Transformation of types of signs for a task of the regression analysis

IOP Conf. Series: Materials Science and Engineering **862** (2020) 052065 doi:10.1088/1757-899X/862/5/052065

Transformation of types of signs for a task of the regression analysis

D Z Narzullaev, B A Abdurakhmanov, A S Baydullaev, Sh T Ilyasov and K K Shadmanov

Tashkent pharmaceutical Institute, 45, Oybek street, Tashkent, 100015, Uzbekistan

E-mail: davr1960@mail.ru

Abstract. Today there are a large number of methods and analysis algorithms of data for solution of problems of image identification, automatic classification, component, correlation and regression analysis which in case of availability of data of the different type nature become inapplicable because they are intended only for processing of quantitative information. The approach to solution of this problem named conversion of types of signs for a problem of regression analysis is described in the offered article. At the same time conversion of type of signs is presented in the form of the separate task allowing to pass from not quantitative signs to quantitative ones and in further processing to use all range of traditional methods of the analysis of an original information. The offered algorithm is implemented in integrated environment of working out of the of Delphi 10 Seattle software.

1. Introduction

It is known that the processing stage of data is preceded by the scaling process, i.e. representation of values of object properties of an initial empirical set in the formalized numerical form which is a subject of studying of the theory of measurements [1, 2]. The application of the scaling process leads to obtaining from the empirical table of experimental data (TED), presented in the formalized form and ready to input to computer memory. From the theory of measurements it is known that for school values of each of types of signs only certain operations and conversions are allowed, i.e. a certain scale allows calculation of limited set of statistical characteristics. For example, for scales of names the only coordinate of position is the mode defining the value which is most often found in this set of numbers. In case of quantitative scales arithmetic operations are used, and a suitable measure of position of the center is the mean value. Generalizing the above stated we may make the following conclusion: as each scale provides calculation of a certain set of statistical characteristics, many classical methods of applied statistics are inapplicable for all volume of information of different type and are used only for processing of quantitative data.

The approach to the analysis of data of different type using methods of conversion of qualitative and nominal characters to quantitative is based on the principle of digitization. Digitization of not quantitative variables is carried out for the purpose of their further, most effective use along with the available quantitative signs in classical statistical models [3-5]. An important point in the procedure of digitization is the choice of a search criterion of numerical tags for gradation of not quantitative variables and methods of its optimization.

Among shortcomings of the existing methods of digitization preventing their wide circulation it is necessary to list the following: when using methods of digitization the relations inherent in initial empirical properties are not considered; the existing methods of digitization do not allow to consider rather fully influence of the quantitative variables entering to the description space along with qualitative and nominal (classification) characters. Therefore the quantitative variables either are excluded from consideration at all – at the same time the adequacy of the description of basic data is broken, or are come down to a nominal kind owing to what the informational content of signs is lost; as a result of unreasonable application of methods of digitization further data processing leads to obtaining empty information.

Work [4] in which process of digitization is considered as an independent research task is devoted to elimination of the above shortcomings. The procedure of transition from less "strong" scales to more "stronger" within the solution of a certain task of data analysis is called method of transformation of types of signs and is based on the following basic provisions: consecutive strengthening of a scale; at enrichment of a scale a priori assumptions of the researcher of concrete subject domain of the hidden relations which are not reflected in the initial system of signs are taken into account.

The main objective of this article is development of methods and algorithms of digitization of not quantitative signs at solution of a task of regression analysis.

2. Basic concepts and notation

Let initial information be presented in the form of the TED "an object sign (property)" by N*p size where N is number of objects lines, and p is number of signs columns in TED:

$$X=(X_1, X_2, ..., X_N), \quad (1)$$

in which $X_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}\right)$ – vector of values of the analyzed signs $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}$, registered on *i*-m the surveyed object. In the presence in the initial empirical matrix presented in the form (1) of the signs measured in various scales at the initial stage of the analysis of data the necessary condition is transformation of types of signs. Hereinafter under the concept of "transformation of types of signs" we will consider transition from various types of signs to the one kind: quantitative, qualitative or classification. Follows from above stated that scaling process is preceded to the stage of data processing, i.e. representation of values of properties of objects of an initial empirical set in the formalized numerical form. Application of process of scaling leads to receiving of experimental data of TED from the empirical table represented in the formalized form and ready to input to computer memory.

In practice the following types of scales are the most widespread: names, order, quantitative. In turn, in relation to the corresponding scales the following types of signs usually are distinguished [1]:

- Classification (nominal) signs are measured in a scale of names. This scale allows to break objects into groups (classes) uniform in property of this sign x. At the same time any streamlining of groups by sign x it is not entered. If all possible classes are known in advance, then they say about a categorized nominal scale (a scale of categories), and classes are called categories of sign x. Each value of sign x we can assign a certain code (gradation) – number which plays a role of a name of value of sign or the equivalence class corresponding to this value. The objects refer to one class of equivalence are considered as indiscernible on values of this sign.
- Qualitative (ordinal, serial) signs are measured in an order scale. This scale differs from classification in existence of the relation of a linear order for the set equivalence classes. The *x* sign, measured in an ordinal scale, allows to order (by *x* sign) objects on extent of manifestation of the property described by this sign, but does not give a quantitative measure for its expression.
- Quantitative signs are measured in scales of intervals, relations, differences. Such signs can take a continuous number of values from some range of permissible values, or permissible values can be numbered.

At the solution of real tasks in many cases results of researches are presented by the table of experimental data including at the same time values of signs of various type. We call such basic data polytypic, and the considered space – as space of polytypic signs. Scale values of each of types of signs allow only certain operations and transformations [2].

3. Statement of the problem

In [4] the generalized formula for the choice of imperative ordinal scales is offered as:

$$\frac{\bar{\mathbf{d}}^{\mathrm{T}}\mathbf{V}\bar{\mathbf{d}}}{\bar{\mathbf{d}}^{\mathrm{T}}\mathbf{Z}\bar{\mathbf{d}}} \rightarrow \min, \tag{2}$$

where $d \in P^g$, P^g - set of shifts of g of natural numbers. Concrete expressions for matrixes of V and Z depend on type of a solvable task. In [5] the algorithm of minimization (2) for a problem of recognition of images is offered. Now let's pass to the problem definition of the multiple regression analysis. Let initial TED be presented in the form (1), and there are signs of various types. Let's consider model of multiple linear regression for quantitative signs $x_1, x_2, \ldots, x_l$. Let the sign x_1 be selected as dependent variable and it needs to be evaluated linear combination of signs $x_2, \ldots, x_l$, considered as independent variables. Then multiple linear regression analysis [6-13] comes down to search of such hyperplane

$$x_1 = \beta_{12}x_2 + \beta_{13}x_3 + \ldots + \beta_{1l}x_l + \beta_{10},$$

for which the ratio

$$\sum_{i=1}^{N}(x_1^i - \beta_{12}x_2^i - \ldots - \beta_{1l}x_l^i - \beta_{10})^2 = \min \tag{3}$$

is fulfilled. Multiple linear regression used in the analysis of data as model of the description of one sign from TED by means of a linear combination of the others is connected with classical statistical model of regression as conditional average [6]. In this model value of signs are considered as selections by N volume for random variables with the arbitrary distribution having the final moments of the second order. In this case expression (3) determines the plane of mean square regression which is the best linear approach to the surface of regression by a method of the smallest squares

$$x_1 / x_2, \ldots, x_l = f(x_2, \ldots, x_l),$$

where $x_1 / x_2, \ldots, x_l$ - conditional average value for x_1 at fixed $x_2, \ldots, x_l$.

The regression model is considered as unconditional average also along with regression model as conditional average in classical statistics. Here a random variable is only the dependent variable, and independent variables are considered as determined. These two models of regression differ only in statistical properties of estimates of parameters of the plane of regression while computing aspects match both for classical statistical models, and considered in data analysis for the linear regression. Therefore all analytical expressions for definition of the plane of regression according to data of different type can be used also at creation of any of two classical models.

Within this article the task of type conversion of signs for multiple linear regression analysis allowing to consider influence of not quantitative signs from TED for obtaining more exact description of dependence of one of signs from the others is set.

4. Proposed approach

Let's determine parameters of the equation of regression [14] by elements of a covariation matrix

$$\underline{\Lambda} = \begin{bmatrix} \lambda_{11}\ \lambda_{12}\ \ldots\ \lambda_{1l} \\ \lambda_{21}\ \lambda_{22}\ \ldots\ \lambda_{22} \\ \ldots \\ \lambda_{l1}\ \lambda_{l2}\ \ldots\ \lambda_{l2} \end{bmatrix},$$

where λ_{ij} - covariations of *i*- and *j*- signs.

Similarly [4] we will designate through Λ matrix determinant $\underline{\Lambda}$, and through Λ_{ij} - additional of minor element λ_{ij} in determinant Λ. Then we may receive equality:

$$\Lambda_{11} = \Lambda_{11,l+1,l+1} \bar{a}^T \mathrm{S}' \bar{a}, \tag{4}$$

where

$$\boldsymbol{S}' = \boldsymbol{T} - \boldsymbol{M} \Lambda^{-1}_{1.1,l+1.l+1} \boldsymbol{M}^{\boldsymbol{T}},$$
$$\bar{a}^T \mathrm{S}' \bar{a} = \sigma^2_{l+1}(1 - r^2_{l+1.2\ldots l}). \tag{5}$$

We may show that [4]:

$$\sigma^2_{1,2,\ldots,l+1} = \frac{\Lambda}{\Lambda_{11}} = \sigma^2_{1,23\ldots l} \frac{\bar{a}^T \mathrm{S} \bar{a}}{\bar{a}^T \mathrm{S}' \bar{a}}. \tag{6}$$

The criterion of search of an imperative scale for a task of the regression analysis will take the form:

$$\frac{\bar{a}^T \mathrm{S} \bar{a}}{\bar{a}^T \mathrm{S}' \bar{a}} \to \min, \tag{7}$$

where $\bar{a} \in \mathrm{B}^{\mathrm{g}}$, where from the following equality follows:

$$\frac{\bar{a}^T \mathrm{S} \bar{a}}{\bar{a}^T \mathrm{S}' \bar{a}} = 1 - r^2_{1l+1.23,\ldots,l}. \tag{8}$$

The imperative scale determined for model of linear regression by criterion (7) provides the maximum value of private coefficient of correlation of quantitative sign x_1 and the digitized sign x^{ξ}_{l+1}.

Let's assume that as dependent variable not quantitative sign x^{ξ}_{l+1} is considered. Then the numerical tags setting for x_{l+1} imperative scale, have to minimize size $1 - r^2_{l+1.1\ldots l}$. We may show that in this case criterion of search of an imperative scale for not quantitative sign x_{l+1} will have the form:

$$\frac{\bar{a}^T \mathrm{S} \bar{a}}{\bar{a}^T \mathrm{T} \bar{a}} \to \min, \tag{9}$$

where $T = \left[t_{ij} \right],\ i,\ j = \overline{1, g},]$,

$$t_{ij} = \begin{cases} -\dfrac{n_i n_j}{N},\ i \neq j, \\ \dfrac{n_i}{N^2}(N - n_i),\ i = j, \end{cases}$$

$$\bar{a} \in \boldsymbol{B}^{\boldsymbol{g}}.$$

In case of search of integer numerical tags criteria (7) and (9) will take the form:

$$\frac{\bar{D}^T \mathrm{S} \bar{D}}{\bar{D}^T \mathrm{S}' \bar{D}} \to \min \text{ и } \frac{\bar{D}^T \mathrm{S} \bar{D}}{\bar{D}^T \mathrm{T} \bar{D}} \to \min, \tag{10}$$

where $\bar{D} \in \mathrm{P}^g$.

For optimization of criteria (10) the approach stated in [5] is used.

Thus, use of methods of transformation of types of signs for a task of the multiple linear regression analysis allows to consider influence of not quantitative signs from the table of experimental data for obtaining more exact description of dependence of one of signs from the others. At the same time the possibility of use of classification or qualitative sign as dependent is not excluded.

5. Conclusion

The following main results are received in the work:

- Methods and algorithms of type conversion of signs are developed for a problem of regression analysis;
- The search algorithm of imperative scales of an order of gradation of not quantitative signs for a problem of regression analysis is offered;
- The developed search algorithms of imperative scales of an order of gradation of not quantitative signs for a problem of regression analysis are implemented in integrated development environment of Delphi 10 Seattle software and included in structure of an analysis system of data of SITO-PC.

References

[1] Alexandrov V V and Gorsky N D 1983 *Algorithms and programs for the structural method of data processing* (Leningrad: Nauka)
[2] Pfanzagl I 1976 *Measurement theory* (Moscow: Mir)
[3] Did E 1985 *Data Analysis Methods* (Moscow: Finansi i statistika)
[4] Nikiforov A M and Fazilov Sh Kh 1988 *Methods and algorithms for converting feature types in data analysis tasks* (Tashkent: FAN)
[5] Narzullaev D Z Shadmanov K K and Ilyasov Sh T Converting Feature Types in Analysis of Different Types of Data *International Journal of Innovative Technology and Exploring Engineering* vol 9 issue 4 421-426 doi: 10.35940/ijitee.D1441.029420
[6] Ayvazyan S A, Yenyukov I S and Meshalkin L D 1985 *Applied statistics. Dependency research* (Moskow: Finansi i statistika)
[7] Ayvazyan S A, Mkhitaryan V S and Zekhin V A 2012 *Workshop on multidimensional statistical methods* (Moskow: Moscow State University of Economics, Statistics and Informatics)
[8] Lbov G S 1981 Methods for processing heterogeneous experimental data (Novosibirsk: Nauka)
[9] Lbov G S, Gerasimov M K and Polyakova G L 2010 A method of interval prediction based on logical regularities // *Proceedings of the IASTED International Conference on Automation,*

Control, and Information Technology - Control, Diagnostics, and Automation, ACIT-CDA
[10] Lbov G and Berikov V 2010 Construction of an event tree on the basis of expert knowledge and time series *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*
[11] Bing Z, Wei Z et al 2018 An Empirical study on Predicting Blood Pressure using Classification and Regression Trees *[J] IEEE Access* 99 1-1
[12] Wang S and Jia Sh 2019 Signature handwriting identification based on generative adversarial networks *Journal of Physics: Conference Series* **1187** 042047 doi:10.1088/1742-6596/1187/4/042047
[13] Fang J L and Wu W 2018 Research on Signature Verification Method Based on Discrete Fréchet Distance *IOP Conf. Series: Materials Science and Engineering* **359** 012003 doi:10.1088/1757-899X/359/1/012003
[14] Kramer G 1975 *Mathematical Statistics Methods* (Moskow: Mir)