The Method for Linear Regression Models Constructing Based on the Sharing of Measured Data and Expert Assessments

Vladimir Gvozdev
dept. Computer Science and Robotics
Ufa State Aviation Technical University
Ufa, Russia
wega55@mail.ru

Oxana Bezhaeva
dept. Computer Science and Robotics
Ufa State Aviation Technical University
Ufa, Russia
obezhaeva@gmail.com

Dinara Akhmetova
dept. Computer Science and Robotics
Ufa State Aviation Technical University
Ufa, Russia
dinara.akhmetova.92@gmail.com

Alexander Levkov
dept. Computer Science and Robotics
Ufa State Aviation Technical University
Ufa, Russia
projector@gmail.ru

***Abstract*—The paper discusses the classical approaches to the construction of regression models on the basis of independent and dependent random variables distribution laws. Regression dependencies are one of the main tools for constructing empirical descriptive models, at present moment the theoretical apparatus for constructing regression models has been developed. In the paper analysis of approaches to the construction of the one-dimensional regression dependencies is carried out The analysis allows to conclude that the known methods for constructing of regression models focused on the processing of jointly observed measured data. The authors of the paper propose the method for constructing linear based on the sharing of measured data and expert assessments. Methodological basis for ensuring the comparability of measured data and expert analysis is to convert it to a form of random variables distribution laws. Transformation of expert assessments to the form of continuous random variables distribution laws allows to developed the formal procedure for the construction of linear regression based on the sharing of measured data and expert assessments. Due to the proposed method for constructing of linear regression dependencies in the paper is shown an example of the assessments calculating of paired correlation coefficient and parameters of linear regression dependencies**

Keywords*—*regression models, jointly observed measured data, the sharing of measured data and expert assessments, random variables distribution law, linear regression dependencies, paired correlation coefficient

I. Introduction

Regression dependencies are one of the main tools for constructing empirical descriptive models of inertialess objects. At present moment the theoretical apparatus for constructing regression models has been developed, on the basis of which software-implemented tools have been also developed. At the same time, linear models occupy a special place in the problem of regression analysis. This is due to the fact that it creates a basis for studying the main properties of objects in case of a small number of measurement data.

The information basis for constructing of regression dependences is made the table of jointly observed values of independent and dependent random variables. At the solution of practical tasks, it is necessary to come up with a situation when the formation of such tables meets both organizational, and technical difficulties. Examples of this are the situations related to the solving of the problems of the of territorial systems environmental safety management, when different indicators characterizing the impact on objects and responses are accumulated in different departments, regulated by various regulatory documents. In these conditions the lack of measuring data has to be compensated by expert assessments. In this regard, the development of methods for constructing the linear regression dependencies based on measurement data and expert assessments has the scientific and practical interest.

II. Analysis of Approaches to the Construction of the One-Dimensional Regression Dependencies

The classical approach to the construction the of regression models is based on the processing of correlation tables (tables of jointly observed values) of independent and dependent random variables.

$$A^{(0)}: \{x^*, y^*\}_1^N \underset{K}{\vec{\;}}\; \hat{y} = \varphi(\vec{Q}, x) \qquad (1)$$

Here $\{x^*, y^*\}$ is the set of jointly observed values of the metric characteristics of the values of independent x and dependent y random variables (table of jointly observed values);

$A^{(0)}$ - operator for calculating of the vector parameters $\vec{Q}$ of the regression model as a result of solving an optimization problem, in which K is a criterion characterizing the discrepancy between the measured $\{y^*\}_1^N$ and model data $\left\{\widehat{y(x^*)}\right\}_1^N$ (for example, the method of the smallest squares based on use of the Euclidean metrics as K [1]; the method of the smallest modules [2] based on use of modules sums of a divergence of measured data K).

In [3], a method for constructing non-parametric one-dimensional regression dependencies is described, based on an analysis of the distribution laws of the independent $F(x)$ and dependent $F(y)$ variables:

$$A^{(1)}: \{F(x), F(y)\} \rightarrow \varphi(x) \qquad (2)$$

The basis of the method is the solution of the inverse problem of determining the distribution law of the function of a random argument.

The solution of the direct task is based on the relation:

$$F(y) = \int_{y<\varphi(x)} f(x)\, dx \qquad (3)$$

where $f(x)$ – distribution density of a random argument;

$F(y)$ – distribution law of random argument function;

$y = \varphi(x)$ – interconnection characteristic of independent and dependent random variables.

The method based on the solution inverse task of (3) allows to construct nonparametric regression function in case of absence of the jointly observed values table. The initial data for the implementation of the method are sample data $\{x\}_1^N$, $\{y\}_1^M$ (in general case with different volume) independent and dependent random variables.

Restriction of the method is, first, need of a priori justification of existence of monotonous dependence between X and Y, secondly, the fact that parametrical assessments are constructed as strict on an interval functional dependences.

In [4] described the method for identifying objects with unobserved input. The method is based on the implementation of a modified smallest squares method, which allows to construct linear regression dependencies in the case when there are measured data $\{y\}_1^N$ {y}, and an independent random variable is represented by the distribution law $F(x)$. The method uses the apparatus of order statistics. In fact, the method described in [4] represents a special case of solving the inverse task of determining the distribution law function of the random argument.

In general, the performed analysis allows us to conclude that the known methods for constructing regression dependencies are focused on the processing of measurement data. Not identified the methods focused on the sharing of measured data and expert assessments.

III. Assessments construction of the random variables distribution laws based on measured data and expert assessments

When solving practical problems faced with a situation where there are measured data $\{y^*\}_1^N$, while the measured data characterizing X are absent, but there are expert assessments of the possible values range $x\epsilon[l_x, h_x]$, either in the form of the expected value $\widehat{M}(x)$, or in the form of both the expected value and the possible values range $\{\widehat{M}(x), [\,l_x, h_x\,]\}$.

The proposed approach is based on the transformation of expert assessments to the form of the random variable distribution law of the $\hat{F}(x)$, which allows in the future to construct a linear regression equation, based on the solution of the inverse task of determining the distribution law of a random argument function (2).

In [5-13] is described the method for construction of assessments of the random variables distribution laws based on various forms of expert assessments. The conceptual basis of the method is made the well-known entropy maximization principle [14, 15].

It is shown that if expert assessment is presented in the form of possible values range of a random variable $x\epsilon[a_x, b_x]$ it is necessary to choose the uniform distribution law as assessment:

$$\hat{F}^{(1)}(x) = \frac{x}{b_x - a_x} \qquad (4)$$

If the expert assessment is represented as the expected value of the random variable $M[x]$, in addition, a one-sided restriction is imposed on the possible values interval of the random variable $x\epsilon(a_x, \infty)$, it is necessary to choose the an indicative distribution law as assessment.

$$\hat{F}^{(2)}(x) = 1 - e^{-\lambda x} \qquad (5)$$

where $\lambda = (M[x] - a_x)^{-1}$.

If the expert assessment is represented as the expected value of the random variable $M[x]$ and possible values range of the random variable $x\epsilon[l_x, h_x]$ and it is known that, at the interval boundaries the values of the distribution density $f(l_x) = f(h_x) = 0$, then it is necessary to choose the triangular distribution as assessment.

$$f(x) = \begin{cases} \frac{2(x-l_x)}{(h_x-l_x)(M[x]-l_x)}, \ when\ l_x \le x \le M[x] \\ \frac{2(h_x-x)}{(h_x-l_x)(h_x-M[x])}\ , when\ M[x] \le t \le h_x \end{cases} \qquad (6)$$

Here $M[x]$ is the distribution density mode, characterizing the expected value of X.

If the expert assessment is represented as the expected value of the random variable $M[x]$ and possible values range of the random variable $[l_x, h_x]$, and restrictions are not imposed on value of distribution density $f(x)$ on the interval borders, it is necessary to accept as assessment:

$$F(x) = \int_{a_x}^{x} e^{\mu_0 + \mu_1 \tau} d\tau \qquad (7)$$

And parameters μ_0, μ_1 are found in the result of the solution of the equations system:

$$\begin{cases} \int_{a_x}^{b_x} e^{\mu_0 + \mu_1 \tau} d\tau = 1 \\ \int_{a_x}^{b_x} x e^{\mu_0 + \mu_1 \tau} d\tau = M[x] \end{cases} \qquad (8)$$

Thus, the transformation of expert assessment to the type of the continuous random variable distribution law makes it possible to create continuous regression dependences according to the scheme (2) on the basis of sharing measured data and expert assessment. Regression dependences represent strictly on interval functional dependences continuous on an interval, and borders of an interval are defined by area of permissible values of an independent random variable.

IV. Construction of One-Dimensional Regression Dependencies Based on Measured Data and Expert Assessments

Solution of the inverse task of estimating a distribution law of random argument function (2) provides, with distinct types of $\hat{F}(x)$, $\hat{F}(y)$, nonlinear, strict on an interval of acceptable values X, assessment of nonparametric regression dependence. The linearization $\hat{y} = \varphi(x)$ generally comes down to the solution of the task:

$$A^{(4)}: \{\varphi(x_i), y^{(l)}(x_i)\} \underset{K}{\rightarrow} min \qquad (9)$$

Here $\varphi(x_i)$ - the value of non-parametric dependence in points x_i, ; $y^{(l)}(x_i)$, $-$ the value of linear dependence in the same points. K - optimization criterion

$A^{(4)}$ - operator, which implements the decisions of the optimization problem.

If as K is accepted the Euclidean metrics, then determination of linear dependence parameters $\hat{y} = a + b_x$ is used the method of the smallest squares. The parameters of the linear model in this case are determined by the relations [16].

$$a = m_y - bm_x; b = r_{xy}\frac{\sigma_y}{\sigma_x} \qquad (10)$$

where m_x, m_y - mathematical expectation X and Y respectively; $\sigma_x, \sigma_y -$ standard deviation X and Y respectively; $r_{xy} -$ pair correlation coefficient between X and Y.

An alternative approach to determining the parameters of linear dependence is based on the use the Dominance metric as K. In this case, the determination of the parameters is reduced to solving the optimization task:

$$\max_{x_i} |\varphi(x_i), y^{(l)}(x_i)| \to min \qquad (11)$$

On the figure 1 the distribution laws of independent and dependent random variables are shown.

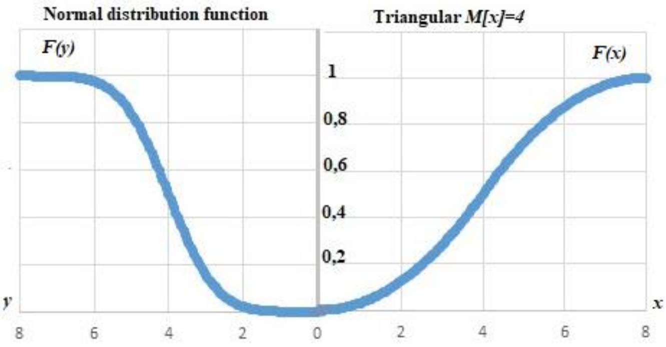

Fig 1. The distribution laws of independent and dependent random variables

Figure 2 shows the results of estimating the regression dependencies obtained as a result of solving (2) in the case when $\hat{F}(x)$ corresponded to the (6) and $\hat{F}(y)$ corresponds to the normal distribution law (Figure 1). In the example, the theoretical dependence $\varphi(x)$ is linear [17].

The conducted researches showed that use of the Euclidean metrics allows to receive more accurate linear approximation of nonparametric regression dependences strict on interval, than use of domination metrics.

When determining the linear dependence of the parameter values by solving the optimization problem (11) has been used known optimization method without calculating derivatives – the deformable polyhedron method [18].

The focus of the determination of the linear dependencies parameters based on the Euclidean metric is the calculation of the paired correlation coefficients (10).

Transformation of expert assessments to the form of random variable distribution law (4) - (6) makes it possible to propose the following procedure for estimating of the linear dependence parameters, excluding need of the solution of an optimizing task (9).

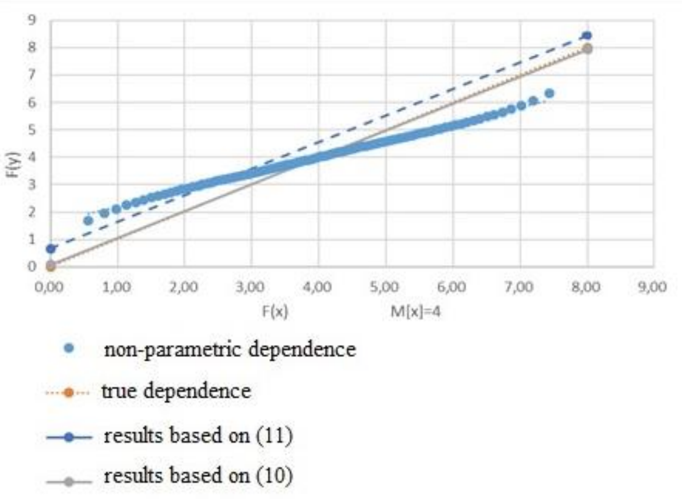

Fig 2. The results of the approximation $\boldsymbol{\varphi(x)}$

Taking into account that the basis for calculating of the pair correlation coefficients is a table of jointly observed values of independent random variables $\{x^*, y^*\}_1^N$ $*\}$ construct this table using $\hat{F}(x)$, $\hat{F}(y)$. For this, generating for each q-th realization of z^q is expected random numbers:

$$x^{*(q)} = F_x^{\ -1}(z^{(q)}); y^{*(q)} = F_y^{\ -1}(z^{(q)}) \quad (12)$$

z^q – uniformly distributed random variable $z^q \in [0,1]$.

Here $F_x^{\ -1}(z^{(q)})$, $F_y^{\ -1}(z^{(q)})$ – inverse functions of the distribution laws $F(x)$, $F(y)$ accordingly.

Based on the generated table $\{x^*, y^*\}$, is not difficult to calculate the values $r_{xy}, m_x, m_y, \sigma_x, \sigma_y$ [9].

Figure 3 shows the graphical scheme of the construction of the table of jointly observed values $\{x^*, y^*\}_1^N$.

In table 1, as an example, assessments of paired correlation coefficient r_{xy} are given, as well as parameters of linear regression dependencies:

- in the first case, when the theoretical distribution functions $F(x)$, $F(y)$ were normal;
- in the second case when the theoretical distribution functions $F(x)$, $F(y)$ were exponential with the same $\sigma_x = \sigma_y = 1$ (in this case, the theoretical values of the

pair correlation coefficient and parameters of the regression dependence are $r_{xy} = 1; . a = 0; b = 1$).

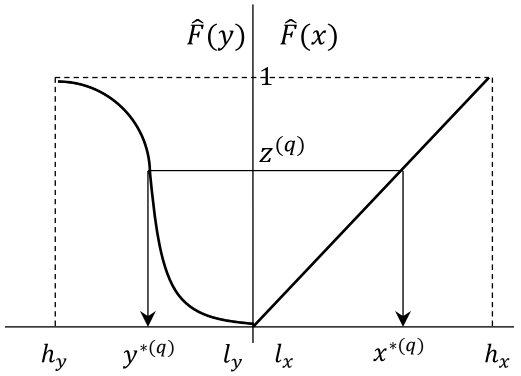

Fig 3. graphical scheme of the construction of the table of jointly observed values $\{\boldsymbol{x}^*, \boldsymbol{y}^*\}_1^N$

In fact, the assessments $\hat{F}(x)$ were specified either as an interval $[l_x, h_x]$, or as an expected value $M[x]$ and an interval $[l_x, h_x]$ $\{M[x], [l_x, h_x]\}$.

In Table 1 the assessments of paired correlation coefficient and parameters of linear regression dependencies are shown.

TABLE I. TABLE 1 ASSESSMENTS OF PAIRED CORRELATION COEFFICIENT $\boldsymbol{r_{xy}}$ AND PARAMETERS OF LINEAR REGRESSION DEPENDENCIES

Type and parameters of theoretical distribution law $\boldsymbol{F(x)}$	**Assess-ment type $\boldsymbol{\hat{F}(x)}$**	**Type and parameters $\boldsymbol{F(y)}$**	**Assess-ment $\boldsymbol{r_{xy}}$**	**Parameters of linear model**	
				a	***b***
Normal $M[x] = 4$ σ_x=1	[0, 8]	Normal $M[y] = 4$ σ_y=1	0,98	2,34	0,98
Normal $M[x] = 4$ $\sigma_x = 1$	{4; [0, 8]}	Normal $M[y] = 4$ σ_y=1	0,99	0,03	0,99
Exponential $\lambda = 1$	[0, 8]	Exponential $\lambda = 1$	0,89	0,44	0,89
Exponential $\lambda = 1$	{4; [0, 8]}	Exponential $\lambda = 1$	0,95	0,36	0,95

The described procedure for constructing linear regression dependencies can be used in the case when the type and parameters of the distribution laws $F(x)$, $F(y)$ are apriority unknown. Moreover, in the general case, the types $F(x)$ and $F(y)$ do not coincide. In this case, the results will coincide with the assessments obtained by the described method of linearizing nonlinear strict on the interval functional dependencies.

V. CONCLUSION

Transformation of expert assessments to the form of continuous random variables distribution laws allows to developed the formal procedure for the construction of linear regression based on the sharing of measured data and expert assessments.

The main results of the research are:

- Analysis of approaches to the construction of the one-dimensional regression dependencies. The analysis allows to conclude that the known methods for constructing of regression models focused on the processing of jointly observed measured data only.
- The authors of the paper propose the method for constructing linear regression dependencies based on the sharing of measured data and expert assessments. Methodological basis for ensuring the comparability of measured data and expert analysis is to convert it to a form of random variables distribution laws.
- Due to the proposed method for constructing linear regression dependencies is shown an example of calculating the assessments of paired correlation coefficient and parameters of linear regression dependencies: in the case, when the theoretical distribution functions were normal and another case when the distribution functions were exponential.

ACKNOWLEDGMENTS

The research is supported by a grant № 19-08-00937 “Methods and models of intelligent decision support for the software projects management implemented in the environment of the industrial enterprises” from the Russian Foundation for Basic Research.

REFERENCES

[1] Yu. V Linnik. “The method of the smallest squares and basis of the mathematic-statistic theory of observations processing”. - 2nd edition, Moscow, 1962. - 349p. (in Russian)

[2] V. I. Mudrov, V. L. Kushko. “The method of the smallest modules”, Moscow, 1983, - 304 p. (in Russian)

[3] M.B. Guzairov, V.E. Gvozdev, B.G. Ilyasov, A.E. Kolodenkova “Statistical research of territorial systems”, Moscow, 2008 - 187 p. (in Russian)

[4] A. N. Efimov “Order statistics - its properties and applications”, Moscow, 1980 - 64p. (in Russian)

[5] V. E. Gvozdev, N. K. Crioni, B. G. Ilyasov., O. Ya. Bezhaeva, D. V. Blinova “Elements of system engineering. Technologies for the requirements construction to the hardware-software complexes based on expert-statistical methods”, Moscow, 2017, - 295 p. (in Russian)

[6] V. E. Gvozdev, N. I. Rovneiko “Numerical simulation of the requirements formation processes to software”. Vestnik USATU, 2012. – Vol. 16, № 3, P.153-159. (in Russian)

[7] E. R. Kozhanova., A. A. Zaharov “Application of wavelet analysis to determine the parameters of the normal distribution law” // Proceedings of International Conference on Actual Problems of Electron Devices Engineering (APEDE), vol. 2, 2014, pp. 280-283

[8] N. I. Yussupova, G. Kovács., M. Boyko, D. Bogdanova “Models and Methods for Quality Management Based on Artificial Intelligence Applications”, Acta Polytechnica Hungarica, 2016. - Vol. 13, № 3, P. 45-60

[9] V. E. Gvozdev, D. V. Blinova, D. R. Akhmetova “Statistical analysis of time of establishing steady phases of functioning of complex hardware-software systems” // Proc. of International Russian Automation Conference, RusAutoCon 2018, Russia, Sochi, 2018. P.1-5.

[10] V. E. Gvozdev, O. Y. Bezhaeva, A. S. Subhangulova “Analysis of linear relations objects random parameters on the basis of measurement data” // Proceedings of the 16th Workshop on Computer Science and Information Technologies/ Sheffield, England. 2014. Vol. 2. P. 13-15.

[11] M. B. Guzairov, V. E. Gvozdev, D. V. .Blinova, A. S. Davlieva “Control of component alterations according with the target efficiency of data processing and control system” // Proceedings of the International Conference Information Technology and

Nanotechnology. Session Data Science (Samara, Russia, 24-27 April, 2017): 2017. P. 11-16

[12] V. E. Gvozdev., L. R. Chernyakhovskaya., A. S. Davlieva “Decision support in management of hardware-software complex functional safety on the basis of ontological engineering” // International Russian Automation Conference, RusAutoCon, 2018. P.1-5.

[13] V. E. Gvozdev, M. B. Guzairov, D.V. Blinova, A. S. Davlieva “Control of component alterations according with the target efficiency of data processing and control system”, CEUR Workshop Proceedings, 2017, Vol. 1903, pp. 11-16

[14] A. M. Kagan, U. V. Linnik., C. P. Rao “Characterization problems of mathematical statistics”, Moscow, 1972. – 652p. (in Russian)

[15] Kuzin, L.T. “Basics of cybernetics”. Vol. 1, Moscow, 1973. – 503p. (in Russian)

[16] E. S. Ventzel “Probability Theory”, Textbook. for Universities. - 6th edition, Moscow, 1999, — 576 p. (in Russian)

[17] V. S. Pugachev “Probability Theory and Mathematical Statistics”. – Moscow, 2002.- 496 p. (in Russian)

[18] A.P. Dambrauskas “Simplex search”, Moscow, 1979. - 176 p. (in Russian)