the set of splines of degree p with knots $t_1, \ldots, t_k$ defined in $[a, b]$. Then $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ is a vector space with dimension $p + k + 1$.

Sketch of the proof:

- It is a vector space.
- Number of parameters: $(p + 1)(k + 1) = pk + p + k + 1$.
- Number of linear restrictions: pk.
- Dimension: $(pk + p + k + 1) - pk = p + k + 1$.

A basis for cubic splines

- The set $S[p=3; a=t_0, t_1, \ldots, t_k, t_{k+1}=b]$ of cubic splines has dimension $3+k+1=k+4$.
- A basis for this vector space is as follows:

$$s_1(x)=1, s_2(x)=x, s_3(x)=x^2, s_4(x)=x^3,$$

$$s_j(x)=(x-t_j)_+^3,\ j=1,\ldots,k,$$

where for any real number u, $u_+=\max\{0,u\}$ is the positive part of u.
- This is not the only possible basis for the set of cubic splines. In fact we will study other bases that are more suitable for numeric manipulation (B-splines bases, for instance).

Proposition

Let $N[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ be the set of natural splines of degree p with knots $t_1, \ldots, t_k$ defined in $[a, b]$. Then $N[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$ is a vector space with dimension k.

Sketch of the proof:

- It is a vector space.
- Number of parameters: $(p + 1)(k + 1) = pk + p + k + 1$.
- Number of linear restrictions: $pk + 2l = pk + p + 1$.
- Dimension: $(pk + p + k + 1) - (pk + p + 1) = k$.

Proposition

Given $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, n$, $n \geq 2$, $a < x_1 < \cdots < x_n < b$, *there is a unique natural spline* $s(x)$ *of degree* p *with knots* x_i, $i = 1, \ldots, n$, *interpolating these data:*

$$s(x_i) = y_i, \; i = 1, \ldots, n.$$

Sketch of the proof:

- Dimension of $N[p; a = x_0, x_1, \ldots, x_n, x_{n+1} = b]$: n.
- Number of linear restrictions for interpolation: n.
- Dimension of the solution: $n - n = 0$. Consistent linear system of equations (n equations, n unknowns) with a unique solution.

Practice:

Interpolating natural cubic spline using the R function `spline`.

Smoothing splines

Now we focus on cubic splines.

Proposition

Let $n \geq 2$ and let $s(x)$ be the natural cubic spline interpolating the data (x_i, y_i), $i = 1, \ldots, n$, with $a < x_1 < \cdots < x_n < b$. Let $g(x)$ be other function in $\mathcal{M} = W_2^2[a, b]$ that also interpolates the data: $g(x_i) = y_i, i = 1, \ldots, n$. Then

$$\int_a^b (s''(x))^2 dx \leq \int_a^b (g''(x))^2 dx.$$

Equality holds if and only if $g(x) = s(x)$ for all $x \in [a, b]$.

Proof

Let $h(x) = g(x) - s(x)$. Then $h(x_i) = 0$, $i = 1, \ldots, n$. Integrating by parts,

$$I = \int_a^b s''(x)h''(x)dx = \left\{ \begin{array}{c} u = s''(x) \Rightarrow du = s'''(x)dx \\ dv = h''(x)dx \Rightarrow v = h'(x) \end{array} \right\} =$$

$$(h'(x)s''(x))|_a^b - \int_a^b h'(x)s'''(x)dx = -\int_a^b h'(x)s'''(x)dx.$$

The last equality holds because $s''(a) = s''(b) = 0$, given that $s(x)$ is a natural cubic spline. The same fact implies that $s'''(x) = 0$ if $x \in [a, x_1)$ or $x \in (x_n, b]$.

Proof (cont.)

On the other hand, given that $s(x)$ is a cubic spline, $s'''(x)$ is constant between any pair of consecutive knots: $s'''(x) = s'''(x_i^+)$ if $x \in [x_i, x_{i+1})$, $i = 1, \ldots, n-1$. Then,

$$I = -\int_a^b h'(x)s'''(x)dx = -\sum_{i=1}^{n-1} s'''(x_i^+) \int_{x_i}^{x_{i+1}} h'(x)dx =$$

$$-\sum_{i=1}^{n-1} s'''(x_i^+)(h(x_{i+1}) - h(x_i)) = 0.$$

It follows that

$$\int_a^b (g''(x))^2 dx = \int_a^b ((g''(x) - s''(x)) + s''(x))^2 dx =$$

$$\underbrace{\int_a^b (h''(x))^2 dx}_{\geq 0} + \int_a^b (s''(x))^2 dx + 2 \underbrace{\int_a^b s''(x)h''(x)dx}_{=I=0} \geq \int_a^b (s''(x))^2 dx.$$

Proof (cont.)

Equality hods if and only if $\int_a^b (h''(x))^2 dx = 0$, that is equivalent to say that $h''(x) = 0$ for al $x \in [a, b]$, and equivalent to $h(x)$ being a linear function in $[a, b]$. This jointly with the fact that $h(x_i) = 0$, $i = 1, \ldots, n$ and $n \geq 2$, $h(x) = 0$ for all $x \in [a, b]$, that is $g(x) = s(x)$ for all $x \in [a, b]$.

Proposition

Let $n \geq 2$ and consider the data set (x_i, y_i), $i = 1, \ldots, n$, with $a < x_1 < \cdots < x_n < b$. Given a parameter value $\lambda > 0$, the solution of the problem

$$\min_{\tilde{m} \in W_2^2[a,b]} \Psi(m) = \left\{ \sum_{i=1}^{n} (y_i - \tilde{m}(x_i))^2 + \lambda \int_a^b (\tilde{m}''(x))^2 dx \right\},$$

is a natural cubic spline with knots $x_1, \ldots, x_n$.

Remark: This Proposition establishes that the optimization problem in a infinite dimensional space, $W_2^2[a, b]$, can be reduced to a finite dimensional space: the set of natural cubic splines.

Proof

Let $g(x) \in W_2^2[a, b]$ be a function not being a natural cubic spline with knots x_i, $i = 1, \ldots, n$.

Let $s_g(x)$ be the natural cubic spline with knots x_i, $i = 1, \ldots, n$, interpolating the points $(x_i, g(x_i))$, $i = 1, \ldots, n$.

So $s_g(x_i) = g(x_i)$, $i = 1, \ldots, n$, and therefore

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 = \sum_{i=1}^{n}(y_i - s_g(x_i))^2.$$

On the other hand we know that

$$\int_a^b (s_g''(x))^2 dx < \int_a^b (g''(x))^2 dx.$$

Therefore $\Psi(s_g) < \Psi(g)$, and it follows that the optimum of $\Psi(m)$ is obtained when m is a natural cubic spline with knots $x_1, \ldots, x_n$.

- The problem stated in the last Proposition can be written as

$$\min_{s\in N(3;a,x_1,\dots,x_n,b)} \left\{ \sum_{i=1}^{n} (y_i - s(x_i))^2 + \lambda \int_a^b (s''(x))^2 dx \right\}.$$

- Let $s(x) = \sum_{j=1}^{n} \alpha_j N_j(x) = \alpha^{\mathsf{T}} \mathbf{N}(x)$, where $\alpha = (\alpha_1, \dots, \alpha_n)^{\mathsf{T}}$ and $\mathbf{N}(x) = (N_1(x), \dots, N_n(x))^{\mathsf{T}}$, be the expression of $s(x)$ in the basis $\{N_1(x), \dots, N_n(x)\}$ of $N(3; a, x_1, \dots, x_n, b)$.

- Then, $s''(x) = \sum_{j=1}^{n} \alpha_j N_j''(x) = \alpha^{\mathsf{T}} \mathbf{N}''(x)$ and

$$\int_a^b (s''(x))^2 dx = \int_a^b s''(x) s''(x)^{\mathsf{T}} dx =$$

$$\alpha^{\mathsf{T}} \left(\int_a^b \mathbf{N}''(x) (\mathbf{N}''(x))^{\mathsf{T}} dx \right) \alpha = \alpha^{\mathsf{T}} A \, \alpha,$$

  where A is a $n \times n$ matrix with generic (i,j) element $\int_a^b N_i''(x) N_j''(x) dx$.
- Let $Y = (y_1, \ldots, y_n)^{\mathsf{T}}$ and let $\mathbf{N_x}$ be the $n \times n$ matrix with generic (i,j) element $N_j(x_i)$.
- Then

$$\sum_{i=1}^{n} (y_i - s(x_i))^2 = (Y - \mathbf{N_x}\alpha)^{\mathsf{T}} (Y - \mathbf{N_x}\alpha).$$

- Then the penalized least squares problem can be expressed as

$$\min_{\alpha\in\mathbb{R}^n} \Psi(\alpha) = (Y - \mathbf{N_x}\alpha)^{\mathsf{T}}(Y - \mathbf{N_x}\alpha) + \lambda\alpha^{\mathsf{T}}A\alpha,$$

that has an explicit solution, as we show now.

- Taking the gradient

$$\nabla\Psi(\alpha) = -2\mathbf{N_x^{\mathsf{T}}}(Y - \mathbf{N_x}\alpha) + 2\lambda A\alpha,$$

and solving in α when this gradient is equal to 0, we have that the optimal value of α is

$$\hat{\alpha} = \left(\mathbf{N_x^{\mathsf{T}}}\mathbf{N_x} + \lambda A\right)^{-1}\mathbf{N_x^{\mathsf{T}}}Y.$$

- Therefore, the vector of fitted values is

$$\hat{Y} = \mathbf{N_x}\hat{\alpha} = \mathbf{N_x}\left(\mathbf{N_x^{\mathsf{T}}}\mathbf{N_x} + \lambda A\right)^{-1}\mathbf{N_x^{\mathsf{T}}}Y = SY.$$

- We conclude that the spline estimator of $m(x)$ is a linear smoother.
- Therefore we can apply here all we know about linear smoothers: smoothing parameter choice (now the smoothing parameter is λ) by leave-one-out cross-validation or generalized cross-validation, effective number of parameters, estimation of the residual variance.

Asymptotic properties of the spline estimator of $m(x)$

Local behaviour

- Let $\hat{m}_\lambda(x)$ be the spline estimator of $m(x)$ when using λ as smoothing parameter.
- When $\lambda \longrightarrow 0$ and $n\lambda^{1/4} \longrightarrow \infty$ as $n \longrightarrow \infty$, It can be proved that

$$\text{Bias}(\hat{m}_\lambda(x)) = O(\lambda), \ \text{Var}(\hat{m}_\lambda(x)) = O\left(\frac{1}{n\lambda^{1/4}}\right),$$

$$\lambda_{\text{AMSE}} = O\left(n^{-4/9}\right), \ \text{MSE}(\lambda_{\text{AMSE}}) = O\left(n^{-8/9}\right).$$

- The local asymptotic properties are similar to those of local cubic regression.

Global behaviour

It can be established the following relationship between the spline estimator and a Nadaraya-Watson type kernel estimator with varying bandwidth:

for $x \in (a, b)$,

$$\hat{m}_\lambda(x) \approx \frac{1}{nf(x)h(x)} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h(x)}\right) y_i = \frac{\frac{1}{nh(x)} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h(x)}\right) y_i}{f(x)},$$

where

$$K(u) = \frac{1}{2} e^{|u|/\sqrt{2}} \sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right),$$

is an *order 4 kernel* (a symmetric kernel having zero moment of order 2), and

$$h(x) = \lambda^{1/4} f(x)^{-1/4}.$$

B-splines

- We introduce now a basis for $S[p; a = t_0, t_1, \ldots, t_k, t_{k+1} = b]$, the vector space of spline functions of degree p and knots $t_1, \ldots, t_k$, that is both numerically and computationally convenient.
- They are the bases of B-splines, that are recursively defined.
- In addition to the k knots $t_1, \ldots, t_k$, we introduce $2M$ auxiliary knots:

$$\tau_1 \le \cdots \le \tau_M \le t_0,\; t_{k+1} \le \tau_{k+M+1} \le \cdots \le \tau_{k+2M}.$$

- The choice of the auxiliary knots is arbitrary and they can be

$$\tau_1 = \cdots = \tau_M = t_0,\; t_{k+1} = \cdots = \tau_{k+M+1} = \tau_{k+2M}.$$

- Rename the original knots: $\tau_{M+j} = t_j$, $j = 1, \ldots, k$.

- Define the basis of B-splines of order 1 (degree 0) as:

$$B_{j,1} = I_{[\tau_j,\tau_{j+1}]},\ j = 1, \dots, k + 2M - 1.$$

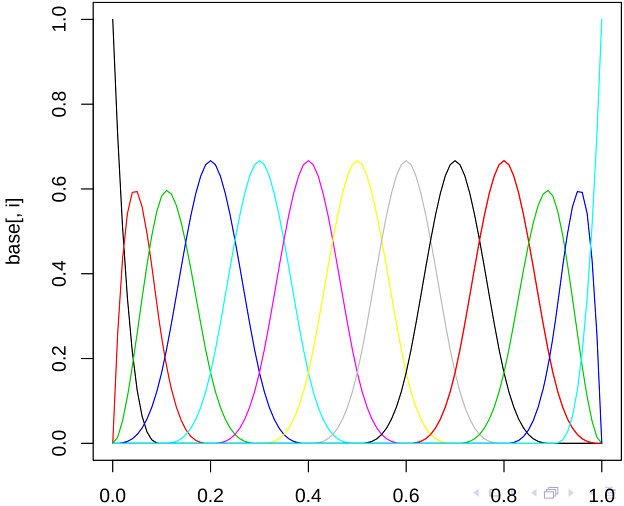
- For $m = 2, \dots, M$, the basis of B-splines of order m (degree $m - 1$) as

$$B_{j,m} = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j} B_{j,m-1} + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}} B_{j+1,m-1},$$

  for $j = 1, \dots, k + 2M - m$, where these quotients are 0 when the denominator is 0.

- For $m = M = 4$ the functions $\{B_{j,m},\ j = 1, \dots, k + 4\}$ are a basis for the vector space of cubic splines with knots $t_1, \dots, t_k$ defined in $[a, b]$, that is called basis of cubic B-splines.

Example: The 13 functions in the basis of cubic B-splines defined in $[0,1]$ having 9 knots in $0.1,\ldots,0.9$, $\tau_1=\cdots=\tau_4=0$, $\tau_{14}=\cdots=\tau_{17}=1$.



Properties of cubic B-splines bases

1. $B_{j,4}(x) \geq 0$ for all $x \in [a, b]$.
2. $B_{j,4}(x) = 0$ if $x \notin [\tau_j, \tau_{j+4}]$.
3. If $j \in \{4, \ldots, k+1\}$, $B_{j,4}^{(l)}(\tau_j) = 0$, $B_{j,4}^{(l)}(\tau_{j+4}) = 0$, for $l = 0, 1, 2$.

The second property is the responsible of the computational advantages of using bases of B-splines.

- Consider again the penalized least squares problem, optimizing now in the set of cubic splines, written as a linear combination of the basis of B-splines:

$$\min_{\beta\in\mathbb{R}^{n+4}} \Psi(\beta) = (Y - \mathbf{B_x}\beta)^{\mathsf{T}}(Y - \mathbf{B_x}\beta) + \lambda\beta^{\mathsf{T}}B\beta,$$

  where B is the $(n+4)\times(n+4)$ matrix with generic (i,j) element $\int_a^b B_i''(x)B_j''(x)dx$, and $\mathbf{B_x}$ is the $n\times(n+4)$ matrix with generic (i,j) element $B_j(x_i)$.
- Now the solution is $\hat{\beta} = \left(\mathbf{B_x^{\mathsf{T}}}\mathbf{B_x} + \lambda B\right)^{-1}\mathbf{B_x^{\mathsf{T}}}Y$.
- Observe that now the matrices $\mathbf{B_x^{\mathsf{T}}}\mathbf{B_x}$ and B are band matrices, with elements (i,j) equal to 0 if $|i-j|>4$.
- Then the required matrix inversion to compute $\hat{\beta}$ is easier (for instance, Cholesky decomposition can be used).
- Observe that the optimal cubic spline must be a natural cubic spline.

Practice:

- Basis of B-splines.
- Smoothing splines for Country Development Data.

Spline regression using B-splines

- When solving the penalized least squares problem, and then looking for the best (natural) cubic spline, in practice it is not necessary to look for those having knots at every observed x_i, $i = 1, \ldots, n$.
- In fact the computational cost of doing that is high for large values of n.
- It is enough to take a sufficiently large number k of knots, taking the knot t_j as the $(j/(k+1))$ quantile of the data $x_1, \ldots, x_n$.
- Regarding the convenient value of k, Ruppert, Wand, and Carroll (2003) suggest

$$k = \min\left\{35, \ \frac{1}{4} \times (\text{number of different } x_i\text{'s})\right\}.$$

- Function `smooth.spline` uses by default $k = O(n^{1/5})$ when $n \geq 50$.

- All we have said before for B-splines applies also when $k < n$ knots are used:

$$\min_{\beta \in \mathbb{R}^{k+4}} \Psi(\beta) = (Y - \mathbf{B_x}\beta)^{\mathsf{T}}(Y - \mathbf{B_x}\beta) + \lambda\beta^{\mathsf{T}} B\beta,$$

  where B is the $(k+4) \times (k+4)$ matrix with generic (i,j) element $\int_a^b B_i''(x)B_j''(x)dx$, and $\mathbf{B_x}$ is the $n \times (k+4)$ matrix with generic (i,j) element $B_j(x_i)$.
- Two tunning parameters, λ and k, that can play the role of *smoothing parameters*:
  - If we take $\lambda = 0$, the number k of knots acts as smoothing parameter.
  - If k is fixed and *large* ($k = O(n^{1/5})$, for instance) then λ is the only smoothing parameter. This is the common option.

Practice:

- Spline regression using B-splines for Country Development Data.
- Run `Example.spline.regression.R` first.
- Then go to `Pract_ASM_Sess_6.R`
- Compare with smoothing splines.

Generalized nonparametric regression with splines

- Assume now that the response variable Y has a distribution, conditional to $X = x$, given by
$$(Y|X = x) \sim f(y|\theta(x)),$$
where $\theta(x) \in \mathbb{R}$ is a smooth function of x free of constraints.
- Given a sample (x_i, y_i), $i = 1, \ldots, n$, following this model, the problem of maximum log-likelihood, penalized for lack of smoothness, is
$$\max_{\theta \in W_2^2[a,b]} \left\{ \sum_{i=1}^{n} \log(f(y_i|\theta(x_i))) + \lambda \int_a^b (\theta''(x))^2 dx \right\}.$$

$$\max_{\theta \in W_2^2[a,b]} \left\{ \sum_{i=1}^{n} \log(f(y_i|\theta(x_i))) + \lambda \int_a^b (\theta''(x))^2 dx \right\}.$$

- Similar arguments to those used before prove that the optimal function $\theta(x)$ is a natural cubic spline with knots $x_1, \ldots, x_n$.
- Nevertheless, now the solution has not a closed expression. Numerical optimization is required.
- A possible algorithm:
  Penalized iteratively re-weighted least squares (P-IRWLS).
- The idea is to replace the linear fit at each step of the IRWLS algorithm used to fit a GLM model, by a spline smoothing.
- See Wood (2006), page 136, for details.
- An alternative: To fit a GLM using a B-spline basis matrix as regressors matrix.

Practice:

- GLM and B-splines for Country Development Data.
- Run `Example.IRWLS.logistic.R` first.
- Then go to `Pract_ASM_Sess_6.R`

Green, P. and B. Silverman (1994).
Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.
London: Chapman & Hall.

Hastie, T., R. Tibshirani, and J. Friedman (2001).
The Elements of Statistical Learning. Data Mining, Inference, and Prediction.
Springer.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003).
Semiparametric Regression.
Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wasserman, L. (2006).
All of Nonparametric Statistics.
New York: Springer.

Wood, S. (2006).
Generalized Additive Models: An Introduction with R.
Chapman and Hall/CRC.