Capabilities of the adaptive regression modeling package SSOR

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

[1]*Ulyanovsk State Technical University, Severnyi Venets str., 32, 432027, Ulyanovsk, Russia*

Abstract

The original statistical package «The system of searching for optimal regressions» is presented in the paper. The package allows performing high-precision statistical (regression) modeling of processes or phenomena with the subsequent use of models for the forecast of their output characteristics. The approach implemented in the package reduces the dimension of the model, increases the accuracy of parameter determination and improves the quality of the forecast. The effectiveness of the approach is directly proportional to the dimensionality, the degree of noisiness, and the multicollinear nature of the initial data. The features of the package make it a perspective mathematical tool for high-precision statistical calculations.

Keywords: regression modeling; prediction; methods of structural identification; model quality criteria; software package

1. Introduction

The software package «The system of searching for optimal regressions» (SSOR) is a specialized system implementing the strategy of adaptive regression modeling (ARM) [1].

At the initial stage, the ARM–approach provides the application of linear regression analysis (RA), which assumes the model postulating, least–squares method (LS) estimation, statistical analysis of the model and its components. LS–estimates $\hat{\beta}$ and forecasts $\hat{Y}$ are considered to be the best linear estimates (BLE) under certain assumptions. Unfortunately, these assumptions are violated in many cases. This leads to a distortion of LS–estimates $\hat{\beta}$, entails an uncontrolled increase in random and systematic errors of forecasts $\hat{Y}$.

At the following stages the ARM–approach includes checking the compliance of the RA–LS hypotheses, ranking the violations by the degree of distortion of the properties of the best linear estimates or depending on the purpose of the model (forecast, description or description and forecast), consistent adaptation to violations by applying appropriate computational procedures, repeated check of violations and ranking if necessary.

The main difficulties in the practical implementation of the RM–approach are as follows: selection of a global (or integrated) criterion of optimality; satisfactory solution of the problem of structural identification in conditions of high dimensionality; selection of the optimal route for checking the application conditions of the RA–LS scheme and the corresponding adaptation. The SSOR package resolves these problems [3, 8].

The main purpose of the package is to obtain regression models of processes, phenomena or functioning of objects with their subsequent use for forecasting output characteristics (responses) [9, 14]. The need for such a system is generated by great difficulties in performing such work, which requires both a multivariate calculation, and the application of various methods for estimating parameters and structural identification and analysis of residuals in the selected scenario for verifying compliance with the LS assumptions.

2. Adaptive RM

In developing and using forecast models the main goal is to achieve the properties of the best linear estimation (consistency, unbiasedness, efficiency) for the predicted value $\hat{Y}$. These properties are primarily ensured by the selection of the corresponding (optimal according to the given criterion) structure of the model from the set (on the basis of the postulated model) of competing structures. Thus, the problem of not only parametric, but also structural identification is solved.

In most cases, the postulated model

$$y_i = \beta_0 + \beta_1 x_{i1} + \quad + \beta_{p-1} x_{i,p-1} + \varepsilon; \quad i = \overline{i,n} \qquad (1)$$

is not optimal (adequate) to observations. If linear dependence (1) is considered to be suitable, the dimensionality of the model will be the main problem.

On the one side, for fear of losing significant factors, a researcher tries to include as many of them as possible in the right-hand side of the model (1). Therefore, as a rule, the model is overdetermined, which leads to: a) economic costs; b) the inclusion of non-informative, low-information and duplicating variables. The latter leads to an increase in the variance of the forecast $\hat{Y}$ for the forecast model and to a decrease in the accuracy of the estimation of the $\hat{\beta}$ -coefficients in the parametric model.

On the other side, an underdetermined model that does not contain significant factors leads to a systematic error Δ in the forecast. Here the problem arises of measuring the displacement Δ with the magnitude of the random error of the forecast in the overdetermined model. Most often the random error is greater than the systematic error. In addition, in some estimation methods other than LS, the model is deliberately burdened with bias to reduce the forecast error.

Thus, putting forward the hypothesis (1), a researcher encounters a set of competing models (structures) containing $x_0 (x_0 = 1)$ and some regressors from the set $\{x_1, \quad , x_{p-1}\}$. Since each variable $x_j \left(j = \overline{1, p-1}\right)$ can either enter the equation or not we'll have 2^{p-1} models. From this set of structures one or more competing models must be selected according to a given quality criterion.

If a standard regression analysis is used, in applied statistics after analyzing the model as a whole and its individual terms we use one-criterion search for the optimal structure. If it is impossible to apply a complete search of structures, one or another known type of incomplete search is used according to one of the model quality criteria (mean square error σ, selective coefficient of multiple correlation *R*, *F*-criterion etc.).

The most preferable for structural identification is the error in the control sample. This criterion to the maximum extent reflects the real random and systematic errors of the forecast (response) and does not have a systematic move in relation to the dimension. The error in the control sample is, naturally as «true», as the «true» control values y_i are burdened, in their turn, with various errors.

The application of the RM approach requires the development of multi-criteria search algorithms. In the general case, in order to obtain an adequate data processing model, it is necessary to solve the multicriteria optimization problem by successive adaptation to violations of the RA–LS conditions.

In some cases two-criterion methods are quite effective. The SSOR a step-by-step regression method (inclusion-exclusion scheme) is implemented using in addition to the *F*-criterion a number of other measures of comparison [10].

3. Criteria for the quality of a model

One of the most important tasks in the analysis of data is the problem of choosing a criterion for comparing competing descriptions.

The SSOR, in addition to the quality criteria of the model on the training sample (t, σ, F, R), the possibility of calculating the error on the control sample and the error on the «sliding» control sample is given [1,19].

The quality of the RA model is usually determined by the following criteria:

- mean square error σ, which is used both to assess the adequacy of the model, and to compare different models with each other;

- selective multiple correlation coefficient *R*, which is used as a linear link measure (1): the larger the value of R ($0 \le R \le 1$), the stronger the connection, i.e. the better the approximating function corresponds to observations, the high value of *R* also guarantees the suitability of the forecast model;

- *F*-criterion, when $F > 4F_T(\alpha; p-1, n-p)$ (F_T – critical value, taken from the table for the *F*-criterion) model is recognized as worthy of attention for its use for forecasting.

These quality criteria characterize the adequacy of the model only with respect to the sampling points used for its construction (training sample). This is the first stage in the study of the model in which an experimenter must be convinced that the model corresponds to observations.

If the model is intended for forecasting, then one must be sure of its suitability for determining the region that does not coincide with the sampling points y_i.

Control points are used to assess external adequacy (forecast accuracy). The initial sample is divided into training and control. The first sample builds a model or set of models; the second one estimates its adequacy or discrimination by statisticians is made.

The error in the control sample is based on an analysis of the discrepancies between the forecast $\hat{Y}$ and the known observed value Y for objects that did not participate in obtaining the model.

Since when working with small samples there is no possibility to divide them into a training sample and control one with a sufficiently large number of points, we suggest to use a criterion based on a «sliding» control sample to assess external adequacy. If, sequentially, we deduce each of the sampling objects from it, assuming that this object is a control one, and recalculating the model parameters again, the differences between y_i and $\hat{y}_i$ for a sliding control point Δ_i = «Observation minus the forecast» ($i = \overline{1,n}$; where n – total number of objects) can be used to calculate the error on a «sliding» control sample.

Consecutive exclusion of objects, corresponding to the removal of certain rows from the data matrix makes it possible to formulate an artificially new sample (check or control) of the same volume as the original one.

All procedures of structural-parametric identification included in the package realize the calculation of the statistics considered and the search for the optimal structure of the model. More detailed criteria for comparing competing models are considered in [19].

4. The software package SSOR

In organizing the optimal RM-strategy it is necessary to take into account the availability of various samples, assumptions, classes of functions, estimation methods, quality measures and their sets for the principle of multicriteria, structural identification methods competing in accordance with the principle of non-conclusive solutions of adaptation strategies to the violation of assumptions.

The practical application of RM first of all requires the full automation of all declared procedures. For this purpose the corresponding software was developed.

The SSOR package includes the following modules:

1) control module;

2 request generation module;

3) library of functional procedures;

4) script block;

5) system configuration block;

6) data editor block;

7) table formation block;

8) guide.

The main tool for positive impact on the predictive properties of the model is the algorithm for finding its optimal structure. The package includes the following structural-parametric identification procedures:

- multiple linear regression,
- comb regression,
- robust estimation,
- complete search of structures,
- incomplete search of structures (search with restriction on the number of included regressors in the model),
- search of normal systems,
- step-by-step regression with inclusion-exclusion,
- random search with adaptation,
- random search with a return.

These procedures can be performed both in automatic mode to process a number of data samples and to process a single sample of data according to the realized optimal scenario [18].

The package implements the procedure for constructing and analyzing the residue schedule, which is a useful statistical tool for testing the adequacy of the estimated regression model to the available data.

Competitiveness of SSOR with other statistical packages can be described as:

• using new methods of structural identification: full search, partial search of overdetermined and normal systems, multi-criteria method of step-by-step regression with inclusion-exclusion;

• using flexible tool for building comparative tables;

• using, in addition to the classical quality criteria of the model in the training sample, the quality criteria of the model in the control sample and the errors in the «sliding» control sample, which allows an external model adequacy assessment (forecast accuracy) to be performed.

At present work is under way to enhance the capabilities of the SSOR and its intellectualization [15, 17].

5. Using the SSOR package

The SSOR package can be used to solve the problems of the least-squares method (problems of recovering dependencies from excessive indirect observations) and regression analysis in any areas (ecology, technological processes, economics, sociology etc.), various tasks requiring restoration of the empirical relationship between the output process parameter and the input set.

In processing aerospace photographs [2] and solving a number of photogrammetric problems [4] the use of SSOR by computational experiments allowed to obtain the following results:

1. Obtaining models for transforming coordinates from small samples with a variance of the accuracy estimation of 1.2-100 times smaller than the variance in the standard approach, which corresponds to an increase in the approximation accuracy when applying PM up to several times.

2. Increase of accuracy in the use of RM is ensured by the procedure of structural identification. The implementation of the latter implies the formation of a set of competing structures based on the initial perspective model and the search for the optimal structure according to a given quality criterion.

3. Search for a model that is optimal by error on a «sliding» control sample leads to a model with better predictive properties than models that are optimal for the mean square error and allows to solve the problem of selecting a set of regressors that is informative by the t-criterion. In 70% of all cases models that are optimal for the mean square error contain low-information terms. Models that are optimal by error on a «sliding» control sample contain only insignificant terms only in 17% of all cases. Models containing little informative terms obtained by mistake on a «sliding» control sample contain one insignificant regressor, while models derived from the mean square error usually have two or more little informative terms. It was estimated that an improvement in external accuracy makes a significant difference in the quality of the error in the «sliding» control sample. The analysis showed that application of this criterion gives a significant improvement in predictive properties compared to the mean square error. The stability of the conclusions with respect to the observations included in the control sample was verified and confirmed by 10 random experiments for each of the three randomly selected images.

The SSOR package was successfully used to process laser [5] and high-dimensional radiointerferometric data [6, 7], for assessing the quality of drinking water [11, 16], for processing socio-economic indicators [12, 13].

6. Conclusion

The SSOR package can be useful in the development of forecast models in high-precision areas of knowledge, in technological processes with input characteristics that contain interdependent, non-informational or little informational factors and in socio-economic phenomena and environmental situations. The application of the package provides an increase in the accuracy of forecasting using the optimal model up to several times.

References

[1] Valeev SG, Kadyrova GR. The system of searching for optimal regressions: tutorial . Kasan : FEN, 2003; 160 p.

[2] Kadyrova GR, Bilibina NA, Bugaevskii LM, Valeev SG. Regression models for transforming images in aerocosmic pictures. Izvestiya Vuzov. Geodezy and Aerophotography 1997; 1: 56–66.

[3] Valeev SG, Kadyrova GR. Automatic system for solving least-squares method tasks. Izvestiya Vuzov. Geodezy and Aerophotography 1999; 6: 124–130.

[4] Valeev SG, Kadyrova GR. Optimal reduction models in photographical astronometry. Izvestiya Vuzov. Geodezy and Aerophotography 2002; 3: 58–69.

[5] Valeev SG, Rodionova TE. The method of stepwise orthogonalization of the basis and its using during least-squares task. Izvestiya Vuzov. Geodezy and Aerophotography 2003; 6: 3–14.

[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] 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.

[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] 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.

[11] 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.

[12] Rodionova TE, Rybkina MV. Using mathematical modeling for the analysis of thesocial sphere influence on the quality of life of the population (on the example of the Ulyanovsk region). Economic analysus: theory and practice 2014; 32(383): 61–66.

[13] Rodionova TE, Rybkina MV, Ananeva NA. Research of inflation impact on socio-economic factors. Quality. Innovation. Education 2015; 9(124): 48–51.

[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, Klyachkin VN. Statistical methods of estimation the drinking water quality. Reports of the Academy of Sciences of the Russian Federation 2014; 2-3(23-24): 101–110.

[17] Kadyrova GR. Possibilities of a software regression modeling system for estimating a model and searching for its optimal structure. Radioelectronic engineering 2015; 2(8): 228–233.

[18] Kadyrova GR. Formation of strategies for finding optimal regressions // Modern problems of design, production and operation of radiotechnical systems 2016; 1(10): 178–180.

[19] Kadyrova GR. Research of the quality measures of models for assessing the state of a technical object. Synthesis, analysis and diagnostics of electronic circuits 2016; 13: 71–83.