On the method of step evaluation in construction descriptive models

T.E. Rodionova[1], G.R. Kadyrova[1]

[1]Ulyanovsk State Technical University, Severniy Venets St. 32, 432027, Ulyanovsk, Russia

Abstract

Mathematical regression models used to describe technical objects or process the considered. Assumptions violations of regression analysis appearing in different practical data processing, are discussed. The method of step evaluation allowing to overcome negative impact of multicollinearity effect is described. The models got by the analysis of laser and radio interferometric observations and data of physic-chemistry index of water source are cited. Received the ratings are compared with the results got when using the method of least squares and the step estimation method. The choice of the optimal model is made according to the criteria of minimum displacement. The possibility of applying the method of step evaluation to construct descriptive models is proved.

Keywords: descriptive models; multicollinearity; the method of least squares; regression analysis; methods of structural identification; step evaluation

1. Introduction

Let's consider a descriptive (parametrical) regression model applied for the description of relationships of cause and effect of the phenomenon. Let the mathematical model look like:

$$\mathbf{Y} = \eta(X, \beta) + \varepsilon \quad (1)$$

Where Y is a dependent variable; $\mathbf{X} = \left(x_0 x_1 \quad x_{p-1}\right)^{\mathrm{T}}$ - is a vector of independent variables; $\beta = \left(\beta_0 \beta_1 \quad \beta_{p-1}\right)^{T}$ is a vector of unknown parameters defined by the result of the experiment; ε is a vector of random errors. Variables X and Y, included into the model, are the result of a passive experiment, i.e. the measured or calculated values. Vector β in model (1) is supposed to non-changeable in time, i.e. a mathematical model is considered stationary according to parameters [1,13].

In practice, to estimate parameters of such mathematical models methods of regression analysis are used, in particular the method of least squares. Thus it's necessary to consider possible infringements of conditions of application of this method. The application of regression modelling in the task considered means research and selection of optimal methods to obtain the best linear estimates of parameters and check the effectiveness of the resulting model according to the relevant criteria [3,12].

We can identify the following violations of applying an ordinary method of least squares in solving practical problems. They are as follows:

- Models contain insignificant (noise) terms;
- The model parameters correlate with each other (multicollinearity effect);
- The residuals can also be further distorted by autocorrelation and other systematic errors.

In general, the choice of the way to adapt violations of conditions of regression analysis by the method of least squares depends on the type of model investigated [9, 14]. In this case, the object of attention is directly the parameters of the model, rather than the results of prediction. The ultimate goal of adaptation is the best linear estimates, e.g. evaluations not burdened by notable systematic and random errors. They can be such values, at least, in case of statistical significance and, what in most important, when parameters of the model are independent on each other. It's obvious that adaptation to the first violation mentioned by simply removing insignificant terms is difficult for a very simple reason: some of them can be interconnected with significant ones.

To overcome the effect of multicollinearity and reduce the number of insignificant terms in descriptive models it's proposed to use the step evaluation method.

2. The description the step evaluation method

According to the method the phased partitioning is carried out not by individual variables (like in step regression) but by their groups consistently formed as subsets of variables with insignificant pair correlation coefficients r_{ij}. That means that the groups are formed not by the degree of correlation with the sequentially formed by responses, but in the form of separate structures in almost orthogonal basis [2,4]. A brief description of the algorithm of the step estimation method:

1. The estimation valuation an original model using one of computing schemes of least squares method is calculated:

$$\Delta = (X^T X)^{-1} X^T, \quad (2)$$

Its covariance matrix is

$$D(\Delta) = (X^T X)^{-1} \sigma^2, \quad (3)$$

and different statistics, allowing to estimate the statistical value of each term and whole model, including values of t-statistics and coefficients of pair correlations r_{ij} are calculated.

2. The first subset of corrections Δ_1, that got insignificant values r_{ij} is formed using comparison of values r_{ij}.

3. The parameters of the orthogonal structure are estimated

$$Y = X_1\Delta_1 , \quad (4)$$

the first vector of residues is calculated

$$e_1 = Y_1 - \hat{Y}_1, \quad (5)$$

which is regarded as another response vector forming the next subset of corrections from the set of remained ones.

4. Parts 2 and 3 are repeated until are process of forming subsets $\Delta_1,\Delta_2,\ldots\Delta_k$ is finished.

To improve the quality of the estimates obtained, the orthogonal transformation of the Householder is included in the calculation scheme. In this case, numerical stability, characteristic for orthogonal transformations, is combined with flexibility, which makes it easy to adapt to the consistent accumulation of data, which is very important for solving large-dimensional problems. In addition, the requirement for computer memory is reduced, execution speed and accuracy are increased. The protection from "machine zeros" and overflow should be noted. With the help of the first strategy of this algorithm, an attempt was made to evaluate the interrelated regressors separately by evaluating them at different stages of this method (MSE1). As a drawback of such an algorithm, it can be noted that among the estimated parameters there are also insignificant statistics according to the Student's statistics. As a defect of such an algorithm, it can be noted that among the estimated parameters there are also insignificant statistics according to the Student's statistics.

The second strategy of this method including to the final model only those regressors that turned out to be significant according to the t-criterion at each stage of the work (MSE2). It's very similar to the step regression method, but due to the fact that the calculation is based on individual subsets, it's possible to estimate many times the parameters of the initial model – since the regressor insignificant at one stage may turn out to be significant in subsequent ones. This is very important for the problem of parametric estimation, where is necessary to obtain the most possible complete model. The defect of this strategy is the lack of analysis of the interdependence of the included parameters.

The third strategy is the combination of the first and second ones. The selection to the set of evaluated parameters is performed immediately according to two grounds; namely the significance and orthogonality (MSE3).

3. The description of the initial data and revealed violations of regression analysis assumptions

For approbation of this method of estimation of the parameters of a mathematical model the following data were used: the data on the laser location of the moon; VLBI observations of extragalactic sources; the results of physical and chemical control of drinking water.

The processed biennial radar data wear got by using to angle reflector from the spaceship "Appolo-15" in McDonald observatory (Texas, USA) from August 1971 to November 1973 (549 observations at all). The source data in the form of coefficients of the conditional equations were prepared by the staff of the Institute of Theoretical Astronomy of USSR Academy of Sciences.

The considered VLBI observations are 1262 conditional equations for determining 203 corrections of the constant theory of the orbital motion and rotation of the Earth. To these data have been added 4 coupling equations, were added to these data. They determine the equality of the parallel transfer of the earth's coordinate system and the rotation of the earth's and celestial coordinate systems to zero, as well as the constraints imposed on the basis vectors. The data for the calculations were prepared by Professor V.E. Zharov (The State Astronomical Institute named after P.K. Shternberg, Moscow State University) [6,7].

As a third example, the results of physico-chemical control of drinking water (responses $y_1 - y_7$) and water from water source (estimated parameters $x_1 - x_8$), used to clean water were considered [8,10]. The original file is a result of the parameters control during a year.

As a first problem in the data processing we can name the problem of the sufficiency of the observations scope. In the research we are dealing with the following situation: laser data on the Moon to determine 24 unknown corrections contain 549 conditional equations, e.g. they exceed the number of estimated amendments 22 times; according to radiointerferometric data on the Earth, we have the ratio of 203 unknown corrections and 1289 conditional equations (including 27 coupling equations), so the number of observations is only 6 times as many as the number of parameters; according to the water source 365 observations are available to determine 8 water parameters (the number of observations is 45 times as many as the number of parameters). In the regression analysis between the number of determined parameters p and the number of observations n must be satisfied, during the experiment the ratio n =5p÷15p.

Data research began with the analysis of the model obtained by the multiple regression method. The number of insignificant parameters of the model and the matrix of pair correlation coefficients were considered. For this purpose, the SPOR package was used, which makes it possible to obtain regression models and determine their quality measures [5,11,15].

The presence of abnormal observations in the sample can be considered as the second problem facing a researcher. In the considered initial data four abnormal observations were removed from the laser observation file of the Moon, 18 outliers were found in the file with VLBI observations, and in the data for the water source for different responses, the amount of emissions varies from 1 to 4.

The next problem is directly related to the matrix of its original data: among the arguments (variables) should not be linearly dependent ones. However, in practice, this assumption is not always observed. When this condition is violated, the linear functional or statistical relationship exist between the analyzed variables. This phenomenon is called multicollinearity and has very negative consequences for estimation the regression coefficients. In computational mathematics, these concepts correspond to the degeneracy and poor conditionality of the matrix X^TX, i.e. for the latter there doesn't exist $(X^TX)^{-1}$ and its determinant is close to zero. Consequences of this violation are particularly serious for models whose estimated parameters are subject to

physical interpretation. One of the ways to solve the multicollinearity problem may be that the equation must contain only terms that uncorrelate with each other.

In the analysis of radiointerferometric data, it was revealed that the matrix of correlation coefficients contained 76 coefficients exceeding modulo 0.5. 30 of these values of coefficients are greater than 0.95, which indicates an almost linear relationship between the estimated parameters. During the research of data of water source according to y_1-y_7, from 1 to 3 correlating parameters of the model were revealed. It should also be mentioned that in the mathematical models under consideration the data on the factors and on the response have a different physical meaning and different physical dimensions. This causes computational inconvenience, because you have to work with both very large and very small numbers which can lead to computational errors.

Thus, the presence of insignificant terms in the obtained models, as well as the presence of a mutual correlation between the estimated parameters of the anomalous observations makes it possible to conclude that the assumptions of the regression analysis are violated.

4. The step evaluation method used for adaptation to identified violations

To eliminate the effect of multicollinearity and the presence of insignificant parameters in the models, the step estimation method described above was applied. Further the results of application the step evaluation method for processing different data sets are given. The main task in creating descriptive models is to determine the maximum number of parameters with the highest accuracy. For laser data, the application of the step orthogonalization method (MSE1) allowed to estimate all parameters of the model. During selection only significant parameters of the step estimation method (MSE2), 10 corrections were obtained, and the MSE3 algorithm gave estimates for 9 corrections. The step regression method, which was used for comparison allowed us to estimate only 9 parameters out of 24 possible ones.

For radiointerferometric data: the application of the step regression method resulted in a model, containing estimates of 6 significant parameters out of 203 possible ones; 188 amendments were identified in the MSE1 strategy of the step-by-step assessment method, the MSE2 strategy gave an assessment of 136 amendments, and the MSE3 strategy gave 51 amendments.

Table 1. Numbers of parameters included in the model for different processing schemes

Response	SR	MSE1	MSE2	MSE3
y_1	1, 2, 3, 5	3, 4, 5, 6, 7, 8	-	-
y_2	2, 5, 6, 7, 8	3, 4, 5, 6, 7, 8	2, 4, 6, 7	2, 3, 6, 7
y_3	3, 5, 7	3, 4, 5, 6, 7, 8	1, 2, 4, 6, 7	1, 2, 3, 4, 6
y_4	2, 3, 5, 6, 7, 8	3, 4, 5, 6, 7, 8	1, 2, 3, 4, 6, 7,8	4, 6, 8
y_5	1, 2, 7, 8	3, 4, 5, 6, 7, 8	1, 2, 7	1, 2, 3, 4, 5, 6, 7, 8
y_6	2, 3, 4, 5, 6, 7	3, 4, 5, 6, 7, 8	4, 6, 7	-
y_7	2, 5, 7	3, 4, 5, 6, 7, 8	2, 5, 6, 7	5, 6, 7

Table 1 lists the sets of parameters included in the model obtained by different computational schemes (SR is a step regression) as part of the treatment of water purification data. It can be seen from the table that the strategy of step evaluation method (The MSE1 is the selection of only orthogonal parameters) allows to estimate more model parameters than a step regression, which is very important in describing the technological process. For the considered data set, the MSE2 strategy (only significant ones selection at each step) and MSE3 (choosing significant and simultaneously orthogonal parameters at each step) did not allow to obtain models better than a step regression. The table shows the structure of the model and which of the eight regressors are significant and are the part of the model. The above data allows us to conclude that for various samples in the SR model different parameters were introduced, while neither of the models included either of the controlled parameters x_7 and x_8. In the model obtained by the MSE1 strategy for all indicators of quality operation of the object y_1-y_7 the set of indicators is the same and practically in all cases x_7 and x_8 significant [16].

Comparing the estimates for the same parameters, obtained by various estimation methods, we can conclude that we have obtained values sufficiently close to each other. If we take the values obtained by the step regression method as the standard, then a very small number of estimates obtained by other methods is greatly different from the standard. Considering the ratio of the standard errors of the above estimates obtained by different estimation methods, we see that the accuracy of estimation of the unknown parameters in the considered methods of SR and MSE1 practically coincides. Thus, it can be concluded that the obtained models are applicable to the description of this technological process.

The next stage of the research is the task of choosing the best descriptive model. When solving it, it should be borne in mind that the internal criteria, i.e. criteria that don't use any additional information, in the presence of interference, can not solve the problem of choosing the best descriptive model. When using external measures, it's very important to split the initial sample into two parts. It's necessary to take into account the physical meaning and time of observation, since the initial data is a combination of several samples. It's proposed to choose a model by the criterion of minimum displacement-in consistency, which demands the model obtained from the training set, to be at least as possible different from the models obtained for the test sample. Analyzing the obtained results of processing radiointerferometric, laser observations and data on water purification, we can conclude that the methods of step estimation are effective.

5. Conclusion

The numerical experiments carried out make it possible to make the following conclusions:

- the step evaluation method allows to evaluate a larger number of model parameters;
- estimations of the step evaluation method are close to the estimations of step regression. Thus, the step evaluation method can be used for evaluating the parameters of a mathematical model, as well as for describing technical objects and technological processes. Analyzing the obtained values of the minimum displacement criteria, for the indicated observations (both radiointerferometric and laser observations and data on water purification), it can be concluded that the step evaluation methods are effective and allow us to describe the object under investigation with sufficient accuracy.

References

[1] Valeev SG, Kadyrova GR. Optimal regression search system: tutorial. Kazan: FEN, 2003; 160 p.
[2] Valeev SG, Rodionova TE. The method of stepwise orthogonalization of the and its using during least-squares taskю Izvestiya Vuzov. Geodezy and Aerophotography 2003; 6: 3–14.
[3] Valeev SG, Rodionova TE. Analysus of methods for parameters rating at multicollinear values. Izvestiya Vuzov. Geodezy and Aerophotography 1999; 5: 20–28.
[4] Valeev SG, Rodionova TE. Software for solving task of structure-parametrical ranking during data processing. Izvestiya Vuzov. Geodezy and Aerophotography 2004; 1: 25–34.
[5] Valeev SG, Kadyrova GR. Automatic system for solving least-squares method tasks. Izvestiya Vuzov. Geodezy and Aerophotography 1999; 6: 124–130.
[6] Valeev SG, Rodionova TE, Zharov VE. Methodic of statistical processing of RSDB-observings. Izvestiya Vuzov. Geodezy and Aerophotography 2008; 1: 13–18.
[7] Valeev SG, Rodionova TE, Zharov VE. Computational experiments for processing of RSDB-observings. Izvestiya Vuzov. Geodezy and Aerophotography 2008; 2: 94–100.
[8] Rodionova TE. Using adaptive-regression modelling for describing the functioning of technical object. Izvestiya of the Samara Russian Academy of Sciences scientific center 2014; 16(6-2): 572–575.
[9] Kadyrova GR. Estimation and prediction of the state of a technical object based on regression models of regressions. Automation of management processes 2015; 4(42): 90–95.
[10] Rodionova TE, Klyachkin VN. Statistical methods of estimation the drinking water quality. Reports of the Academy of Sciences of the Russian Federation 2014; 2-3: 101–110.
[11] Valeev SG, Kadyrova GR, Turchenco AA. Software system for optimal regression searching. Issues of modern science and practice. Technical science 2008; 4(14): 97–101.
[12] Kadyrova GR. Modification of the stepwise regression method for obtaining mathematical models for predicting the behavior of an object. Automation of management processes 2016; 3(45): 65–70.
[13] Valeev SG, Rodionova TE. Sequential orthogonalization of a basis in problems of the least squares method. Messenger of the Ulyanovsk state technical university 1999; 1(6): 4–9.
[14] Kadyrova GR. Software System of searching for optimal regression models of forecast . Way of science 2014; 7 (7): 10–11.
[15] Kadyrova GR. The system of searching for the optimal model. State of affairs and development prospects. Modern science potential 2015; 4(12): 8–10.
[16] Rodionova TE. Comparison of regression indicator models of the drinking water quality. Materials of 3-rd science-practical internet-conference “Interdisciplinary research in the field of mathematical modeling and informatics”. Tolyatti, 2014; 159–162.