EMI, Vol. 5, Issue 1, 2013
ISSN: 1804-1299 (Print), 1805-353X (Online) Article

VAGUENESS OF REGRESSION MODELS PART I – LINEAR REGRESSION FUZZIFICATION

NEURČITOST REGRESNÍHO MODELU ČÁST I – FUZZIFIKACE LINEÁRNÍ REGRESE

Miroslav Pokorný
Moravian University College Olomouc, Department of Computer Science and Applied Mathematics, Czech republic
miroslav.pokorny@mvso.cz

Jana Nowaková
VŠB-Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science, Czech republic
jana.nowakova@vsb.cz

Abstract:
The theoretical background for abstract formalization of the vague phenomenon of complex systems is the fuzzy set theory. In the paper, vague data is defined as specialized fuzzy sets - fuzzy numbers and a fuzzy linear regression model is described as a fuzzy function with fuzzy numbers as vague regression parameters. To identify the fuzzy coefficients of the model, the genetic algorithm is used. The linear approximation of the vague function together with its possibility area is analytically and graphically expressed. Suitable numerical experiments are performed, namely the task of two-dimensional fuzzy function modelling and the time series fuzzy regression analysis.

Keywords:
Complex systems, vague property, fuzzy set, fuzzy number, fuzzy linear regression, genetic algorithms, possibility area

Abstrakt:
Teoretickým zázemím pro abstraktní formalizaci neurčitosti kompleních soustav je fuzzy množinová teorie. V příspěvku je definována vágnost pozorovaných dat pomocí specializovaných fuzzy množin – fuzzy čísel a je prezentován lineární regresní model ve formě fuzzy funkce s fuzzy čísly jako vágními regresními parametry. Pro identifikaci fuzzy koeficientů je použit genetický algoritmus. Lineární aproximace vágní funkce spolu s neurčitým (vágním) pásmem jsou vyjádřeny analyticky I graficky. Jsou uvedeny dva numerické experimenty – fuzzy lineární regresní model dvourozměrné funkce a fuzzy regresní analýza časové řady.

Klíčová slova:
Komplexní systém, vágnost, fuzzy množina, fuzzy číslo, fuzzy lineární regrese, genetický algoritmus, neurčité pásmo

JEL Classification: *C4*

1 Introduction

Regression models are often used in engineering practice wherever there is a need to reflect more independent variables together with the effects of other unmeasured disturbances and influences. In classical regression, we assume that the relationship between dependent variables and independent variables of the model is well-defined and sharp. In the real world, however, hampered by the fact that this relationship is more or less non-specific and vague. This is particularly true when modelling complex systems which are difficult to define, difficult to measure or in cases where it is incorporated into the human element.

The suitable theoretical background for abstract formalization of vague phenomenon of complex systems is fuzzy set theory. In the paper are defined vague data as specialized fuzzy sets - fuzzy numbers. Next, a fuzzy linear regression model as a fuzzy function with fuzzy numbers as vague parameters is identified using the genetic algorithms.

2 Regression Analysis

Widely used linear regression model of investigated system [1] is given by a linear combination of values of its input variables

$$Y^*(x_j) = A_0 + A_1 x_{1j} + ... + A_j x_{nj} \quad (1)$$

Conventional regression model (1) is based on the assumption that the dependency of input and output variables is approximately linear and the system characteristic is defined by sharp, precise. Deviations between observed and estimated values of the dependent variables are the result of errors of observation. The origin of the deviation between the observed and estimated values of the dependent variables may not be significant extent caused by poor local variables of system structure. Also in cases where it is incorporated into the human element.

The causes of these variations are in not very sharp nature of the system parameters. Such fuzzy phenomenon must also reflected in fuzziness of the corresponding parameters of the model.

3 Fuzzy Linear Regression

Regression models reflecting the vagueness of the modelled systems are called fuzzy regression models [2], [3], [4], [5], [6]. The indeterminate nature of fuzzy regression model is represented by the estimated fuzzy output values $\tilde{Y}^*(x_j)$ and the fuzzy regression coefficients $\tilde{A}$ in the form of specialized fuzzy sets - fuzzy numbers. Shape of fuzzy linear regression model is given by

$$\tilde{Y}^*(x_j) = \tilde{A}_0 + \tilde{A}_1 x_{1j} + ... + \tilde{A}_j x_{nj} = \tilde{\mathbf{A}}.\mathbf{x}' \quad (2)$$

where x´ is a transpose column vector x = $(x_1, x_2, ..., x_n)$ and $\tilde{\mathbf{A}}$ is a parameter vector whose elements are fuzzy numbers. In the fuzzy regression function is $\tilde{\mathbf{A}}$ the multi-dimensional fuzzy sets (fuzzy relation) as the Cartesian product of fuzzy sets of fuzzy parameters

$$\tilde{\mathbf{A}} = \tilde{A}_1 \times \tilde{A}_1 \times ... \times \tilde{A}_n$$

with membership function in the form

$$\mu_{\tilde{\mathbf{A}}}(\mathbf{a}) = \bigcup_{i=1}^{n} \left\{ \mu_{\tilde{A}_i}(a_i) \right\}, \quad \mathbf{a} = (a_1, a_2, ..., a_n)$$

The shape of the membership function of fuzzy numbers output value of fuzzy linear regression model (1) is

calculated by Zadeh´s extensional principle [7] in the form

$$\mu_{\tilde{Y}}(y) = \bigcup_{\mathbf{a}|\mathbf{a}t'=y} \mu_{\tilde{\mathbf{A}}}(\mathbf{a}) \; ; \; \{\mathbf{a}|\mathbf{t}'=y\} = 0$$
$$0 \quad ; \text{ jinak} \tag{3}$$

Membership function $\mu_{\tilde{A}_i}(a_i)$ is approximated in the form of triangular fuzzy numbers [7]

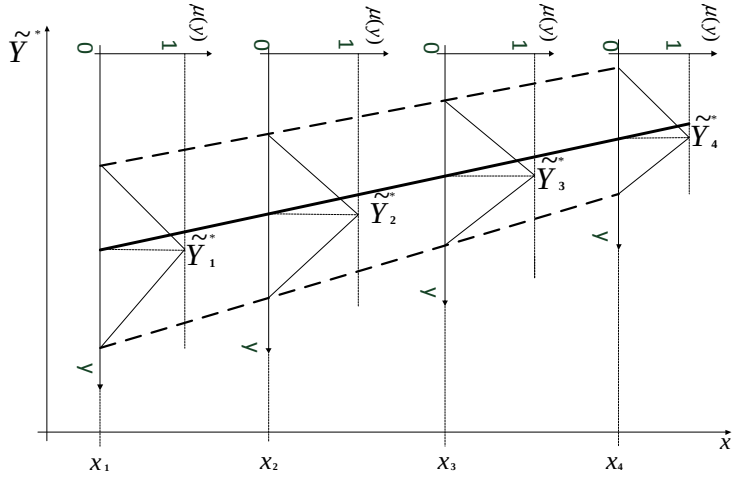
$$\mu_{\tilde{A}_i}(a_i) = 1 - \frac{|\alpha_i - a_i|}{c_i} \; ; \quad \alpha_i - c_i \le a_i \le \alpha_i + c_i \tag{4}$$
$$0 \quad ; \text{ jinak}$$

where α_i is the mean value (core) of fuzzy number $\tilde{A}_i$ and c_i is half of the width of the carrier bearing $\tilde{A}_i = \{\alpha_i, c_i\}$. The term of membership functions for the output fuzzy sets (3) can be written in the form [1]

$$\mu_{\tilde{Y}}(y) = 1 - \frac{|y - \alpha.x'|}{\sum_{i=1}^{n} c_i |x_i|} \; ; \; \alpha.x' - \sum_{i=1}^{n} c_i |x_i| \le y \le \alpha.x' + \sum_{i=1}^{n} c_i |x_i| \tag{5}$$
$$0 \quad ; \text{ jinak}$$

and such triangular approximation of membership function is used in next Figure 1

Figure 1: One-dimensional linear fuzzy regression function



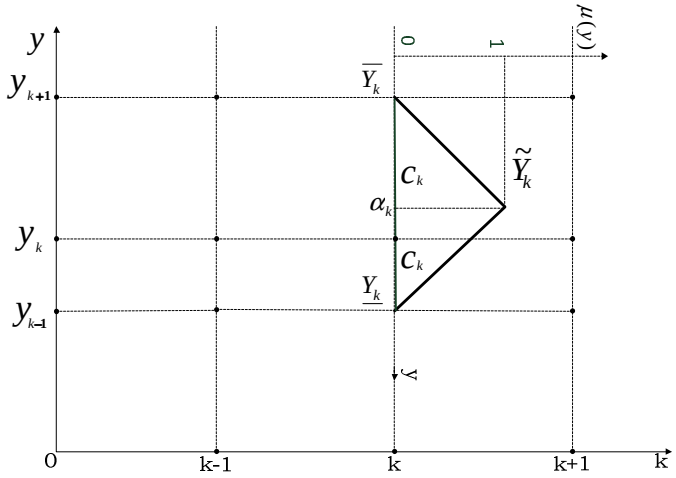
4 Fuzzy Linear Regression Model Identification

To define the type of fuzzy regression model we use the verse in which the input variables $x_{i,j}^0$ are mentioned as crisp numbers and observed values $\tilde{Y}_k^0$ as the triangular fuzzy numbers, respectively. Thus, let us consider the fuzzy number $\tilde{Y}^*$ as estimate and fuzzy number $\tilde{Y}^0$ as observed value of model output variable respectively. The fuzziness Δy_k of observed value $\tilde{Y}_k^0$ at the discrete time k we can determine using the observed values at the time (k+1) and (k-1), respectively (see Figure 2). It means, the fuzzy number $\tilde{Y}_k^0$ is mentioned of the symmetrical triangular type. The fuzziness is done using estimation by formula [1]

$$\Delta y_k = \frac{1}{2}\left|y_{k+1} - y_{k-1}\right| \quad (6)$$

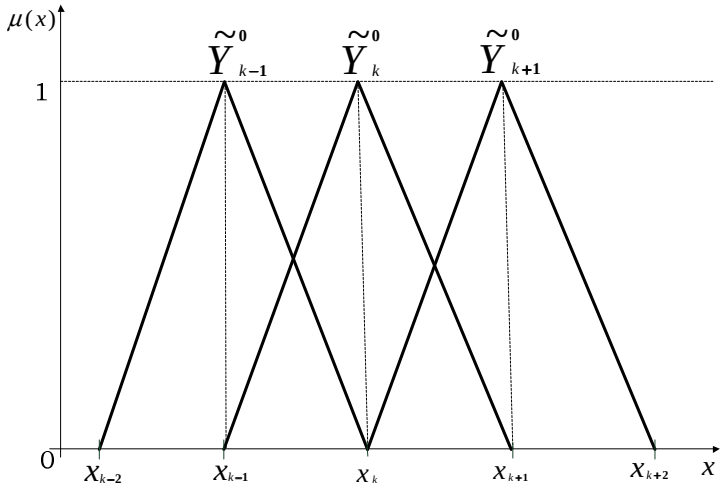
where all values in interval $\langle y_{k-1}, y_{k+1}\rangle$ are expected as members of membership function $\tilde{Y}_k^0$ carrier (Figure 2).

Figure 2: Input variable fuzzification



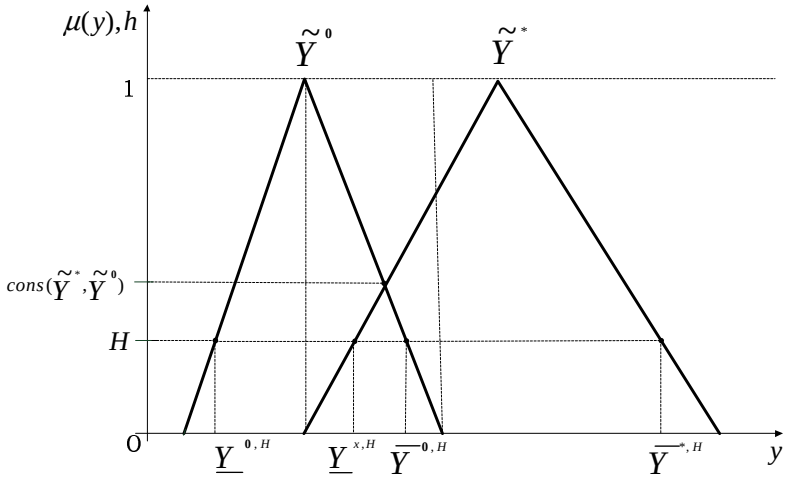
In this way, the membership function of fuzzy numbers of observed values $\tilde{Y}_k^0$ are mutually situated as we can see in Figure 3

Figure 3: Input variable fuzzy values



Fitness of linear regression fuzzy model to the given data is measured through the Bass-Kwakernaak´s index H – see Figure 4 [1].

Figure 4: Adequacy of linear regression fuzzy model



We assume the good estimation of output value under the condition is fulfilled

$$\max_{y}\left\{\mu_{\tilde{Y}^0}(y) \wedge \mu_{\tilde{Y}^*}(y)\right\} = Cons(\tilde{Y}^0, \tilde{Y}^*) \geq H \tag{7}$$

The fitness of estimated value to sampled value is done using α – cut and α – level set at the fitness h = H (see Fig.C)

$$Y_j^{0,H} = \left\lfloor \underline{Y}_j^{0,H}, \overline{Y}_j^{0,H} \right\rfloor \tag{8}$$

$$Y_j^{*,H} = \left\lfloor \underline{Y}_j^{*,H}, \overline{Y}_j^{*,H} \right\rfloor$$

The relation (3) is satisfied under the condition

$$\underline{Y}_j^{*} \leq \overline{Y}_j^{0}, \quad j = 1, 2, \ldots, m \tag{9}$$

$$\underline{Y}_j^{0} \leq \overline{Y}_j^{*}, \quad j = 1, 2, \ldots, m$$

Boundary of intervals $Y_j^{*,H}$, $j = 1, 2, \ldots, m$ we can express

$$\underline{Y}_j^{*,H} = -(1-H)\sum_{i=1}^{n} c_{ij}\left|x_{ij}\right| + \alpha^T x_j \tag{10}$$

$$\overline{Y}_j^{*,H} = (1-H)\sum_{i=1}^{n} c_{ij}\left|x_{ij}\right| + \alpha^T x_j$$

Next we can set the optimization problem

a) minimization of fuzzy model vagueness

$$\min J_j = \min \sum_{i=1}^{n} c_{ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m \tag{11}$$

b) subject to

$$\alpha^T x_j + (1-H)\sum_{i=1}^{n} c_{ij}|x_i| \geq y_j^0 + (1-H)\Delta y_j^0 \quad (12)$$

$$-\alpha^T x_j + (1-H)\sum_{i=1}^{n} c_{ij}|x_i| \geq -y_j^0 + (1-H)\Delta y_j^0$$

$$c_{ij} \geq 0$$

To solve the minimization problem under condition many authors use the linear programing method [1], [8]. Nevertheless, this problem is solved using genetic algorithm method in presented paper.

5 Genetic Algorithms Utilization

As mentioned before, the classical method of linear programming used for the identification of fuzzy regression coefficients [1] was substituted by using a genetic algorithm (GA) [9]. Mainly, the reason is that the authors are oriented to use unconventional methods of artificial intelligence in order to prove their quality and efficiency in solving complex tasks. Genetic algorithms are a representative of evolutionary methods; their higher computational complexity is nowadays eliminated by high-performance computing. They are widely used in the search for optimal solutions. They can be well used for the identification of fuzzy regression models where they deal with the task of finding the optimal fuzzy regression coefficients as triangular fuzzy numbers.

The identification of fuzzy regression coefficients – fuzzy numbers $\tilde{A}_0, \tilde{A}_1, ..., \tilde{A}_n$ - was divided into two tasks

(1) the identification of the mean value (core) α_i of fuzzy number $\tilde{A}_i$ and

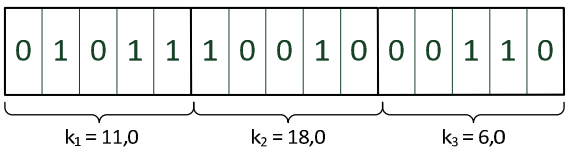
(2) the identification of c_i as a half of the width of the carrier bearing $\tilde{A}_i = \{\alpha_i, c_i\}$.

The tasks are solved by using the genetic algorithm in series. First the identification of α_i and then the identification of c_i are done.

As it was mentioned before, the genetic algorithm (GA) is an unconventional optimisation method, which is used for minimization of the target optimization function (fitness function). It is used instead of conventional methods, such as the linear programming method.

GA is a seeking procedure which looks for the best solutions according to the fitness function based on the processes observed in Nature, on the principle of natural selection and genetic laws, i.e. selection, crossover and mutation. The basis of GA is to use a character string, also called a chromosome, in which parameters of an optimized model are stored. An example of a chromosome which is composed of three parameters k1, k2 and k3 expressed by three 5-bit binary words is shown in Figure 5.

Figure 5: Binary Coded Parameters in Structure of Chromosome



Individual bits represent the string of chromosome genes; at the particular optimization step their specific values represent binary codes of three parameters of the model. Each chromosome is evaluated by the size of its fitness function, the value of which determines the distance of a solution (which is represented by a particular chromosome) from the optimal solution.

The set of evaluated n-chromosomes represents one population, the best individuals (solutions) of which are genetic operations of selection and are picked out for follow-up populations. Selected individuals are subjected to genetic operations of crossover, in which two individuals (parents) interchange gene circuits and generate two new chromosomes - offspring with different combinations of k1, k2 and k3. The descendants, who were generated this way, then form a new population where individuals (solutions) appear to have better characteristics (better fitness function value) than the best individual in the population of parents. Then, an appropriate follow-up offspring population is created (solution step, iteration) and the genetic crossing procedure is repeated. Good convergence for finding an optimal individual (solution) is supported by a genetic operation - mutation. The features of genetic operations as selection, crossover and mutation are defined by setting the internal parameters of the genetic algorithm in the way that the convergence of a solution to optimum is favorable.

The procedure of the genetic algorithm is usually finished by a solution step (population) in which the values of the fitness function of the best current individual and the best individual in the last step vary less than the specified limit (stop-criterion). As an optimal solution is then determined the best chromosome of the last population. Corresponding (coded) parameter values are used in the optimal model.

The main tasks while designing a genetic algorithm are the method of encoding the optimized parameters to a chromosome string and the definition of its fitness function.

Optimization of the fuzzy linear regression model is a two-step process when two genetic algorithms, designated G1 and G2, are used.

For the identification of the mean value (core) α_i of fuzzy number $\tilde{A}_i$ the minimization of the fitness function is defined in the form

$$\min J_1 = \min \frac{1}{J} \sum_{j=1}^{J} [\tilde{Y}^0(x_j) - \tilde{Y}^*(x_j)]^2 \tag{13}$$

and the genetic algorithm GA1 is used. For the identification of c_i as a half of the width of the carrier bearing $\tilde{A}_i$ the minimization of the fitness function is defined in the form

$$\min J_2 = \min \sum_{i=1}^{n} |c_i| \tag{14}$$

and the genetic algorithm GA2 with three constraints ( ) is used. Minimization of the fitness function J2 is based on the previous identification of the role of the mean value (core) αi and uses the already identified values of αi for determining the width of the carrier bearing αi.

The genetic procedures GA1 and GA2 respectively are provided using the specialized Genetic Algorithm and Optimtoolbox of the program system MATLAB [10].

6 Numerical Examples

To ilustrate the shape of multivalue dependency possibility area the artificial two dimensional linear function in the form

$$y = 48x_1 - 120x_2 + 20 \tag{15}$$

was chose and identified. The set of observed value $\tilde{Y}^0$ with ten members using ( ) was created. For creating the set of $\tilde{Y}^0$ the values of x_1 and x_2 were chosen randomly from the standard uniform distribution on the open interval (0; 1) but multiplied by random integer.

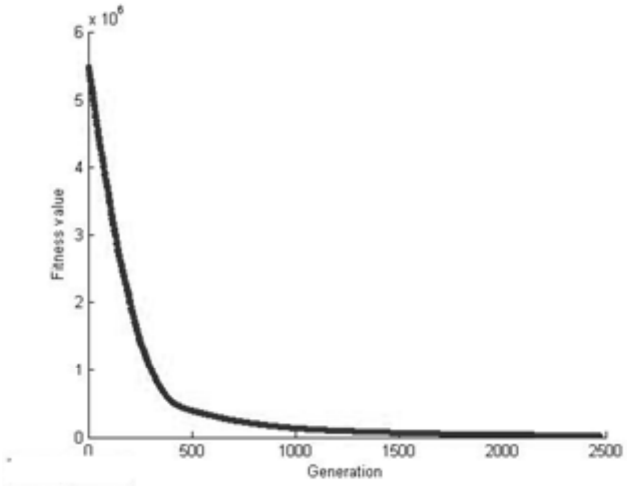
Then the minimization of fittness function J_1 ( ) by embedded function of genetic algorithm in Optimtool in Matlab environment [ ] was used. The parameters of GA were elected in the next form: population type - double vector, population size – 100, scaling function – rank, selection - stochastic uniform, mutation function -

constraint dependent, crossover function – scattered, migration – and stop criterion - no changes in fitness function.

The shape of convergence of values of minimization of fittness function J_1 is depicted in Figure 6. The outputs of the minimization by described GA are the estimated values of the mean values (cores) α_0, α_1 and α_2 of $\tilde{A}_0, \tilde{A}_1$ and $\tilde{A}_2$.

The next step was to determine the c_0, c_1 and c_2 of $\tilde{A}_0, \tilde{A}_1$ and $\tilde{A}_2$. For this task the minimization of fitness function (14) GA was used with the same parameters as in task of determining of α_i.

Figure 6: Course of GA convergence



As we now have the complete information to assemble the estimated fuzzy numbers $\tilde{A}_0, \tilde{A}_1$ and $\tilde{A}_2$ we can define

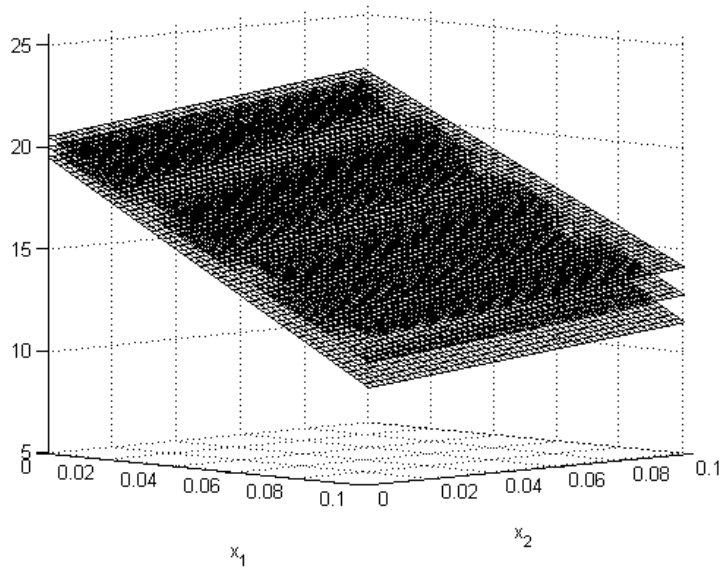
$$\begin{aligned} \tilde{Y}(y, \Delta y) &= \tilde{A}_0(\alpha_0, c_0) + \tilde{A}_1(\alpha_1, c_1)x_1 + \tilde{A}_2(\alpha_2, c_2)x_2 \\ y &= \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \\ \Delta y &= c_0 + c_1 x_1 + c_2 x_2 \end{aligned} \quad (15)$$

With knowledge of ( ) we are able to create the surfaces, which are defined as the upper and lower boundary

$$\begin{aligned} \underline{Y} &= y - \Delta y \\ \overline{Y} &= y + \Delta y \end{aligned} \quad (16)$$

The area between the created lower and upper surface boundary could be called possibility area. The regression analysis of function (15), i.e. its linear area and possibility area we can see in Figure 7.

Figure 7: Possibility area of two-dimensional fuzzy regression Function



Next the more practice and useful example is presented, namely task of the time series fuzzy regression analysis. The time-trend and seasonal cycles including their possibility areas are calculated and expressed, namely time-development of unemployment [11].

Unemployment behaves in the next way: when production declines, unemployment rises and vice versa (due to fluctuation - decrease/ increase of demand). Thus, unemployment secondarily shows a seasonal character.

The results of unemployment time series analysis are shown in the form of the fuzzy regression models of time series of Figure 8 and Figure 9. The figures represent their fuzzy trend and fuzzy seasonal cycle together with their possibility areas.

Figure 8: Amount of Unemployment - Fuzzy Linear Regression Function

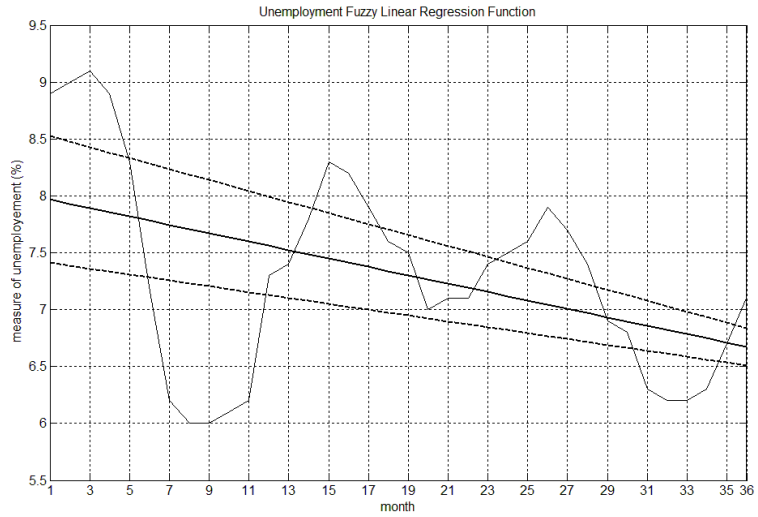
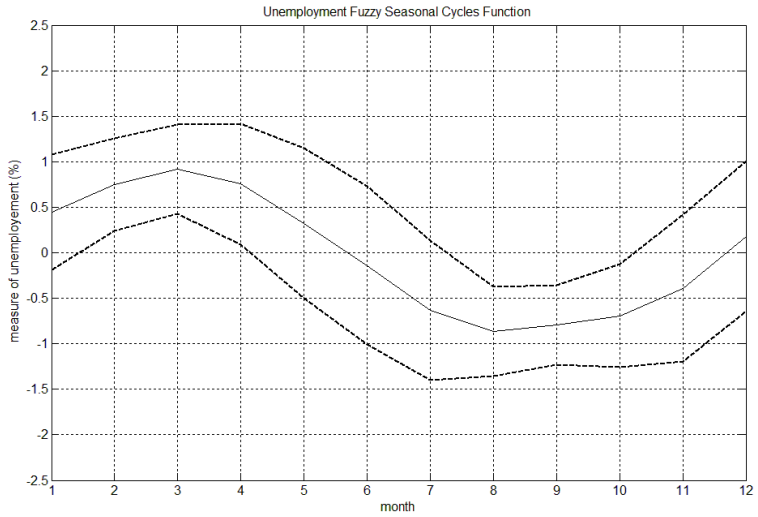


Figure 9: Amount of Unemployment - Fuzzy Seasonal Cycles Function



7 Conclusions

In classical regression, we assume that the relationship between dependent variables and independent variables of the model is well-defined and sharp. In the real world, however, hampered by the fact that this relationship is more or less non-specific and vague. The suitable theoretical background for abstract formalization of vague phenomenon of complex systems is fuzzy set theory. In the paper are defined vague data as specialized fuzzy sets - fuzzy numbers and there is a fuzzy linear regression model as a fuzzy function with fuzzy numbers as vague parameters. To identify the fuzzy coefficients of model the genetic algorithm is used. The linear approximation of vague function together with its possibility area are analytically and graphically expressed. Two numerical examples are presented namely fuzzification of two dimensional linear regression function and practical time-series fuzzy linear analysis. In the next Part II of contribution the non-linear fuzzy regression modelling [12] will be presented.

Acknowledgements

This work has been supported by Project GAČR P403-12-1811: Unconventional Managerial Decision Making Methods Development in Enterprise Economics and Public Economy.

References

[1] KACPRZYK,J. FEDRIZZI,M. (Ed.). *Fuzzy Regression Analysis. Studies in Fuzziness and Soft Computing,* ISBN-13: 978-3790805918. Publisher: Physica-Verlag HD, 1992.
[2] BARDOSSY,A. Note on fuzzy regression. *Fuzzy Sets and Systems*, volume 37, pages 65-75, 1990.
[3] BUCKLEY, J.J. JOWERS, L.J. Fuzzy Linear Regression I. *Studies in Fuzziness and Soft Computing*, volume 22, ISBN 978-3-540-76289-8. Springer, 2008.
[4] HESMATY,B, KANDEL,A. Fuzzy Linear Regression and Its Application to Forecasting in Uncertain Environment. *Fuzzy Sets and Systems*, volume 15, pages 159-171.
[5] SHAPIRO,A.F. Fuzzy regression models. [online]. [cit. 2013-0$-10]. Dostupné z: http://www.soa.org/library/research/actuarial-research-clearing-house/2006/january/arch06v40n1-ii.pdf
[6] TANAKA,H.,UEJIMA,S.,ASAI,K. Linear regression analysis with fuzzy model. IEEE Transactions and Systems, Man and Cybernetics, 12:, 1982.
[7] NOVÁK,V.,PERFILIEVA,I.,MOČKOŘ,J. *Mathematical Principles of Fuzzy Logic.* Kluwer, Boston. 1999. ISBN 0-7923-8595-0.
[8] CETINTAV,B. ZDEMIR,F. *LP Methods for Fuzzy Regression and a New Approach* E.Krause, Ed. In Synergies of Soft Computing and Statistics for Intelligent Data Analysis, volume 22, ISBN 978-3-642-33041-4. Springer, 2013.
[9] C.R.REEVES, J.E.ROWE. Genetic Algorithms: Principles and Perspectives Kluwer Academic Publishers, New York. 2002,ISBN 1-4020-7240-6
[10] MATLAB - *The MathWorks-MATLAB and Simulink for Technical Computing.* [online]. [cit. 2013-07-10]. Dostupné z: http://www.mahworks.com.
[11] POKORNÝ,M.,POSPÍŠIL,R.,NOWAKOVÁ,J. On Macroeconomic Values Investigation Using Fuzzy Linear Regression Analysis. *E+M Ekonomika-Management*.TU Liberec (podáno k tisku)
[12] POKORNÝ,M. *Fuzzy nelinearní regresní analyza*. Doctoral Thesis (in Czech) VUT Brno, FEL, Brno, 1993.