JP Journal of Biostatistics
Volume …, Issue …, 2010 Pages …
This paper is available online at http://pphmj.com/journals/jpb.htm
© 2010 Pushpa Publishing House

FREE-KNOT SPLINES WITH RJMCMC FOR LOGISTIC MODELS AND THRESHOLD SELECTION

M. DENIS and N. MOLINARI

Institut Universitaire de Recherche Clinique (IURC)
University of Montpellier
1, 641, avenue Gaston Giraud
34093 Montpellier, France
e-mail: marie.denis@inserm.fr

Hôpital Carremeau, CHU Nîmes
Place du Pr. R. Debré
30029 Nîmes cedex 9, France

Abstract

In medical statistics, the logistic model is a popular choice for the analysis of the dependence between a response variable and one or more explanatory variables. The response variable is the log odds and it is a linear function of explanatory variables. This type of modeling is restrictive, as the behaviour of the log odds can be best represented by a smooth non-linear function. Thus, we use a representation *B*-spline, where the number and location of knots are seen as free variables, is used to improve the fitting. For a piecewise linear spline, knots are points where the slope is changing in the shape of the function. Therefore, a quick change of slope allows to interpret the knot location as a threshold value. The use of MCMC simulation techniques is a very important computational tool in Bayesian statistics. These methods belong to a class of algorithms for sampling from target distributions on a space of fixed dimension. The Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm, allows simulations from target distributions on spaces of

2010 Mathematics Subject Classification: Kindly Provide.
Keywords and phrases: Kindly Provide.

Received September 10, 2010

varying dimension. One of the main purposes of the present investigation is to use this RJMCMC method for modeling the log odds by a *B*-spline representation with an unknown number of knots at unknown locations. The method is illustrated with simulations and a real data set from an in vitro fertilization program.

1. Introduction

Logistic regression is a powerful and flexible means to analyze the relationship between a dependent dichotomous variable (e.g. which only takes two possible values) and one or more risk factors (e.g. explanotary variables). It is a method very used in applied research, but it assumes that these explanotary variables have a linear effect on the model. This assumption is restrictive; in fact in most of problems the underlying processes are complex and not well understood. Using spline functions seems to be an interesting alternative to study this relationship. It permits to detect the possibility of non-linear effects of the explanotary variables. The name spline function was introduced by Schöenberg (1) in 1946. The real explosion in the theory, and in practical applications, began in the early 1960s. Spline functions are used in many applications such as interpolation, data fitting, solving numerically ordinary and partial differential equations (finite element method), and in curve and surface fitting. For survival data analysis, Sleeper and Harrington (2) introduced spline function into the Cox model. Kooperberg et al. (3) developed the hazard regression (HARE) method which uses piecewise linear regression splines to model the hazard function. The diversity of applications exists due to the great flexibility of splines. But, the main difficulty of splines is the selection of the number and location of knots. In this paper, we utilize the Reversible Jump Markov Chain Monte Carlo (RJMCMC) technique introduced by Green (4) to handle this difficulty.

In recent years the use of MCMC simulation techniques has been a very important computational tool in Bayesian statistics. These methods belong to a class of algorithms for sampling from target distributions on a space of fixed dimension. The RJMCMC algorithm allows simulations from target distribution on spaces of varying dimension. One application is the comparaison of models: the “true” model is unknown but is assumed to come from a specified class of parametric model $\{\mathcal{M}_0, \mathcal{M}_1, \cdots\}$.

One of the main purposes of the present investigation is to use this RJMCMC method for modeling the logit function by a *B*-spline representation with an unknown number of knots at unknown locations. Considering the spline knots as free parameters implies more flexibility and improves data approximation. Moreover, the use of spline allows both defining threshold values and removing the linearity assumption of the logit function. If the estimation of the logit function is based on piecewise linear splines, the knot location corresponds to a break point in the linearity, so a quick change of slope can be interpreted as a point separating the variable range in two parts and the knot location corresponds to a threshold value. Finally, the RJMCMC algotrithm gives directly the knot number without using a model selection criterion and it allows to estimate a wide range of features for the function of the interest. This approach has been introduced by (9) and developped by severals authors ((1)).

The paper is organized as follows. In section 2, a short review of the spline functions and the logistic model is given. In section 3, we shall introduce the Reversible Jump MCMC algorithm, and we give two applications in section 4 with simulations and a real data set from an in vitro fertilization program.

2. The Model

2.1. Spline functions

Let $(r_0 =)a < r_1 < r_2 < \cdots < r_k < b(= r_{k+1})$ be a subdivision of k distinct points on the interval $[a, b]$ on which the x variable is valued. We denote the points $(r_1, ..., r_k)$ as the k interior knots, r_0 and r_{k+1} as the boundary knots. The spline function $s(x)$ used to transform the x variable is a polynomial of degree d (or order $d + 1$) on any intervals $[r_{i-1}, r_i]$, and has $d - 1$ continuous derivatives on the open interval $[a, b]$. These functions provide great flexibility for fitting data, which is controlled by the number of knots. Spline functions belong to a linear functionnal space of dimension $d + 1 + k$. The most popular basis function for this linear space is given by Schoenberg's B-splines, or Basicsplines, and is denoted by $\{B_1^d(\cdot, r), ..., B_{d+1+k}^d(\cdot, r)\}$ for a fixed sequence of knots $r = (r_1, ..., r_k)'$. Their structure is advantageous as it requires less computation as compared to other basis functions such as the truncated power basis (Eubank (15); Ramsay and Silverman (16)). De Boor (5) proposes a recursive algorithm to compute B-splines of any degree from B-splines of a lower degree.

We can define B-spline basis functions by:

$$B_j^1(x,\ r) = \begin{cases} 1 & \text{si } r_j \le x \le r_{j+1} \\ 0 & \text{sinon} \end{cases}$$

$$B_j^s(x,\ r) = \frac{r - r_j}{r_{j+s-1} - r_j} B_j^{s-1}(x,\ r) + \frac{r_{j+s} - r}{r_{j+s} - r_{j+1}} B_{j+1}^{s-1}(x,\ r),$$

where $j = 1, \ldots, k + d + 1$ and $s = 2, \ldots, d + 1$. Thus, each basis function is non zero in a limited interval spanned by a $d + 1$ adjacent knots which leads to stable estimates and reduces computation. These are piecewise polynomials with continuity constraints on the polynomial and its first $d - 1$ derivaties at the interior knots.

So, a spline function can be written

$$s(x,\ \beta,\ r) = \sum_{i=1}^{d+k+1} \beta_i B_i^d(x,\ r), \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_{d+1+k})'$ is the vector of the spline coefficients and $r = (r_1, \ldots, r_k)'$ is the vector of the interior knots. We can extend to the multivariate case by using additive models. With additive modeling (6) one can to decompose a function of the form $h(X) = h(X_1, \ldots, X_p)$ by a sum of functions of the individual components of X, where $Y = h(X)$ is the response variable and $X = (X_1, \ldots, X_p)$ the explanotary variables.

Let $(y^i,\ x^i)$ the observations, where each x^i is a p-vector $(x_1^i, \ldots, x_p^i)$. So, h is defined by:

$$Y = h(X) = h(X_1, \ldots, X_p) = \sum_{i=1}^{p} h_j(X_j), \tag{2}$$

and, an estimator s of f can be given by:

$$s(x) = s(x_1, \ldots, x_p) = \sum_{j=1}^{p} s_j(x_j,\ \beta^j,\ r^j), \tag{3}$$

where $\beta^j = (\beta_1^j, \ldots, \beta_{k_j+d_j+1}^j)'$, $r^j = (r_1^j, \ldots, r_{k_j+d_j+1}^j)'$ for $j = 1, \ldots, p$ and each function s_j is defined according to the equation (1). If X_i and X_j are two variables, and the response variable depends on the combination of levels of X_i and X_j, then X_i and X_j are said to interact. We incorporate a term to model this interaction; thus the model can be represented by an additive model including multiplicative interaction of order 1, as follows

$$s(x) = s(x_1, \ldots, x_p) = \sum_{j=1}^{p} s_j(x_j, \beta^j, r^j) + \sum_{i<j} s_{ij}(x_i \times x_j, \beta^{ij}, r^{ij}). \tag{4}$$

2.2. Spline logistic regression model with free-knots

The logistic model is used to study the relationship between a dichotomous variable (or response variable) and one or more explanotary variables. This model estimates the probability of a certain event occurring. The specific form of the logistic regression is

$$f(x) = \frac{\exp(\alpha_0 + \alpha' x)}{1 + \exp(\alpha_0 + \alpha' x)'}, \tag{5}$$

where $f(x)$ is the expected value of a randomly obtained proportion of the subpopulation corresponding to the vector $x = (x_1, x_2, \ldots, x_p)'$, where α^0 and $\alpha = (\alpha^1, \ldots, \alpha^p)'$ are the regression coefficients which have to be estimated from the data. We can define the logit function g as follows

$$g(x) = \ln \frac{f(x)}{1 - f(x)} = \alpha_0 + \alpha' x. \tag{6}$$

This equation shows a linear relation between the logit function and the explanotary variables. This type of modeling is too restrictive, in fact the behavior of the logit function can be non-linear. The use of splines in this regressive model allows the investigation of non-linear effects with continuous covariates and introduces a nonparametric character. The tuning parameters for regression splines are the number k and the location of knots. In this work, we model the logit function (6) with B-splines with free-knots in order to allow maximum flexibility and improve the fit.

This approach has been used by Denison (9), Lindstrom (14). More precisely, a

Markov chain Monte Carlo algorithm is used to estimate a Bayesian version of the B-spline model. Unlike, Dimatteo (13) and Johnson (17) which use a prior on the coefficients β, we estimate these parameters with least-squares estimator. Thus, we avoid the "delicate re-balancing of the coefficients" like mentioned by Dimatteo.

Other approach exists concerning the spline approximation notably P-splines (Eilers and Marx, 1996; Brezger and Lang, 2006). These methods use a relatively large number of knots and to prevent overfitting, a penalty on the second derivative restricts the flexibilty of the fitted curve. In our work, we use the knot location to interpret the results, and in this context the P-splines are not adapted.

With respect to the spline logistic regression model, it is defined by

$$f(x) = \frac{\exp(s(x, \beta, r))}{1 + \exp(s(x, \beta, r))}. \tag{7}$$

Thus, the logit function (6) can be written as a spline function

$$g(x) = \ln \frac{f(x)}{1 - f(x)} = s(x, \beta, r). \tag{8}$$

Let (y^i, x^i), where $i = 1, ..., n$, the n observed independent pairs. We approxime the logit of the conditional probability of success by a B-spline model.

Using (8), we obtain:

$$\begin{aligned}
\operatorname{logit}(P(Y = 1 | X_1, ..., X_p)) &= \ln \frac{P(Y = 1 | X_1, ..., X_p)}{1 - P(Y = 1 | X_1, ..., X_p)} \\
&= \sum_{j=1}^{p} s_j(x_j, \beta^j, r^j) \\
&= \sum_{l=1}^{k_1+d_1+1} \beta_l^1 B_l^1(x_1, r^1) + \cdots + \sum_{l=1}^{k_p+d_p+1} \beta_l^p B_l^p(x_p, r^p),
\end{aligned} \tag{9}$$

where for $j = 1, ..., p$, $(B_l^j(\cdot, r^j))_{l=1,...,k_j+d_j+1}$ is the B-spline matrix, r^j is the knots vector, β_l^j are the spline coefficients, k_j is the fixed number of knots and d_j the fixed degree of the spline function. The associated likelihood is defined by

$$\mathcal{L}(\beta, r) = \prod_{i=1}^{n}\left(\frac{\exp\left(\sum_{j=1}^{p}\sum_{l=1}^{k_j+d_j+1}\beta_l^j B_l^j(x_j^i, r_j)\right)}{1+\exp\left(\sum_{j=1}^{p}\sum_{l=1}^{k_j+d_j+1}\beta_l^j B_l^j(x_j^i, r_j)\right)}\right)^{y^i}$$

$$\left(\frac{1}{1+\exp\left(\sum_{j=1}^{p}\sum_{l=1}^{k_j+d_j+1}\beta_l^j B_l^j(x_j^i, r_j)\right)}\right)^{1-y^i}. \qquad (10)$$

The estimate of the logit function with a linear spline $d = 1$ implies easy interpretation. So, we let $d_1 = \cdots = d_p = 1.$ In fact, knots are points where the slope is changing in the shape of the piecewise linear function. So, a quick change of slope can be interpreted as a point separating the variable range into two parts and the knot location corresponds to a threshold value.

From a clinical point of view, knot location represents the threshold value of the risk factor for which the probability of a disease occurring suddenly changes. Moreover, we can define the notion of odds ratio on each interval. In practice, only a small number of threshold values are of clinical interest. A good working model provides one or two threshold values which allow the classification of the patients into two or three groups for differentiation of treatment.

3. Bayesian Estimation of The Logit Function

This section presents essential background on Reversible Jump MCMC, proposed by Green (4). The adaptation of this algorithm for the spline regression is given.

3.1. RJMCMC

The reversible jump MCMC algorithm allows simulation from target distributions on spaces of varying dimension, it can be considered as a general framework for Metropolis-Hastings algorithms ((7), (8)).

Consider the following hierarchical model: let *k* be a indicator from a coutable set $\mathcal{K}$ and $\theta^{(k)}$ be the parameter vector. Each k determines a model $\mathcal{M}_k$ defined by $\theta^{(k)}$, with dimension of the parameter space $\Theta^{(k)}$ allowed to vary with *k*.

The joint distribution of $(k,\ \theta^{(k)}y)$, where y is the data vector, is modeled as:

$$p(k,\ \theta^{(k)}y) = p(k)\,p(\theta^{(k)}\,|\,k)\,p(y\,|\,k,\ \theta^{(k)}),$$

i.e. the product of model probability, parameter prior and likelihood. Inference about k and $\theta^{(k)}$ will be based on the joint posterior $p(k,\ \theta^{(k)}\,|\,y)$, which is known as the target distribution. For convenience, we abbreviate $(k,\ \theta^{(k)})$ as z and we note $\pi(dz)$ this target distribution. Given k, z lies in $C_k = \{k\} \times \Theta^{(k)}$, while generally $z \in C = \bigcup_{k \in \mathcal{K}} C_k$.

In Markov Chain Monte Carlo computation, an aperiodic and irreductible Markov transition kernel $P(z,\ dz')$ is constructed and it satisfies detailed balance:

$$\int_A \int_B \pi(dz)P(z,\ dz') = \int_B \int_A \pi(dz')P(z',\ dz), \tag{11}$$

where $A,\ B \in C$. We simulate this chain to obtain a dependent, approximate, sample from $\pi(dz)$.

In our case, we have multiple parameter subspaces $\{C_k\}$ of different dimension.

A method that switches between these subspaces is needed. For all that, different types of move between the subspaces can be defined. If the current state is z, a move of type m to state dz' with probability $q_m(z,\ dz')$ is defined and is accepted with probability

$$\alpha_m(z,\ z') = \min\left\{1,\ \frac{\pi(dz')q_m(z',\ dz)}{\pi(dz)q_m(z,\ dz')}\right\}. \tag{12}$$

For move z to z', we must generate ramdom numbers u and set z' as a determinist function of z and $u : z' = z'(z,\ u)$. The reverse move from z' to z has to be defined symmetrically by generating random numbers u' and setting $z = z(z',\ u')$. The vectors of Markov chain states and proposal random variables $(z,\ u)$ and $(z',\ u')$ must be of equal dimension, that is, the crucial dimension matching condition:

$$n_1 + n_1' = n_2 + n_2',$$

where n_1, n_2 are the dimensions of z, z' respectively, and n'_1, n'_2 are the dimensions of u, u' respectively.

The ratio (12) becomes

$$\alpha_m(z, z') = \min\{1, \text{likelihood ratio} \times \text{prior} \times \text{proposal ratio}\}$$

$$= \min\left\{1, \frac{p(y \mid z')}{p(y \mid z)} \frac{p(z')}{p(z)} \frac{p(k \mid k')q_2(u')}{p(k' \mid k)q_1(u)} \left| \frac{\partial(z', u')}{\partial(z, u)} \right| \right\}.$$

where q_1, q_2 are the distributions of u, u' and $\left| \frac{\partial(z', u')}{\partial(z, u)} \right|$ is the jacobian. Often in practice $n_1 + m_1 = n_2$. Consequently only for the birth step a random u is necessary and we omit in the $\alpha_m(z, z')$ the terms $q_2(u')$ and u'.

3.2. RJMCMC for *B*-spline logistic regression

We have defined in 2.2 the spline logistic regression model with free-knots. We use the RJMCMC algorithm of previous section to select the number and the position of knots to have the best adjustment. For a Bayesian approach, let us formulate the hierarchical model: we take the number of interior knots k as random, from some countable set $\mathcal{K}$. $\mathcal{M}_k$ denotes the model with exactly k interior knots and $r^{(k)} = (r_1, ..., r_k)$ denotes the interior knot locations, with $r_0 = X_{\min}$ and $r_{k+1} = X_{\max}$ as the boundary knots.

As concerns the vector of spline coefficients $\beta = (\beta_i)_{1 \le i \le k+d+1}$ is to be estimed from the data by means of the standard least squares regression theory. A complete Bayesian approach would include these coefficients in the vector of parameters (see Dimatteo (13), Johnson (17)). However, Denison and al. (9) seem shown that the least squares estimation approach leads to no significant deterioration in the performance of the algorithm and avoids an additional computational burden.

We shall generate samples from the joint posterior of $(k, r^{(k)})$. By taking into account the varying dimensionality, we have to develop appropriate reversible jump moves.

For this problem, possible transitions (9) are

1. the addition of a knot (a birth step),

2. the deletion of a knot (a death step),

3. the movement of a knot.

These independent move types are randomly chosen with probability b_k for move *k* to $k+1$ (i.e. birth step), d_k for move *k* to $k-1$ (i.e. death step) and η_k for the move step. These probabilities satisfy $b_k + d_k + \eta_k = 1$ for all *k*.

3.2.1. Prior specifications

Let $k \in \mathcal{K} = \{0, ..., k_{\max}\}$. We use a truncated Poisson distribution, with parameter λ restricted to the countable set $\mathcal{K}$, to specify the prior for *k*:

$$p(k) \propto \frac{\lambda^k \exp(-\lambda)}{k!} 1_{\{0, ..., k \max\}}(k)$$

The r_i are taken to be the order statistics from a uniform random variable with state space the candidate knot locations $\mathcal{R} = \{r_{01}, ..., r_{0K}\}$, where $r_{01}, ..., r_{0K}$ are disributed equidistantly over the interval $]X_{\min}, X_{\max}[$, i.e. the knots are equally spaced. Then the prior distribution for $r^{(k)} = (r_1, ..., r_k)$ is

$$r(r^{(k)} | k) = \frac{k!}{K^k},$$

where *K* is the number of possible emplacements. As concerns the parameter λ, it could be altered depending on the prior beliefs the researchers may have about the smoothess of the logit function. Small values of λ reflects a stong insistence on smoothness.

3.2.2. Move step

The move step consists in choosing a knot uniformly, say r_j, among the set of moveable knots and proposing this knot to be moved to another position $r_j' \in \mathcal{R}$. A knot $r_j \in \{r_1, ..., r_k\}$ is called moveable ((1)), if the number m_j of vacant candidate knots $r_{0i} \in \mathcal{R}$ with $r_{j-1} < r_{0i} < r_{j+1}$ is at least 1. Let $r = r^{(k)}$. The number $n(r)$ of moveable knots then is defined as

$$n(r) = card\{r_j | m_j \geq 1,\ j \in \{1, ..., k\}\}.$$

So, firstly, we draw a knot r_j uniformly among $n(r)$ moveable knots with probability $p(r_j) = \frac{1}{n(r)}$ and, given r_j, we draw uniformly r'_j (the new position) among the set of m_j vacant candidate knots, with probability $p(r'_j \,|\, r_j) = \frac{1}{m_j}$.

The corresponding proposal ratio is given by

$$\text{proposal ratio} = \frac{p(r_j \,|\, r'_j)\, p(r'_j)}{p(r'_j \,|\, r_j)\, p(r_j)} = \frac{n(r) m_j}{n(r') m'_j} = \frac{n(r)}{n(r')}.$$

The prior ratio is 1 because all collections of the same number of knots have the same prior probability. The acceptance probability for such a move step is

$$\alpha = \min\left\{1, \frac{p(y \,|\, (k, r'))n(r)}{p(y \,|\, (k, r))n(r')}\right\},$$

where $p(y \,|\, (k, r'))$ is the spline likelihood.

3.2.3. Changing dimension

Let $z = (k, r^{(k)})$ the current state of parameters. We define $b_0 = d_{k\max} = 1$, $b_{k\max} = d_0 = 0$ and otherwise $b_k = d_k = 1/3$.

In the birth step, given k, we add a new knot $r'_j \in (r_j, r_{j+1})$. r'_j is drawn uniformly with probability $p(r'_j) = 1/(K - k)$ from the set of the $(K - k)$ vacant candidate knots $r_{0i} \in \mathcal{R}$. We have $z' = (k + 1, r')$, where $r' = (r_1, ..., r_j, r'_j, r_{j+1}, ..., r_k)$. For the birth step, the prior ratio is given by:

$$\text{prior ratio} = \frac{\text{prior for } k + 1 \text{ knots}}{\text{prior for } k \text{ knots}} \frac{\text{prior for location of } k + 1 \text{ knots}}{\text{prior for location of } k \text{ knots}}$$

$$= \frac{p(k + 1)}{p(k)} \frac{p(r' \,|\, k + 1)}{p(r \,|\, k)}$$

$$= \frac{p(k + 1)}{p(k)} \frac{k + 1}{K}.$$

The corresponding proposal ratio is given by

$$\text{proposal ratio} = \frac{d_{n+1}(1/k+1)}{b_k(1/K-k)}$$

$$= \frac{d_{k+1}(K-k)}{b_k(k+1)}.$$

In the death step, the proposed knot to delete is simply chosen uniformly from the knots of the current model, so it is drawn with probability $p(r_{j+1}) = 1/(k+1)$.

The acceptance probability for the birth step is

$$\alpha(z, z') = \min\left\{1, \frac{p(y\,|\,z')}{p(y\,|\,z)} \times \text{prior ratio} \times \text{proposal ratio}\right\}.$$

For the death step, it is the same except that the fraction is inverted. The coefficients β are estimated at each step through the function *glm.fit* available in the *R* package.

In the multivariate case, we let $k = \sum_{i=1}^{p} k_i,$ where k_i is the spline degree s_i and $r^{(k)} = (r^{(k_1)}, \ldots, r^{(k_p)})$ the parameter vector for each spline. The same movements are used that in previous algorithm: addition, deletion or movement of a knot. At each iteration, we choose randomly the spline which we are going to modify. The prior for *k* is a truncated Poisson distribution with parameter λ. The choice of this parameter reflects, in this context, the parsimony of the model.

4. Data Analysis

In this section, we illustrate the reversible jump algorithm with two examples: a simulation study and an analysis of a real data set from an in vitro fertilization program.

4.1. Simulation settings

We have simulated 2000 data according to a logistic model defined by:

$$g(x) = \ln\frac{f(x)}{1-f(x)} = \cos(x/8),\ x \in [15, 65],$$

where *x* is genereted from a uniform distribution on [15, 65]. We use the RJMCMC algorithm with splines of degree $d = 1$ and $d = 2$, to estimate this function. Let

$\lambda = 1$ be the parameter of Poisson distribution and $k_{\max} = 5$, in fact a large number of knots is unlikely to be required. In Figure 1, we display the true function along with corresponding spline model estimates with degree $d = 1, 2.$ The MSE is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\{\hat{f}(x_i) - f(x_i)\}^2,$$

where f is the true function and $\hat{f}$ is our estimate to the true function. The MSE are : 0.03 for the spline of degree 2 and 0.04 for the linear spline. Thus, we find a slightly lower MSE for the estimate when using a spline of degre 2 instead of linear spline.

4.2 Analysis of FIV data

Many couples resort to in vitro fertilization (IVF), when they have difficulties conceiving children. The principal advantage of IVF is to control follicular growth, ovulation, sperm quality and the early development of fertilized eggs. The study carried out by Roseboom et al. performed a multiple logistic regression analysis in order to evaluate the relationship between various factors and pregnancy. The study led by Demouzon et al. (2) leading even results: the probablity of pregnancy for each cycle is a_ected by the age of the patients. We want to validate this result by the new method proposed in the previous sections.

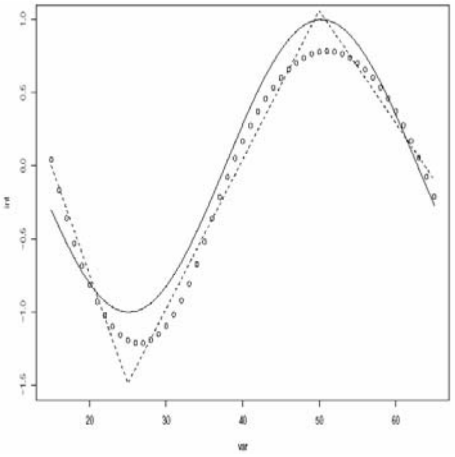

Figure 1. The true curve: -, the estimated logit function by a spline model of degree $d = 1$: -- and by a spline model of degree $d = 2$: ...

The french national register of in vitro fertilization (Fivnat) records all of the IVFs carried out in France. This population-based study is a cohort of 23, 520 couples which underwent IVF for the first time between 1994 and 1996. Couples were followed up until they obtained a possible clinical pregnancy or until the 31st of December 1998. A total of 7892 pregnancies were recorded. For each couple, the age of the woman and the age of the man at the first attempt are available. Generally, we take the degree of the spline $d_i > 0$ and the knot number $k_i \geq 0.$ However, in epidemiology a smaller number of groups is preferred, so $k_i \in \{0, ..., 5\},$ and to allow the interpretation of the results, more precisely to separate the patients in different groups, we let $d_i = 1.$

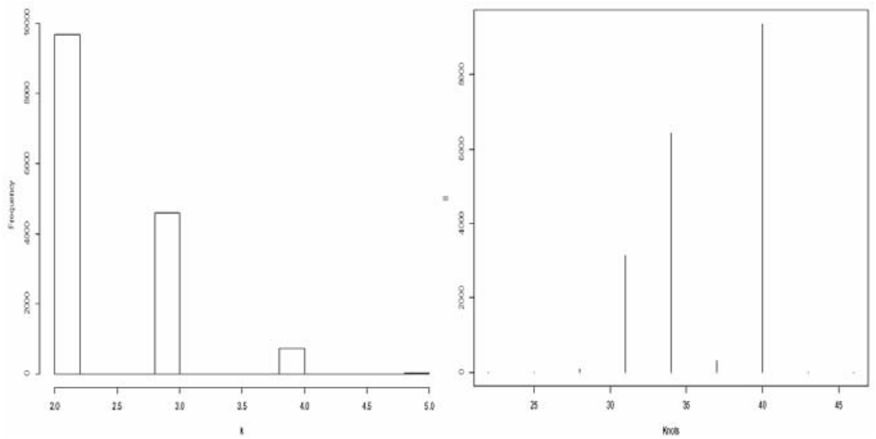

Figure 2. The posterior distribution of *k* (left) and of *r* given *k* (right)

First, we consider the univariate spline model for the age of women. The RJMCMC is used to select the number and location of knots. We let $\lambda = 1$ for the parameter of the prior distribution of *k* and $k_{\max} = 5.$ We choose these values because we wish have a smooth function (i.e. with few knots) and few groups of patients.

As concerns the parameter λ, we have tested others values; the results are the same thus the method seems robust. For the candidate knot location $\mathcal{R}$ the knots are equally spaced of 3 years. The different reversible jump moves have seen in 3.2, the vector β is approximated at each iteration. The estimates are obtained with 20000 iterations and a burn-in time of 5000.

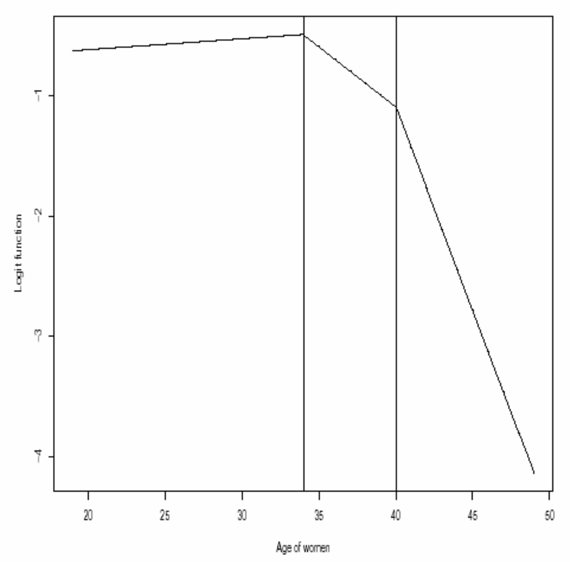

Figure 3. The logit function (i.e. the IVF success rate) according to age of women approximed by a spline model with $k = 2$ and $d = 1$.

The posterior distribution for k is shown at left in Figure 2, it indicates a mode at $k = 2$. From the right part of this figure, we see the posterior distribution of r given $k = 2$. The knot locations selected are 34 and 40. The figure 3 shows the corresponding logit function estimated by a spline of degree one and with two knots located at age of 34 and 40 years. We have fixed the spline degree at $d = 1$ to be able to interpret the results. From figure 3, the knot locations correspond to break points of the logit function. Indeed, before the first knot, the function seems constant, between the two knots it decreases, and after the second knot, it decreases sharply. Thus, the ages of 34, 40 can be interpreted as threshold values for IVF success. These results are consistent with the results found in previous studies ((11), (2)) using the classical criterion BIC.

Secondly, we model the bivariate spline model for the age of women and men.

Let $k_{\max} = 10$, $\lambda = 1$ and $d_1 = 1$, $d_2 = 1$. For each variable, we define a candidate knot site where the knots are equally spaced.

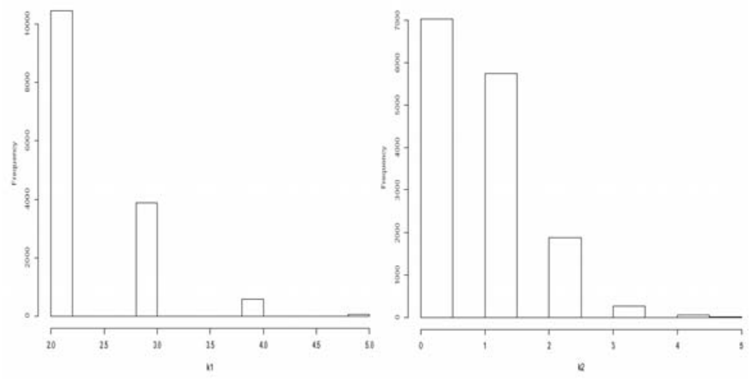

Figure 4. The posterior distribution of k_1 and k_2.

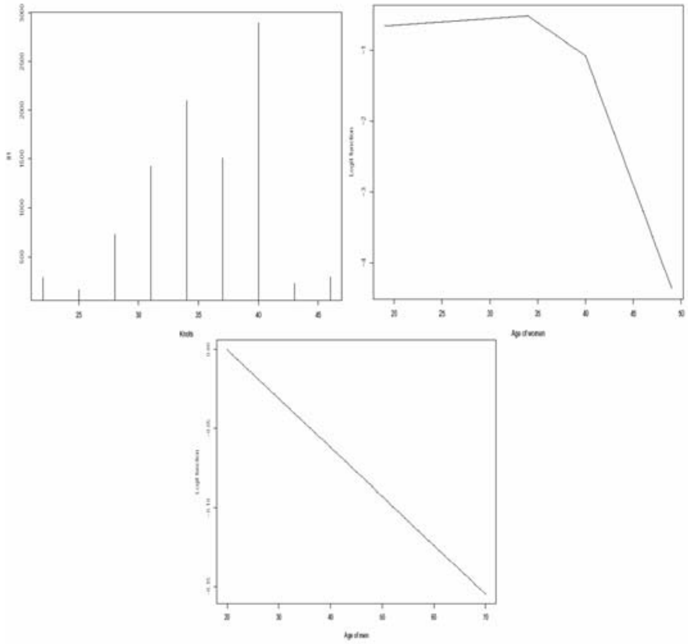

Figure 5. The posterior distribution of r_1 given $k_1 = 2$ (left), the logit function (i.e. the IVF success rate) according to age of women approximated by a spline model with $k_1 = 2$ and $d_1 = 1$ (right) and the logit function according to age of men approximated by a spline model with $k_2 = 0$ and $d_2 = 1$ (down).

Figure 4 shows the posterior distribution of k_1 and k_2. For the age of women, the posterior distribution indicates a mode at 2. Concerning the age of men, we retain any interior knot, the figure 5 shows a linear effect of age of men in IVF success. The left part of the figure 5 illustrates the posterior distribution of r_1 given $k_1 = 2$ (i.e. for the age of women); it indicates two knot locations at 34 and 40 years. These knots are full meaningful and according to the right part of the figure 5: we can assume the ages of 34 and 40 as threshold values for the IVF success. These results are consistent with the previous study using the univariate spline model. These results show the important role played by the age of women in IVF success.

5. Discussion and Future Plans

In summary, the use of *B*-spline to model the logit function helps explain the relationship between response and explanotary variables without imposing a linear link between these variables. In fact, *B*-spline modeling is more flexible. Furthermore, the linear spline model reconsiders the knots as threshold values. Thus we can classify the patients into groups for treatment differentiation. Finally, the advantage of the RJMCMC algorithm is demonstrated by the direct identification of the number of knot without resorting to model selection criterion such as the BIC or AIC.

References

[1] Schoenberg., Contributions to the problem of approximation of equidistant data by analytic functions. Quart. Appl. Math. 4 (1946), 45-99; 112-141.

[2] L. A. Sleeper and D. P. Harrington, Rgression splines in the Cox Model with application to covariate effects in liver disease. Journal of the American Statistical Association 85 (1990), 941-949.

[3] C. Kooperberg, C. J. Stone and Y. K. Truong, Hazard regression. Journal of the American Statistical Association 90 (1995), 78-94.

[4] P. J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82 (1995), 711-732.

[5] C. De Boor, A Practical Guide to Splines. Springer-Verlag: New-York, 1978.

[6] T. Hastie and R. Tibshirani, Generalized Additive Models. Chapman and Hall: London, 1990.

[7] W. K. Hasting, Monte Carlo sampling methods using Markov chains and their

applications. Biometrika 57 (1970), 97-109.

[8] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, Equations of state calculations by fast computing machines. J. Chem. Phys 21 (1953), 1087-1091.

[9] D. G. T. Denison, B. K. Mallick and A. F. M. Smith, Automatic Bayesian curve fitting. J. R. Statist. Soc. B 60 (1998), 333-350.

[10] C. Biller, Adaptive Bayesian regression splines in semiparametric generalized linear models. Sonderforschungsbereich 51 (1998), 4178-4192.

[11] N. Molinari, J. P. Daurés and J. F. Durand, Regression splines for threshold selection in survival data analysis. Statistics in Medicine 20 (2001), 237-247.

[12] J. Demouzon, B. Rossin-Amar, A. Bachelot, C. Renon and A. Devecchi, Influence du rang de la tentative en FIV, Contraception, Fertilité, et Sexologie 26 (1998), 466-472.

[13] I. DiMatteo, C. R. Genovese and R. E. Kass, Bayesian curve-fitting with free-knot splines, Biometrika 88 (2001), 1055-1071.

[14] M. J. Lindstrom, Penalized estimation of free-knot splines, J. Comp. Graph. Statist 8 (1999), 333-52.

[15] R. Eubank, Spline Smoothing and Nonparametric Regression, Dekker, New York., 1988.

[16] J. Ramsay and B. Silverman, Functional Data Analysis, Springer, New York., 1997.

[17] M. S. Johnson, Modeling dichotomous item responses with free-knot splines, Computational statistics and Data Analysis 51 (2007), 4178-4192.

[18] C. S. Li and D. Hunt, Regression splines for threshold selection with application to a random-effect logistic dose-response model, Computational statistics and Data Analysis 46 (2004), 1-9.

[19] S. Zhou and X. Shen, Spatially adaptive regression splines and accurate knot selection schemes, Journal of the American Statistical Association 96 (2001), 247-259.

Paper No. PPH-1009038-JB

Kindly return the proof after correction to:

The Publication Manager
Pushpa Publishing House
Vijaya Niwas
198, *Mumfordganj*
Allahabad-211002 (*India*)

along with the print charges*
by the fastest mail

***Invoice attached**

Proof read by: ………………………………

Copyright transferred to the Pushpa Publishing House

Signature: ………………………………

Date: ………………………………………

Tel: …………………………………………

Fax: …………………………………………

e-mail: ……………………………………

Number of additional reprints required

……………………………………………

Cost of a set of 25 copies of additional reprints @ Euro 12.00 per page.

(25 copies of reprints are provided to the corresponding author ex-gratis)