STUDIA UNIV. "BABEŞ–BOLYAI", MATHEMATICA, Volume **LI**, Number 4, December 2006

SOME INFERENCES AND EXPERIMENTS ON FREE KNOTS SPLINE REGRESSION

PETRU P. BLAGA

Dedicated to Professor Gheorghe Coman at his 70^{th} *anniversary*

Abstract. Inferences and experiments on the simple spline regression with free knots are considered. For the first time an iterative procedure given in [2] to estimate the values of the free knots based on a multiple linear regression is recalled. Point estimators and confidence interval estimators on the spline regression coefficients and variance of the response, confidence interval estimators and (Scheffé [7]) simultaneous confidence interval estimators on the mean value response and prediction value are considered. Inferences are illustrated by some numerical experiments.

1. Introduction

A multiple linear regression model with constant term is given by the functional relation

$$Y = \beta_0 + \sum_{k=1}^{r} \beta_k X_k + \varepsilon,$$

where Y is the response (dependent) variable, $X_1, \ldots, X_r$ are the regressor (independent) variables, and ε represents the error term (random noise).

The multiple linear regression analysis consists in the study of the influence of the variables $X_1, \ldots, X_r$ on the variable Y. This study is realized by the inferences on regression coefficients β_k, and error term ε. In this aim a sample of n data observations

Received by the editors: 01.08.2006.

2000 *Mathematics Subject Classification.* 65D10, 62J05, 62F10, 65C20.

Key words and phrases. Spline regression, multiple linear regression, confidence intervals.

are considered

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; \quad \begin{pmatrix} 1 & x_{11} & \dots & x_{1r} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nr} \end{pmatrix} = (\boldsymbol{1}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_r) = \boldsymbol{X},$$

and the sample multiple linear regression can be written in the matrix form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \dots, \beta_r) \in \mathbb{R}^{r+1}$, $\boldsymbol{\varepsilon}^\top = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$. The classical multiple linear regression model supposes that the random vector $\boldsymbol{\varepsilon}$ follows the normal distribution $\mathcal{N}\left(\boldsymbol{0}; \sigma^2 \boldsymbol{I}_n\right)$, i.e. the components of $\boldsymbol{\varepsilon}$ are independent and identically distributed, each of them following the same normal distribution $\mathcal{N}\left(0; \sigma^2\right)$. A solution $(\boldsymbol{b}, \boldsymbol{e})$, with $\boldsymbol{b} \in \mathbb{R}^{r+1}$, $\boldsymbol{e} \in \mathbb{R}^n$, of the system of equations $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$ is called a fitted multiple linear regression, and the solution satisfying the least-squares criterion

$$\|\boldsymbol{e}\|^2 = \boldsymbol{e}^\top \boldsymbol{e} = \sum_{i=1}^{n} e_i^2 \longrightarrow \text{minim},$$

is called the fitted least-squares multiple linear regression.

It is well-known that the fitted least-squares coefficients are given by

$$\boldsymbol{b} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{1}$$

and these are unbiased estimators of $\boldsymbol{\beta}$. Moreover, we have that

$$s^2 = \frac{1}{n-r-1} \sum_{k=1}^{n} e_k^2 \tag{2}$$

is an unbiased estimator for the parameter σ^2. We remark that the vector of fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$ and the vector of residuals $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ can be expressed by the hat matrix

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top,$$

namely $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, and $\boldsymbol{e} = (\boldsymbol{I}_n - \boldsymbol{H})\,\boldsymbol{y}$, respectively, where $\boldsymbol{I}_n$ denotes identity matrix of order n.

Also, we have the coefficient of determination given by

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}. \tag{3}$$

2. Free knots simple spline regression

The simple spline regression model with distinct free knots $\tau_1, \ldots, \tau_p$

$$Y = \sum_{k=0}^{m} \alpha_k X^k + \sum_{j=1}^{p} \beta_j (\tau_j - X)_+^m + \varepsilon \tag{4}$$

can be reduced to a multiple linear regression model with constant term, if one introduces the new $m+p$ regressor variables $X_k = X^k$, $k = \overline{1,m}$, and $X_{m+j} = (\tau_j - X)_+^m$, $j = \overline{1,p}$.

The spline technique became a very useful in regression analysis, see, for example, [4] and [9]. Some remarks on the number and positions of the knots τ_i are presented in [6] following the suggests given by Wold in [11]: (1) there should be as few knots as possible, with at least four or five data points per segment; (2) there should be no more than one extrem point and one point of inflexion per segment; (3) in so far as possible, the extrem points should be centred in the segment and the point of inflexion should be near the knots.

The transformation on the regressor variables given by Box and Tidwell [3], recalled in [6], was used in [2] to estimate positions of the knots $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$, $\tau_1 < \ldots < \tau_p$.

Let us consider a sample of n pairs of data (x_i, y_i), $i = \overline{1,n}$. The sample spline regression is reduced to a sample multiple linear regression

$$y_i = \alpha_0 + \sum_{k=1}^{m} \alpha_k x_{ik} + \sum_{j=1}^{p} \beta_j x_{i,m+j} + \varepsilon_i, \quad i = \overline{1,n}, \tag{5}$$

where $x_{ik} = x_i^k$, $x_{i,m+j} = (\tau_j - x_i)_+^m$.

Using matrix notation for observations of response variable, coefficients of model, error terms

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_m \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

and design matrix of model

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} & x_{1,m+1} & \dots & x_{1,m+p} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nm} & x_{n,m+1} & \dots & x_{n,m+p} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_1 & \dots & x_1^m & (\tau_1 - x_1)_+^m & \dots & (\tau_p - x_1)_+^m \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^m & (\tau_1 - x_n)_+^m & \dots & (\tau_p - x_n)_+^m \end{pmatrix}$$

the regression model (5) has the matrix expression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{\varepsilon}. \tag{6}$$

Taking into account that the knots of the spline regression are unknown, the following iterative procedure to obtain the knots τ_j is proposed.

For the first time an initial appropriate value $\boldsymbol{\tau}^{(0)} = \left(\tau_1^{(0)}, \dots, \tau_p^{(0)}\right)$ of $\boldsymbol{\tau}$ is considered. Thus, we have an initial spline regression model of type (4) with the attached multiple linear regression model

$$\boldsymbol{y} = \boldsymbol{X}_0\boldsymbol{\delta} + \boldsymbol{\varepsilon}_0, \tag{7}$$

where

$$\boldsymbol{X}_0 = \begin{pmatrix} 1 & x_1 & \dots & x_1^m & \left(\tau_1^{(0)} - x_1\right)_+^m & \dots & \left(\tau_p^{(0)} - x_1\right)_+^m \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^m & \left(\tau_1^{(0)} - x_n\right)_+^m & \dots & \left(\tau_p^{(0)} - x_n\right)_+^m \end{pmatrix},$$

and corresponding vector of errors $\boldsymbol{\varepsilon}_0$. Based on (1), the least-squares estimators of the coefficients $\boldsymbol{\delta}$ of the initial model (7) are given by

$$\boldsymbol{d}_0 = (a_0, \dots, a_m; b_1, \dots, b_p)^\top = \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{X}_0^\top \boldsymbol{y}.$$

For this multiple linear regression model, we have:

- the vector of fitted values (estimated values)

$$\hat{\boldsymbol{y}}^\top = \boldsymbol{X}_0 \boldsymbol{d}_0 = (\hat{y}_1, \dots, \hat{y}_n),$$

- the residual sum of squares

$$\|\boldsymbol{e}_0\|^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \tag{8}$$

- the residual mean squares (unbiased estimator of σ^2)

$$s_0^2 = \frac{1}{n-r-1} \|\boldsymbol{e}_0\|^2 \quad (\text{with } r = m + p), \tag{9}$$

- the coefficient of determination R_0^2 given by (3).

Then, the expanding of

$$h(X; \boldsymbol{\tau}) = h(X; \tau_1, \dots, \tau_p) = \sum_{j=1}^p \beta_j (\tau_j - X)_+^m$$

in Taylor series about the initial value $\boldsymbol{\tau}^{(0)}$ and ignoring terms of higher than first order, we obtain

$$h(X; \boldsymbol{\tau}) = h\left(X; \boldsymbol{\tau}^{(0)}\right) + \left(\boldsymbol{\tau} - \boldsymbol{\tau}^{(0)}\right)^\top \left.\frac{\partial h(X; \boldsymbol{\tau})}{\partial \boldsymbol{\tau}}\right|_{\boldsymbol{\tau} = \boldsymbol{\tau}^{(0)}} + \mathcal{O}\left(\eta^2\right)$$

where $\eta = \max_{j=\overline{1,p}} \left(\left| \tau_j - \tau_j^{(0)} \right| \right)$. Taking into account that

$$\frac{\partial h\left(X;\boldsymbol{\tau}\right)}{\partial \tau_j} = m\beta_j \left(\tau_j - X\right)_+^{m-1}, \quad j = \overline{1,p},$$

it results

$$h\left(X;\boldsymbol{\tau}\right) = h\left(X;\boldsymbol{\tau}^{(0)}\right) + \sum_{j=1}^{p} m\beta_j \left(\tau_j - \tau_j^{(0)}\right) \left(\tau_j^{(0)} - X\right)_+^{m-1} + \mathcal{O}\left(\eta^2\right).$$

Thus, an extended spline regression model is obtained:

$$Y = \sum_{k=0}^{m} \alpha_k X^k + \sum_{j=1}^{p} \beta_j \left(\tau_j^{(0)} - X\right)_+^m + \sum_{j=1}^{p} m\beta_j \left(\tau_j - \tau_j^{(0)}\right) \left(\tau_j^{(0)} - X\right)_+^{m-1} + \tilde{\varepsilon},$$

with a corresponding extended multiple linear regression

$$Y = \alpha_0 + \sum_{k=1}^{m} \alpha_k X_k + \sum_{j=1}^{p} \beta_j X_{m+j} + \sum_{j=1}^{p} \gamma_j X_{m+p+j} + \tilde{\varepsilon},$$

where $\gamma_j = m\beta_j \left(\tau_j - \tau_j^{(0)}\right)$, and the additional regressor variables are given by $X_{m+p+j} = \left(\tau_j^{(0)} - X\right)_+^{m-1}$, $j = \overline{1,p}$. In this way, we have the sample extended multiple linear regression

$$y_i = \alpha_0 + \sum_{k=1}^{m} \alpha_k x_{ik} + \sum_{j=1}^{p} \beta_j x_{i,m+j} + \sum_{j=1}^{p} \gamma_j x_{i,m+p+j} + \tilde{\varepsilon}_i,\ i = \overline{1,n}.$$

We denote by

$$\tilde{\delta}^\top = \left(\alpha_0, \ldots, \alpha_m; \beta_1, \ldots, \beta_p; \gamma_1, \ldots, \gamma_p\right), \quad \tilde{\boldsymbol{\varepsilon}}^\top = \left(\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_p\right),$$

and

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1,m+2p} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{n,m+2p} \end{pmatrix},$$

the vector of coefficients, vector of error terms, and design matrix of the sample extended multiple linear regression. Here, for each $i = \overline{1,n}$, we have

$$x_{ik} = x_i^k,\ k = \overline{1,m};$$
$$x_{i,m+j} = \left(\tau_j^{(0)} - x_i\right)_+^m,\ x_{i,m+p+j} = \left(\tau_j^{(0)} - x_i\right)_+^{m-1},\ j = \overline{1,p}.$$

Thus, the sample extended multiple linear regression has the matrix form

$$\boldsymbol{y} = \tilde{\boldsymbol{X}}\,\tilde{\boldsymbol{\delta}} + \tilde{\varepsilon}. \tag{10}$$

Because $\tilde{\varepsilon} = \varepsilon + \mathcal{O}\left(\eta^2\right)$, it results that $E\left(\tilde{\varepsilon}\right) = E\left(\varepsilon\right) + \mathcal{O}\left(\eta^2\right)\boldsymbol{I}_n \approx \boldsymbol{0}$, and $Var\left(\tilde{\varepsilon}\right) = Var\left(\varepsilon + \mathcal{O}\left(\eta^2\right)\boldsymbol{I}_n\right) = Var\left(\varepsilon\right) = \sigma^2\boldsymbol{I}_n$.

The least-squares estimators of the coefficients $\tilde{\boldsymbol{\delta}}$ are given by

$$\tilde{\boldsymbol{d}} = \left(\tilde{a}_0, \ldots, \tilde{a}_m; \tilde{b}_1, \ldots, \tilde{b}_p; \tilde{c}_1, \ldots, \tilde{c}_p\right)^\top = \left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^\top\boldsymbol{y}.$$

Refering to $\gamma_j = m\beta_j\left(\tau_j - \tau_j^{(0)}\right)$, $j = \overline{1,p}$, we obtain

$$\tau_j = \tau_j^{(0)} + \frac{\gamma_j}{m\beta_j},\quad j = \overline{1,p},$$

and new estimations of coefficients of the linear model (5) can be calculated, considering the new positions of the knots

$$\tau_j^{(0)} := \tau_j^{(0)} + \frac{\tilde{c}_j}{mb_j},\ j = \overline{1,p}.$$

Note that the estimations b_j, $j = \overline{1,p}$, of the coefficients β_j, $j = \overline{1,p}$, obtained on the linear model (6), generally differ from the estimations $\tilde{b}_j$, $j = \overline{1,p}$, of the coefficients β_j, $j = \overline{1,p}$, obtained on the linear model (10).

It is remarked in [6] that the procedure of Box and Tidwell [3] converges quite rapidly, but the round-off error is potentially a problem and successive values of $\boldsymbol{\tau}$ may oscillate widly unless enough decimal places are carried. Convergence problems may be encountered in cases where the error standard deviation of response variable Y is large or when the range of the regressor variable X is very small compared to its expectation.

Table 1 contains the data generated by using the function ([9], p. 45)

$$f(x) = 4.26\left(\mathrm{e}^{-x} - 4\mathrm{e}^{-2x} + 3\mathrm{e}^{-3x}\right), \quad x \in [0, 3.3]. \tag{11}$$

The values of the dependent variable Y are give by

$$y_i = f(x_i) + \varepsilon_i, \quad i = \overline{1,n},$$

where $x_i = (i-1)/30$, $i = \overline{1,100}$, and ε_i, $i = \overline{1,100}$, are independent random numbers following the normal distribution $\mathcal{N}(0; 0.02)$, i.e. the random vector $\boldsymbol{\varepsilon}^\top = (\varepsilon_1, \ldots, \varepsilon_n)$ has multivariate normal distribution with the mean value $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and martrix of covariance is $Var(\boldsymbol{\varepsilon}) = 0.04\boldsymbol{I}_n$.

Table 2 contains the knots τ_i, estimated coefficients a_i and b_i of fitted spline regressions: linear ($m = 1$) with $p = 1, 2, 3$ knots, quadratic ($m = 2$) with $p = 1, 2$ knots, and cubic ($m = 3$) with one knot. The corresponding sum of residual squares (8), residual mean squares (9), and coefficient of determination (3) for each of the fitted spline regresion are given in the same table.

The procedure to obtain the free knots ends if two successive iterations of knots differ less than $\frac{1}{2}10^{-2}$, else the maximum number of iterations is 500. In the second case, the free knots correspond to the minumum norm difference of two successive iterations of the knots of the spline regression no more than 500 iterations.

Figures 1–6 correspond to the six spline regressions and contain for each of them: plot of fitted spline regression (by continuous line), scatter diagram (by circles), positions of knots (by squares), and plot of generator function (11) (by dashed line).

3. **Confidence intervals**

We are interested in giving confidence intervals on the coefficients $\delta_i = \alpha_i$ or β_i of the multiple linear regression (7). If one assumes that the error term $\boldsymbol{\varepsilon}_0$ is normally distributed $\mathcal{N}\left(\mathbf{0}, \sigma^2\boldsymbol{I}_n\right)$, i.e. (7) is a classical multiple linear model, then

i	y_i	i	y_i	i	y_i	i	y_i	i	y_i
1	−0.087	21	−0.516	41	−0.148	61	0.296	81	0.343
2	−0.590	22	−0.789	42	0.242	62	0.231	82	0.373
3	−0.439	23	−0.326	43	−0.005	63	0.511	83	0.397
4	−0.571	24	−0.092	44	0.503	64	−0.085	84	0.005
5	−0.986	25	−0.505	45	0.072	65	0.372	85	0.241
6	−0.614	26	−0.146	46	0.350	66	0.463	86	0.242
7	−0.683	27	−0.020	47	0.298	67	0.426	87	−0.012
8	−0.973	28	−0.545	48	0.079	68	0.392	88	0.037
9	−0.926	29	−0.471	49	−0.163	69	0.281	89	0.397
10	−0.965	30	−0.027	50	0.265	70	0.404	90	0.150
11	−1.032	31	−0.183	51	0.081	71	0.378	91	0.249
12	−0.833	32	0.072	52	0.410	72	0.209	92	0.185
13	−1.069	33	0.132	53	0.393	73	0.180	93	0.036
14	−0.481	34	0.144	54	0.633	74	0.192	94	0.047
15	−0.905	35	0.290	55	0.415	75	−0.048	95	0.243
16	−0.810	36	0.194	56	0.169	76	0.195	96	−0.040
17	−0.572	37	0.325	57	0.375	77	0.261	97	0.302
18	−0.722	38	−0.130	58	0.097	78	0.295	98	0.256
19	−0.701	39	0.129	59	0.294	79	0.516	99	−0.026
20	−0.795	40	0.123	60	0.288	80	0.153	100	0.081

TABLE 1. Values of the dependent variable Y

each of the statistics

$$t_i = \frac{a_i - \alpha_i}{s_i}, \quad i = \overline{0, m},$$
$$t_{m+i} = \frac{b_i - \beta_i}{s_{m+i}}, \quad i = \overline{1, p},$$

	m=1			m=2		m=3
	$p=1$	$p=2$	$p=3$	$p=1$	$p=2$	$p=1$
τ_1	1.715	0.061	0.061	0.061	0.433	0.649
τ_2		1.506	1.360	1.749		
τ_3			2.030			
a_0	0.605	0.433	0.691	−1.119	−0.216	−1.951
a_1	−0.144	−0.082	−0.175	1.260	0.513	2.838
a_2				−0.276	−0.128	−1.133
a_3					0.140	
b_1	−0.888	15.996	15.967	291.236	7.912	6.342
b_2		−0.994	−0.683	0.488		
b_3			−0.408			
$\|\boldsymbol{e}_0\|^2$	4.686	3.797	3.689	4.128	2.904	2.884
s_0^2	0.048	0.040	0.039	0.043	0.031	0.030
$100\,R_0^2$	75.43	80.09	80.66	78.35	84.77	84.87

TABLE 2. Elements of the fitted spline regressions

is T-distributed with $d = n - m - p - 1 = n - r - 1$ degrees of freedom, where

$$s_j^2 = s^2 \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)_{j,j}^{-1}, \quad j = \overline{0, m+p}$$

and $\left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)_{j,j}^{-1}$ denotes the $j+1$-*th* entry of the diagonal of inverse matrix of $\boldsymbol{X}_0^\top \boldsymbol{X}_0$.

Thus, a $100\,(1-\alpha)\,\%$ confidence intervals on the regression coefficients α_i and β_i are given by

$$a_i - t_{d;1-\frac{\alpha}{2}} s_i < \alpha_i < a_i + t_{d;1-\frac{\alpha}{2}} s_i, \quad i = \overline{0, m},$$

$$b_i - t_{d;1-\frac{\alpha}{2}} s_{m+i} < \beta_i < b_i + t_{d;1-\frac{\alpha}{2}} s_{m+i}, \quad i = \overline{1, p},$$

where $t_{d;1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$-quantile of the T-distribution with d degrees of freedom. In the Table 3 are given 95% confidence intervals on the coefficients of the six spline regressions having the elements contained in the Table 2.

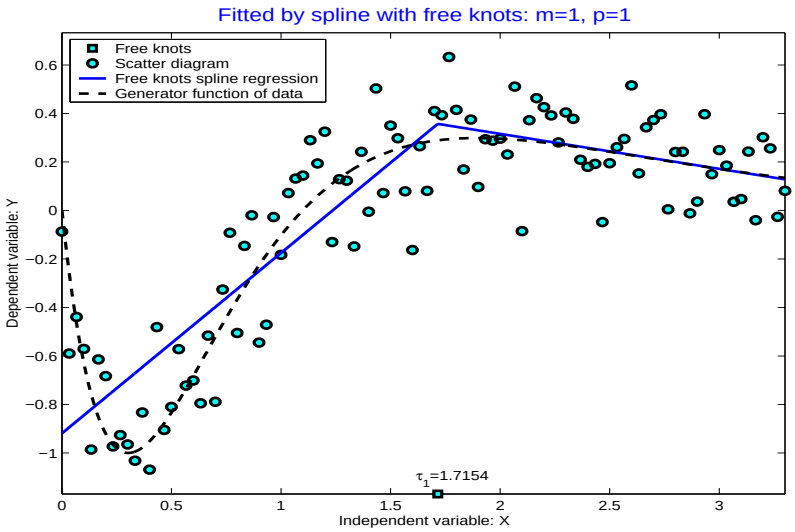

FIGURE 1. Linear spline ($m = 1$) with one knot ($p = 1$)

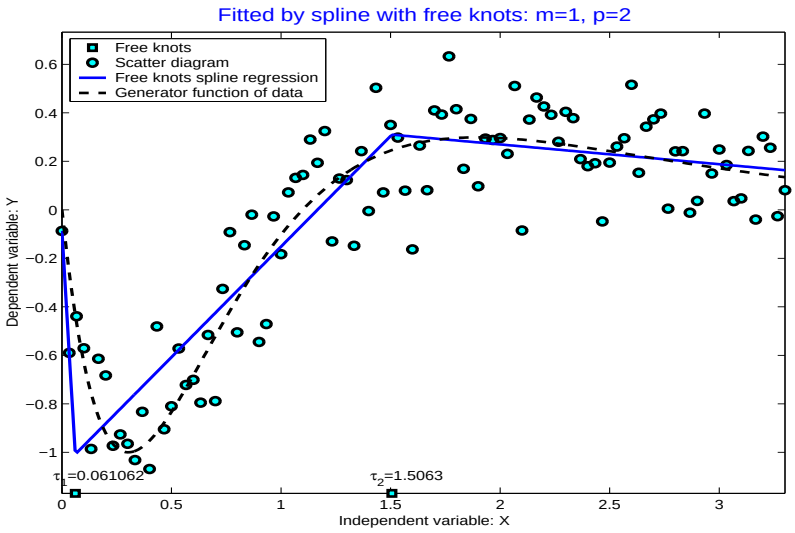

FIGURE 2. Linear spline ($m = 1$) with two knots ($p = 2$)

We have also for the classical multiple linear model (7) that the statistic

$$h^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{d\, s^2}{\sigma^2} = \frac{(n - r - 1)\, s^2}{\sigma^2},$$

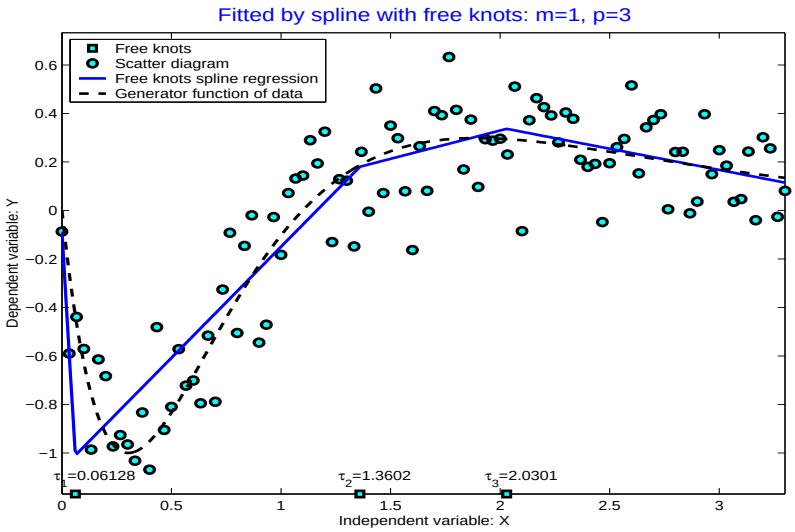

FIGURE 3. Linear spline ($m = 1$) with three knots ($p = 3$)

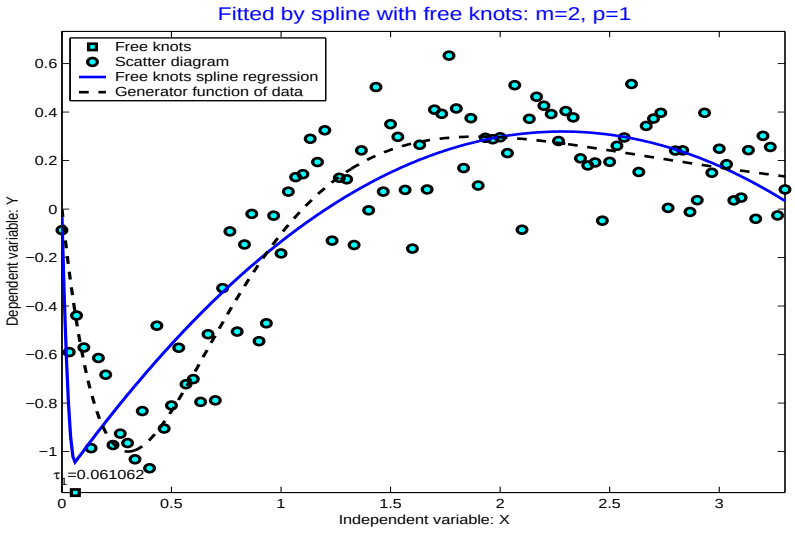

FIGURE 4. Quadratic spline ($m = 2$) with one knot ($p = 1$)

follows a χ^2-distribution with d degrees o freedom. Using this result we have a $100\,(1-\alpha)\,\%$ confidence interval on σ^2

$$\frac{d\,s^2}{\chi^2_{d;1-\frac{\alpha}{2}}} < \sigma^2 < \frac{d\,s^2}{\chi^2_{d;\frac{\alpha}{2}}},$$

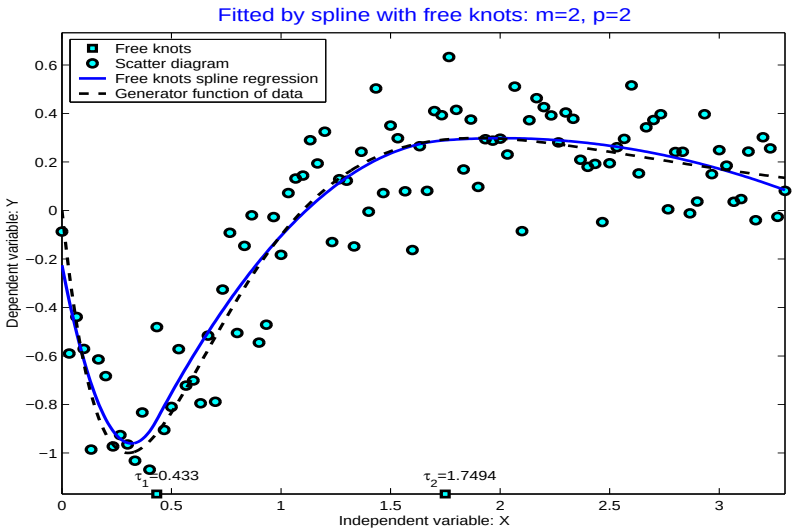

FIGURE 5. Quadratic spline ($m = 2$) with two knots ($p = 2$)

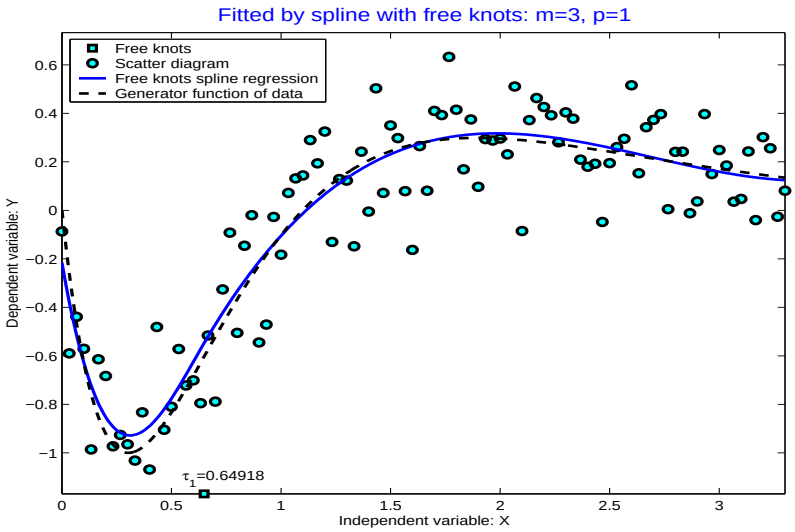

FIGURE 6. Cubic spline ($m = 3$) with one knot ($p = 1$)

where $\chi^2_{d;\gamma}$ denotes the γ-quantile of the χ^2 distribution with d degrees of freedom. The Table 3 contains also 95% confidence intervals on σ^2 of the six examples of spline regressions considered in the previous section.

	m=1		
	$p=1$	$p=2$	$p=3$
α_0	$(0.350, 0.861)$	$(0.238, 0.628)$	$(0.317, 1.066)$
α_1	$(-0.251, -0.038)$	$(-0.165, 0.002)$	$(-0.318, -0.032)$
β_1	$(-1.070, -0.707)$	$(9.696, 22.295)$	$(9.680, 22.253)$
β_2		$(-1.165, -0.822)$	$(-1.015, -0.350)$
β_3			$(-0.745, -0.072)$
σ^2	$(0.0371, 0.0654)$	$(0.0304, 0.0536)$	$(0.0298, 0.0528)$
	m=2		m=3
	$p=1$	$p=2$	$p=1$
α_0	$(-1.248, -0.990)$	$(-0.919, 0.486)$	$(-2.239, -1.663)$
α_1	$(1.083, 1.438)$	$(-0.137, 1.164)$	$(2.212, 3.465)$
α_2	$(-0.328, -0.225)$	$(-0.270, 0.014)$	$(-1.519, -0.747)$
α_3			$(0.070, 0.211)$
β_1	$(176.042, 406.431)$	$(6.066, 9.757)$	$(4.824, 7.860)$
β_2		$(-0.764, -0.211)$	
σ^2	$(0.0330, 0.0583)$	$(0.0234, 0.0415)$	$(0.0233, 0.0412)$

TABLE 3. Confidence intervals for coefficients and variation

From the construction and theoretical results on the multiple linear model (7), it results that an unbiased estimator of the mean response $E(Y \mid \boldsymbol{x})$ at a point $\boldsymbol{x}^\top = \left(1; x, \ldots, x^m; (\tau_1 - x)_+^p, \ldots, (\tau_p - x)_+^p\right)$ is given by

$$\hat{y} = \boldsymbol{x}^\top \boldsymbol{d}_0,$$

and $Var(\hat{y}) = \sigma^2 \boldsymbol{x} \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}$. For the classical multiple linear model, the statistic

$$t_{\boldsymbol{x}} = \frac{\hat{y} - E(Y \mid \boldsymbol{x})}{s\sqrt{\boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}}} = \frac{\boldsymbol{x}^\top \boldsymbol{d}_0 - \boldsymbol{x}^\top \boldsymbol{\delta}}{s\sqrt{\boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}}} \tag{12}$$

is T-distributed with d degrees of freedom. Using the statistic $t_{\boldsymbol{x}}$, the following $100\,(1-\alpha)\,\%$ confidence interval on the mean response $E\,(Y \mid \boldsymbol{x})$ can be obtained

$$\hat{y} - t_{d;1-\frac{\alpha}{2}} s \sqrt{\boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}} < E\,(Y \mid \boldsymbol{x}) < \hat{y} + t_{d;1-\frac{\alpha}{2}} s \sqrt{\boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}}.$$

In a similar manner, to construct a confidence interval for a predicted value y of the response Y, corresponding to a new value x of the regressor X, we have the statistic

$$t_{\boldsymbol{x}} = \frac{\hat{y} - y}{s\sqrt{1 + \boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}}} = \frac{\boldsymbol{x}^\top \boldsymbol{d}_0 - y}{s\sqrt{1 + \boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}}}, \tag{13}$$

where again $\boldsymbol{x}^\top = \left(1; x, \ldots, x^m; (\tau_1 - x)_+^p, \ldots, (\tau_p - x)_+^p\right)$, and $t_{\boldsymbol{x}}$ is T-distributed with d degrees of freedom. Thus, a $100\,(1-\alpha)\,\%$ confidence interval on the predicted response y is given by

$$\hat{y} - t_{d;1-\frac{\alpha}{2}} s \sqrt{1 + \boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}} < y < \hat{y} + t_{d;1-\frac{\alpha}{2}} s \sqrt{1 + \boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}}.$$

We remark that $\boldsymbol{\tau}$ is the last position $\boldsymbol{\tau}_0$ obtained by iterative procedure and s^2 is the corresponding residual sum of square given by (9).

Figures (7), (8) and (9) contain plots of 95% confidence intervals on the mean response and prediction with respect to x for three of the six considered spline regressions. Each figure contains scatter diagram (circles), plot of spline regression function (solid line), plot of confidence interval on mean response (dash-dot line), plot of confidence interval on prediction (dashed line), and positions of the knots (squares).

The construction of simultaneous confidence intervals on the mean response and prediction uses the Scheffé's result. Namely, if C represents a set of points $\boldsymbol{x}^\top = \left(1; x, \ldots, x^m; (\tau_1 - x)_+^p, \ldots, (\tau_p - x)_+^p\right)$, and considering $W = \sup_{\boldsymbol{x} \in C} t_{\boldsymbol{x}}^2$, were $t_{\boldsymbol{x}}^2$ is given by (12) and (13) respectively, then the statistic $W/\,(m+p+1)$ is F-distributed with $(m+p+1, n-m-p-1) = (r+1, d)$ degrees of freedom.

In this way, we have a $100\,(1-\alpha)\,\%$ Scheffé simultaneous confidence interval on the mean response

$$\hat{y} - Ks \sqrt{\boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}} < E\,(Y \mid \boldsymbol{x}) < \hat{y} + Ks \sqrt{\boldsymbol{x}^\top \left(\boldsymbol{X}_0^\top \boldsymbol{X}_0\right)^{-1} \boldsymbol{x}},$$

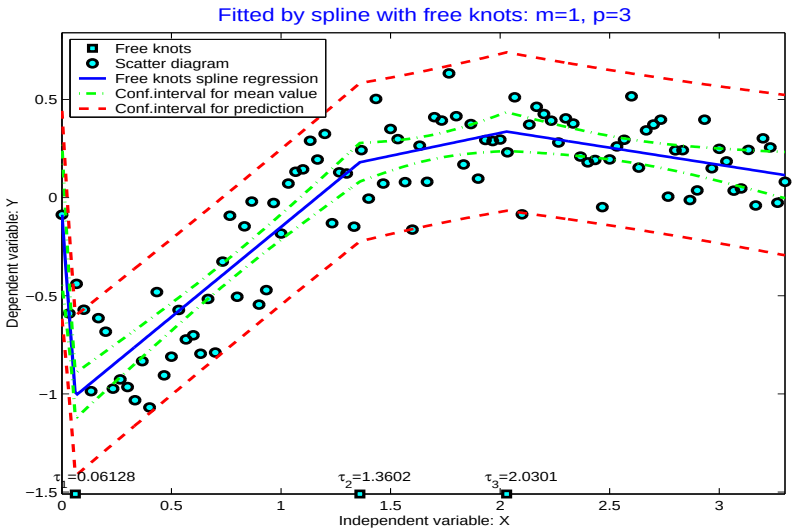

FIGURE 7. Linear spline ($m = 1$) with three knots ($p = 3$)

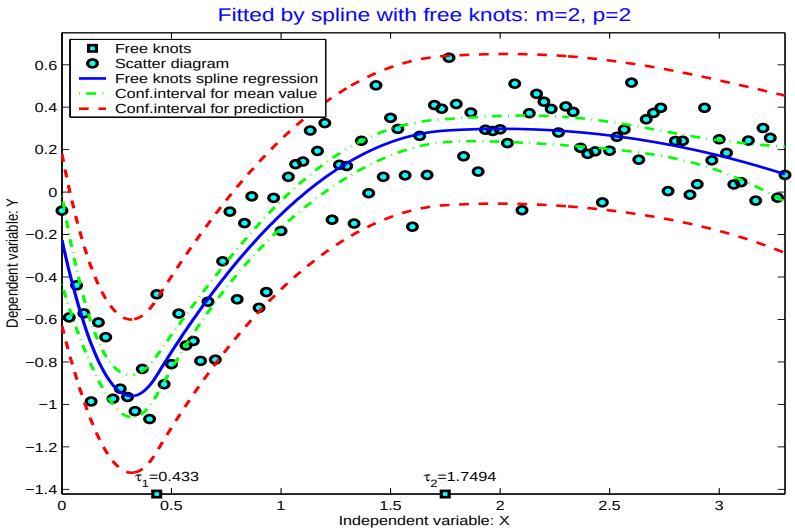

FIGURE 8. Quadratic spline ($m = 2$) with two knots ($p = 2$)

and $100\,(1-\alpha)\,\%$ Scheffé simultaneous confidence interval on the prediction

$$\hat{y} - Ks\sqrt{1 + \boldsymbol{x}^{\top}\left(\boldsymbol{X}_0^{\top}\boldsymbol{X}_0\right)^{-1}\boldsymbol{x}} < y < \hat{y} + Ks\sqrt{1 + \boldsymbol{x}^{\top}\left(\boldsymbol{X}_0^{\top}\boldsymbol{X}_0\right)^{-1}\boldsymbol{x}}.$$

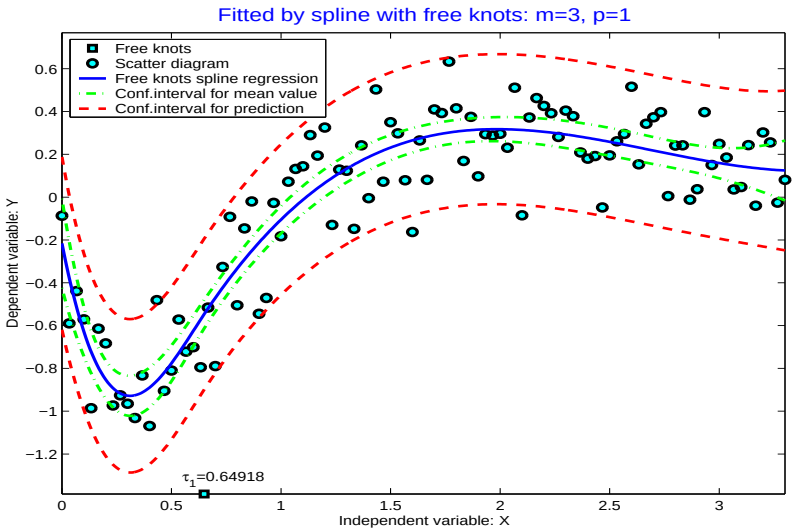

FIGURE 9. Cubic spline ($m = 3$) with one knot ($p = 1$)

Here $K = \sqrt{(r+1)\, f_{r+1,d;1-\alpha}}$, where $f_{r+1,d;1-\alpha}$ is $1-\alpha$-quantile of the F distribution with $(r+1, d)$ degrees of freedom.

Figures (10), (11) and (12) contain plots of 95% simultaneous confidence intervals on the mean response and prediction with respect to x for three of the six considered spline regressions. Each figure contains scatter diagram (circles), plot of spline regression function (solid line), plot of simultaneous confidence interval on mean response (dash-dot line), plot of simultaneous confidence interval on prediction (dashed line), and positions of knots (squares).

References

[1] Agratini, O., Blaga, P., Coman, Gh., *Lectures on Wavelets, Numerical Methods and Statistics*, Science Book House, Cluj-Napoca, 2005.

[2] Blaga, P. P., *Free knots for spline regression*, Annals of Tiberiu Popoviciu Seminar of Functional Equations, Approximation and Convexity, **3**(2005), 3–17.

[3] Box, G. E. P., Tidwell, W., *Transformation of the independent variables*, Technometrics, **4**(1962), 531–550.

[4] Eubank, R. L., *Spline smoothing and nonparametric regression*. Dekker, New York, 1988.

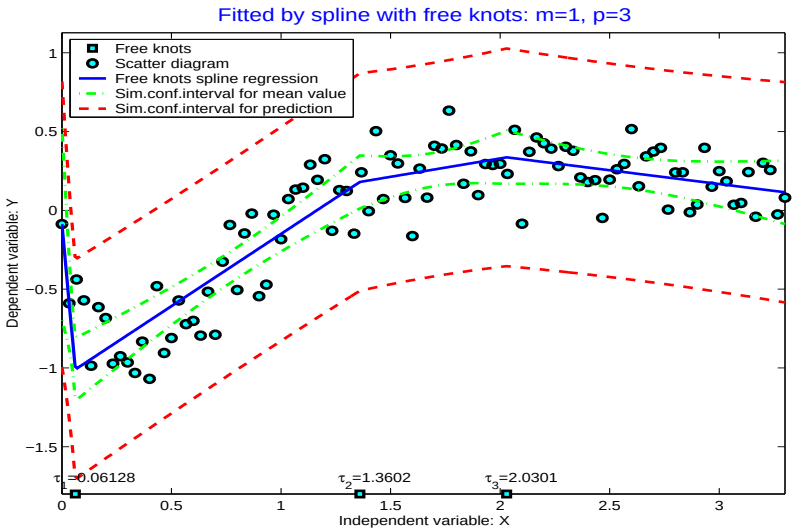

FIGURE 10. Linear spline ($m = 1$) with three knots ($p = 3$)

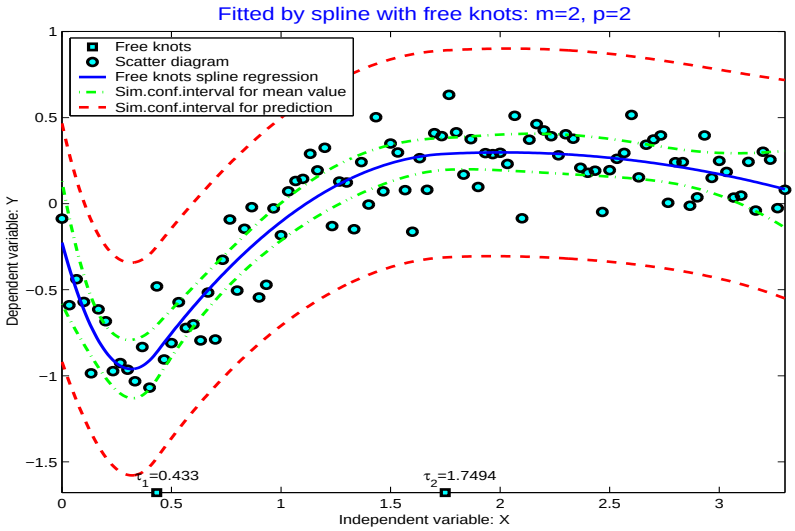

FIGURE 11. Quadratic spline ($m = 2$) with two knots ($p = 2$)

[5] Lehmann, E. L. *Testing statistical hypotheses (Second edition)*, Springer, New York-Berlin, 1997.

[6] Montgomery, D. C., Peck, E. A., Vining, G. G., *Introduction to linear regression analysis (Third edition)*, John Wiley & Sons, New York, 2001.

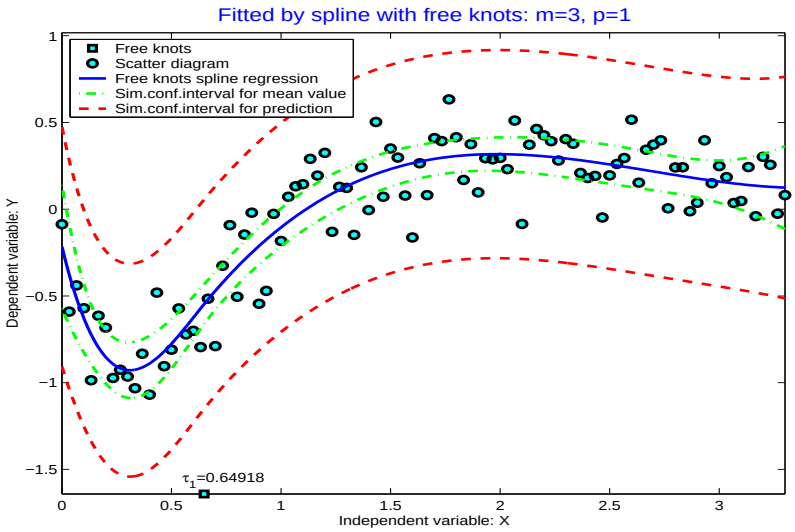

FIGURE 12. Cubic spline ($m = 3$) with one knot ($p = 1$)

[7] Scheffé, H., *The analysis of variance*, Wiley, New York, 1959.

[8] Stapleton, J. H., *Linear statistical models*, John Wiley & Sons, New York-Chichester-Brisbane, 1995.

[9] Wahba, W., *Spline models for observational data*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[10] Wahba, G., Wold, S., *A completely automatic French curve*, Commun. Statist., **4**(1975), 1–17.

[11] Wold, S., *Spline functions in data analysis*, Technometrics, **16**(1974), 1–11.

"BABEŞ–BOLYAI" UNIVERSITY, CLUJ–NAPOCA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
STR. KOGĂLNICEANU 1, 400084 CLUJ–NAPOCA, ROMANIA
E-mail address: blaga@math.ubbcluj.ro