MODELS OF NATURAL CUBIC SPLINES FOR HEAVY METALS EXPOSURE ASSESSMENT

R. Kregzdyte*, I. Patasiene**, M. Patasius***and A. Kazakeviciute**

*Institute for Biomedical Research, Kaunas University of Medicine, Kaunas, Lithuania
**Kaunas University of Technology, Faculty of Social Sciences, Kaunas, Lithuania
***Kaunas University of Technology, Faculty of Informatics, Kaunas, Lithuania

rimak@kmu.lt, cipatasi@cr.ktu.lt, marpata@ktu.lt, agne63@yahoo.com

Abstract: The aim of our study was to build models for estimation of the relationship between internal dose of lead, manganese, cadmium and health indices (arterial blood pressure and blood colour index). As collected data did not follow standard functional form we applied technique of cubic splines. Established models revealed nonmonotonous increase or decrease of arterial blood pressure and blood colour index in dependence on heavy metals concentration in hair.

Introduction

The main goal of environmental health research is to assess the human exposure to environmental pollutants or pollutant dose. Sometimes collected data do not follow standard functional form such as polynomial or common growth curve. The dependence of a response variable y on an explanatory variable x may be expressed as:

$$y = f(x) + e \qquad (1)$$

Nonparametric or semiparametric estimation methods are suitable for finding the estimator. One of the methods is smoothing cubic splines [1, 2, 3].

The aim of our study was to build cubic splines models for estimation of the relationship between internal dose of heavy metals (lead, manganese and cadmium) and health indices (arterial blood pressure (ABP) and blood color index (BCI)).

Material and methods

Data of health survey of 1117 adults differently exposed to heavy metals was analysed. Lead (Pb), manganese (Mn) and cadmium (Cd) concentration in hair was used as indicator of internal dose of different exposure.

As the suitable parametric form of the model describing relationship between internal dose of heavy metals (lead, manganese and cadmium) and health indices was not found, it was decided to search for another form of model. In such case nonparametric smoothing methods were considered. Main methods of this class are moving average, moving line, kernel and cubic splines [2].

In model of multiple linear regression dependence of response variable y of variables $z_1, ..., z_n$ is expressed as a sum of different factors with evaluated parameters:

$$E(y \mid z_1,...z_n) = \beta_0 + \beta_1 z_1 + ... + \beta_n z_n \qquad (2)$$

Using smoothing methods evaluations of variable functions are analyzed instead of evaluations of parameters ($\beta_0, ..., \beta_n$):

$$E(y \mid z_1,...z_n) = f_1(z_1) + ... + f_n(z_n) \qquad (3)$$

One of such functions is a spline, that is a function, continuous till p-th derivative inclusively, that consists of parts of some function (usually this function of a polynom of n-th order).

The simplest spline with p=0 is a broken line. Cubic spline consists of parts of cubic polynom and its $0 \le p \le 2$.

If $x_0, x_1, x_2, ..., x_k$ are values of interval $[a; b]$ such that $a \le x_0 \le x_1 \le x_2 ... \le x_k \le b$, then the cubic spline is a function g(x) that satisfies these conditions:

- in every interval $[x_{i-1}; x_i]$ $(i=1, ..., k)$ $g(x)$ is a cubic polynom;
- $g(x)$ is continuous in the interval [a; b] and in each of inner points of grid x_i $(i=1, ..., k-1)$ equation $g^{(l)}(x_i-0) = g^{(l)}(x_i+0)$, $l=0, 1, 2$ is true, that is the derivatives of the first and the second order are also continuous in each of points x_i.

Evaluation of the function $f(x)$ – cubic spline with given natural initial conditions ($g''(a)=0$, $g''(b)=0$) – can be found as a solution of the problem of minimization of functional

$$\Phi(f) = \sum_{i=0}^{k} (y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx \qquad (4)$$

Of all the functions $f(x)$ the one that minimizes the sums of square of remaining values is found. The first part of the functional (4) is the sum of square of remaining values that is used as distance function between data and evaluated values. The second part of the functional defines the curvature of the function. Parameter $\lambda \ge 0$ is a coefficient that controls the

ISSN: 1727-1983

roughness of the smoothing and $p_i>0$ is a weighting coefficient [4,5,6].

After calculations the spline in interval $[x_{i-1}, x_i]$ can be expressed as:

$$g(X)=m_{i-1}\frac{(x_i-x)^3}{6h_i}+m_i\frac{(x-x_{i-1})^3}{6h_i}+(\tilde{y}_{i-1}-\frac{m_{i-1}h_i^2}{6})\frac{x_i-x}{h_i}+(\tilde{y}_i-\frac{m_ih_i^2}{6})\frac{x-x_{i-1}}{h_i} \quad (5)$$

Here:

$$i=1,...,k \quad (6)$$

$$m_i=g''(x_i) \quad (7)$$

$$h_i=x_i-x_{i-1} \quad (8)$$

and $\tilde{y}_i$ is the calculated value of y_i.

In this expression the second derivatives of the spline (m_i) are unknown and can be calculated by solving the system of equations:

$$(A+H\times P^1\times H^t)\times m=H\times y \quad (9)$$

Here A is a quadratic matrix of (k-1) order:

$$A=\begin{bmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & \cdots & 0 & 0 \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & \cdots & 0 & 0 \\ 0 & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \frac{h_{k-1}}{6} & \frac{h_{k-1}+h_k}{3} \end{bmatrix} \quad (10)$$

m and y are matrixes-columns:

$$m=\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{k-1} \end{bmatrix} \qquad y=\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix} \quad (11)$$

H is a matrix of size (k-1) × (k+1):

$$H=\begin{bmatrix} \frac{1}{h_1} & -\left(\frac{1}{h_1}+\frac{1}{h_2}\right) & \frac{1}{h_2} & \cdots & 0 & 0 \\ 0 & \frac{1}{h_2} & -\left(\frac{1}{h_2}+\frac{1}{h_3}\right) & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\left(\frac{1}{h_{k-1}}+\frac{1}{h_k}\right) & \frac{1}{h_k} \end{bmatrix} \quad (12)$$

H^t – transposed matrix H.

P^{-1} is a quadratic matrix of (k+1) order:

$$P^{-1}=\begin{bmatrix} \frac{1}{p_0} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{p_1} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \frac{1}{p_k} \end{bmatrix} \quad (13)$$

Solving equation system (6) the values of the second derivatives of the cubic spline can be found. Values of the spline $\tilde{y}$ can be calculated from system of equations:

$$\tilde{y}=y-P^{-1}\times H^t\times m \quad (14)$$

The higher value of weighted coefficient p_i, the lesser the difference between $\tilde{y}_i$ and y_i, that is between value of original and smothered function.

After writing values of m_i and $\tilde{y}_i$ in expression (5) values of spline, minimizing the functional (4) can be evaluated in any point $c\in[x_0, x_k]$.

The estimator of function f(x), which defines dependence of health index on heavy metal concentration, is a cubic spline. That is a function g such that on each of the intervals [a,t_0[, [t_0,t_1[,…, [t_k,b], g is a cubic polynomial, and the polynomial pieces fit together at the knots t_i in such a way that g itself and its first and second derivatives are continuous at each t_i and on the whole of the interval [a,b].

Splines models are suitable for finding the threshold concentration tha is important in environmental health studies [7].

Tool for longitudinal data analysis

For more detailed data analysis it is important to collect, to save and to keep data in the database (DB). As the dose of lead, manganese, and cadmium depends on the time it is useful to iterate measurements. Multiple calculations need suitable database. For this purpose relational database is the best type of DB. Related tables allow keeping data with undefined number of records [8, 9]. In our case the number of records in table doesn't need to be the same for each patent as it is difficult to organize measurement for everybody at the same time.

Specialized statistical software package STATA was used for detailed data analysis. STATA, SPSS or other typical software package usually use one main data table. So, it is necessary to prepare sub database needed for data analysis. For this purpose we suggest to design the set of queries and keep them in the SQL queries library.

The schema of complex organizing longitudinal data is shown in Figure 1. The schema shows the sequence of integrated processes from collection of data to analysis of them that generates new knowledge. MS

ISSN: 1727-1983

ACCESS DBMS (Data Base Management system) was used for designing Database, executing queries and exporting/importing data.

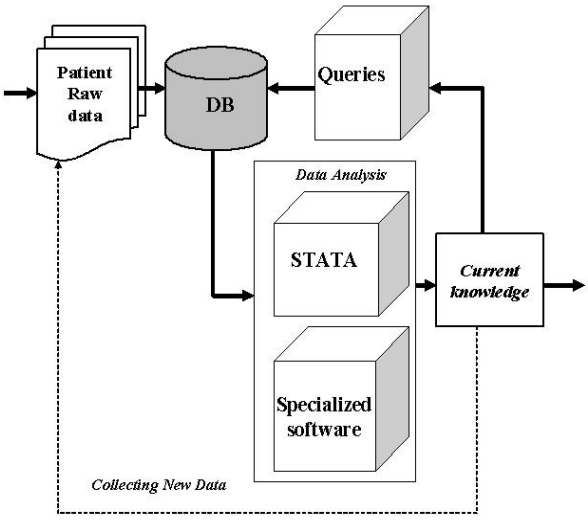

Figure 1: The schema of organizing longitudinal data analysis

The tool helps researcher to organize multiple integrated data analysis.

Results

Cubic splines models for arterial blood pressure were built separately for men and women. The following models were built for men's ABP:

$$Systolic\ \ ABP = g(Pb) + 1.23 \cdot BMI + 0.46 \cdot Age + \varepsilon \qquad (15)$$

$$Diastolic\ \ ABP = g(Pb) + 1.47 \cdot BMI + 0.13 \cdot Age + \varepsilon \qquad (16)$$

Here g(Pb) is natural cubic spline of Pb concentration in hair, BMI is body mass index, ε is error. Two threshold values of Pb concentration for men's systolic ABP were found: systolic ABP increased rapidly until 4 μg/g, after that decreased slowly until 32.5 μg/g and then increased slowly (Figure 2).

Diastolic ABP increased rapidly until 4 μg/g, and after that increased slowly (Figure 3).

The following models were built for women's ABP:

$$Systolic\ \ ABP = g(Pb) + 0.83 \cdot BMI + 0.85 \cdot Age + \varepsilon \qquad (17)$$

$$Diastolic\ \ ABP = g(Mn) + 0.74 \cdot BMI + 0.31 \cdot Age + \varepsilon \qquad (18)$$

Here g(Pb) is natural cubic spline of Pb concentration in hair, g(Mn) is natural cubic spline of Mn concentration in hair, BMI is body mass index, ε is error.

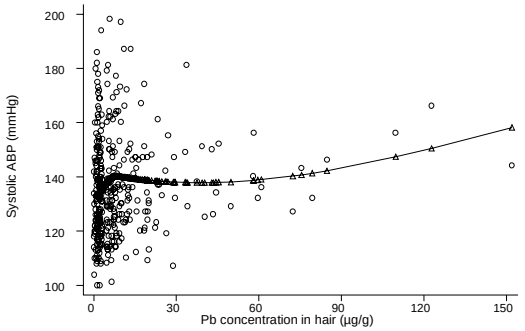

Figure 2: The relationship between systolic arterial blood pressure and Pb concentration in hair of men

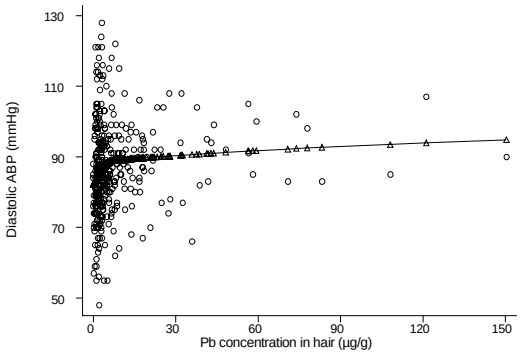

Figure 3: The relationship between diastolic arterial blood pressure and Pb concentration in hair of men

Cubic dependence of systolic ABP on Pb concentration of women was revealed (Figure 4).

Threshold value of Mn concentration for women's diastolic ABP was identified: ABP decreased rapidly until 3μg/g and decreased slowly after that value (Figure 5).

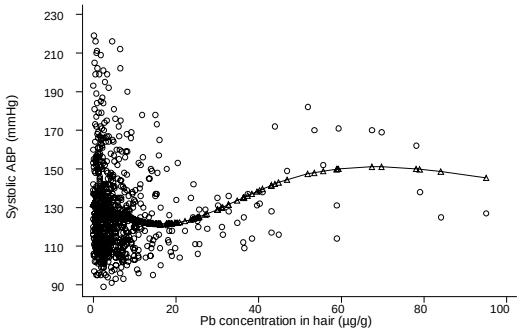

Figure 4: The relationship between systolic arterial blood pressure and Pb concentration in hair of women

ISSN: 1727-1983

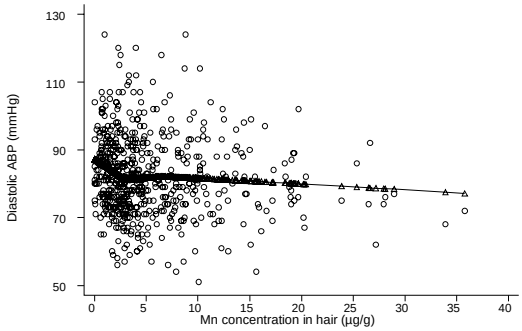

Figure 5: The relationship between diastolic arterial blood pressure and Mn concentration in hair of women

Cubic splines g(Pb) and g(Cd) were calculated for the estimation of heavy metals effect on blood color index. According to models:

$$BCI = g(Pb) - 0.001 \cdot lengh\ of\ \ service + \varepsilon \qquad (19)$$

$$BCI = g(Cd) - 0.001 \cdot lengh\ of\ \ service + \varepsilon \qquad (20)$$

BCI decreased linearly as Pb concentration increased (Figure 6), and the curvilinear dependence of BCI on Cd concentration in hair was determined (Figure 7).

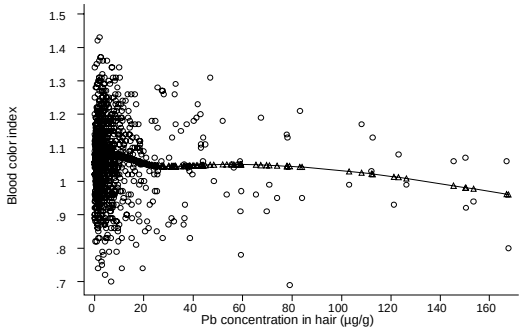

Figure 6: The relationship between blood color index and Pb concentration in hair

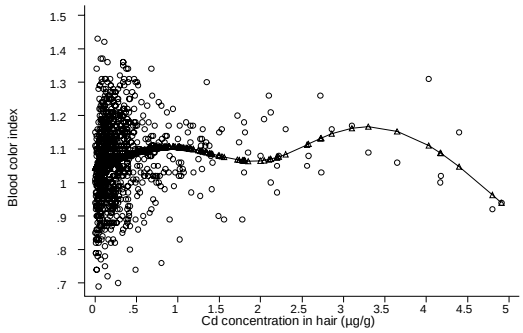

Figure 7: The relationship between blood color index and Cd concentration in hair

Such models could be improved using additional repeated data systematically collected in DB.

Conclusions

Cubic splines are suitable for the estimation of the relationship between heavy metals concentrations in hair and health indices.

Models of cubic splines revealed nonmonotonous increase or decrease of arterial blood pressure and blood color index in dependence on lead, manganese and cadmium internal dose.

Data Base Management System is necessary for systematical collecting and analysis of data.

References

[1] FAIRES, J. D., BURDEN, R. L. (1993): 'Numerical methods', (PWS, Boston), pp. 74-81.

[2] HASTIE, T. J., TIBSHIRANI, R. J. (1991): 'Generalized additive models', (Chapman & Hall, London), pp. 9-25.

[3] PLUKAS, K. (2001): 'Skaitiniai metodai ir algoritmai', eng. Numerical methods and algorithms, (Naujasis lankas, Kaunas), pp. 319-350

[4] MOLINARI N., DURAND J.-F., SABATIER R. (2004): 'Bounded optimal knots for regression splines', *Comput. Statist. Data Anal.*, 45, pp. 159-178

[5] LINDSTROM M.J. (2002): 'Bayesian estimation of free-knot splines using reversible jumps', *Comput. Statist. Data Anal.*, 41, pp. 255-270

[6] LI T.C.M. (2003): 'Smoothing parameter selection for smoothing splines: a simulation study', *Comput. Statist. Data Anal.*, 42, pp. 139-148

[7] LI C.-S., HUNT D. (2004): 'Regression splines for threshold selection with application to a random-effects logistic dose-response model', *Comput. Statist. Data Anal., 46*, pp. 1-10

[8] KAZAMEKAITIS A., PATASIENE I. and KACINSKAS K. (1999): "Municipality Health Care Department information system as a management tool and efforts to create it", Proc. of LITMED 1999 - International Symposium New Frontiers in Public Health, Klaipeda, Lithuania, 1999, p. 13-14

[9] JARKE M., LENZERINI M., VASSILIOU Y. and VASSILIADIS P. (2003): 'Fundamentals of Data Warehouses', (Springer, Berlin), pp.87-121

ISSN: 1727-1983