SAINT PETERSBURG STATE UNIVERSITY
FACULTY OF MATHEMATICS AND MECHANICS

NUMERICAL METHODS FOR STUDYING ERRORS IN SOME STATISTICAL PROBLEMS

BY

MAHMOUD SAIF ABDUL-RAHMAN EID

SPECIALITY: 05.13.18
MATHEMATICAL MODELLING, NUMERAL METHODS AND COMPLEX PROGRAMMING

AUTHOR'S SUMMARY
OF THE THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
PHILOSOPHY OF DOCTORATE (Ph.D.)

SAINT-PETERSBURG
2002

The work has been accomplished
on the Chair of Statistical Simulation,
St. Petersburg State University

Supervisor:
Prof. Ermakov S.M.

Official opponents:
Prof. Nevzorov V.B.
Prof. Sedunov V.V.

Leading organization:
Institute of Engineering Sciences
of Russian Academy of Science

The defense will take place on " " 2002 at on the meeting of the Dissertation Counsel Д.212.232.51 on the defense of thesis submitted for the requirements for the degree of Philosophy of Doctorate in Saint-Petersburg State University. Address: 198504, St. Petersburg, Stary Peterhof, Bibliotechnaja sq., 2, the Faculty of Mathematics and Mechanics.

You can acquaint yourself with the thesis in the M. Gorky Research Library of Saint-Petersburg State University. Address: 199034, St.Petersburg, Universitetskaja nab., 7/9.

The author's summary was sent on " " 2002

Scientific Secretary
of the Dissertation Counsel Prof. B.K. Martynenko

GENERAL OVERVIEW OF THE THESIS

The problem of data analysis is an important part of modern research. Regression analysis, in particular, is important, as it is used to create models that can be used in further research, data smoothing and complex systems behaviour prediction. In all cases, secure estimation (with a given reliability level) of model and prediction errors is of prior interest. This problem has been studied well enough for the so called "classic"case - when approximate normalcy and independence of observation errors, as well as the absence (or smallness) of systematic error are assumed. In other cases, there exist a number of unsolved problems that are of great interest for the theory and applied spheres. We must also note that for an important group of problems on experimental design for regression analysis in the "non-classic"case very few results have been obtained so far.

Topicality of the problem. Due to wide use of computers in data processing, the construction of computational procedures of error estimations for regression analysis in the "non-classical"case, the construction of confidence and tolerance intervals for dependent observations with systematic error and for distributions, different from normal, becomes a topical issue. The development of computational procedures and software for solving the above mentioned problems is highly relevant for construction and proof of probability and simulation models in medicine, biology, economics, engineering, as well as in other fields.

Purpose of research. The purpose of research was the following:

1. Obtaining new theoretical results in the field of regression analysis with systematic error and dependent observation errors. Development of efficient computational procedures to estimate error and construction of random experimental designs.

2. Development of a software system for computational application of the theoretical results obtained.

3. Development of efficient computational methods to construct tolerance intervals for the case of Gamma distribution.

4. Development of software to calculate tolerance intervals.

Scientific novelty. The first software using new methods of dividing systematic and random error components in linear regression analysis with independent observation error has been developed. Random least square estimators in the case of correlated observations have been proposed and studied. For the first time, computational procedures of tolerance intervals construction for Gamma distribution have been developed. The software system using methods developed by the author of the dissertation and previously is new.

Scientific and applied value. The software developed in the thesis and the theoretical observations on the results previously obtained by S. M. Ermakov and S. M. Ermakov and R. Schwabe allow for the construction of new bootstrap procedures for dividing random and determined components of the regression function error. These generalisations are a new field of research in some issues in regression analysis. The computational procedures and the software system constructed on their basis can be used in research of a wide range of probability and simulation models. Computational methods and software for the calculation of tolerance intervals in the case of Gamma distribution may be used in data processing in medicine, biology, economics and other fields.

The following items are subject for defence:

The new theoretical results concerning randomisation of least square estimators in the case of dependent observations and bootstrap procedures related to the randomisation. Methods of calculation of tolerance intervals in the case of Gamma distribution. Algorithms and software for the estimation of errors in least square method and calculation of tolerance intervals.

Approbation of results.

The results obtained in the thesis were reported at the seminar of the Department of Statistical Modelling of St. Petersburg University, the seminar of the Department of Information Management of St. Petersburg University and at the seminar of the Institute of Engineering Science of Russian Academy of Sciences.

Publications.

The contents of the thesis was presented in one published article.

Structure and size of the dissertation.

The dissertation is in English and contains 150 pages of text and 50 pages of Appendix.

The first two chapters of the thesis are devoted to the study generalization of special distribution of the interpolation nodes, the so-called Δ^2-distribution. This distribution was brought into consideration as an instrument to decrease the variance in integral evaluation by Monte Carlo methods.

Let μ be a σ-finite measure on a set $\mathfrak{X}$, $x \in \mathfrak{X}$, $f(x)$ is μ-square integrable function and $\varphi_1(x), ..., \varphi_m(x)$ is a set of orthogonal and normalized with respect to μ functions.

Let us define $\Delta(Q) = det\|\varphi_i(x_j)\|_{i,j=1}^m, \quad Q = (x_1, ..., x_m)$. Let us say that a random point $P = (y_1, ..., y_m), \quad y_i \in \mathfrak{X}$ has the Δ^2 -distribution (with respect to measure μ^n) if the density of this distribution with respect

to $\mu^n(dQ)$) is determined by the formula

$$\Delta^2(Q) = \frac{1}{m!}\Big(det\Big\|\varphi_i(x_j)\Big\|_{i,j=1}^{m}\Big)^2. \tag{1}$$

The density (1) has in particular the next properties.
If $C_1, ..., C_m$ are random values determined from a system of equation

$$\sum_{j=1}^{m} C_j\varphi_j(y_i) = f(y_i), \quad i = 1, ..., m,$$

where y_i are random points with density of distribution (1), then

1. Expectation of C_j is expressed by the formula

$$EC_j = \int f(x)\varphi_j(x)\mu(dx).$$

(in the left hand of the equality is j-th Fourier coefficient of function f on the system $\{\varphi_i\}$.)

2. For the systematic component of the variance of C_j the next inequality holds

$$DC_j \leqslant \int \Big[f(x) - \sum_{i=1}^{m} \varphi_j(x)EC_j\Big]^2\mu(x).$$

the sign of the equality takes place in the case when the system $\{\varphi_i\}$ satisfies the next condition (regularity):

$$\mu^m\{Q : \Delta(Q) = 0\} = 0.$$

Δ^2-distribution has also applications in regression analysis, if we suppose that that f has random errors - the values

$$\varsigma(x_i) = f(x_i) + \varepsilon_i,$$

are observed, and the measure μ is concentrated in a finite set of N points in region $\mathfrak{D} \in R^s$.

The properties of Δ^2-distribution are studied in the first two chapters of the thesis specifically in connection to the problems of regression analysis.

Chapter 1 is devoted to the case of non-correlated errors of observations. First, some well-known facts relevant for further discussion

(see 1.1-1.4) are given and then, methods and algorithms of Δ^2-distribution modelling are described.

Two experimental methods are discussed. The first one is the so called active experiment. In this case the random vectors, $Q = (y_1, ..., y_1)$ are considered as random experimental designs (random replicates). Note that the exact D-optimal design is the mode of Δ^2- distribution. If it is possible to repeat measurements (the experiment) in the same y_j points, then well-known procedures of dividing the systematic and random components of the variance may be used.

The second approach deals with the case when experimental data exists in advance, and repetition of measurements in the points that were used before is impossible or inexpedient. For this case the procedure of random choice of data for dividing the error components is also mentioned. The procedure is based on the earlier got results (Ermakov, Schwabe). In the thesis, the implementation and the software for this procedure is given. Some examples of applications of this procedure to the model data are cited.

The direct (without orthogonality φ_i suppositions) proof of unbiasedness of the least square estimators is contained in the final paragraph of chapter 2. Namely the next lemma is proved:

Lemma. 1.1 *Let $\varsigma(x)$ be values of random function ς, observed in the points x are designed linear independent on the support of measure,*

$$\varsigma(x) = f(x) + \varepsilon(x), \quad E\varepsilon(x) = 0, \quad E\Big(\varepsilon(x_i)\varepsilon(x_j)\Big) = \begin{cases} \sigma^2 & \text{if } x_i = x_j \\ 0 & \text{if } x_i \neq x_j \end{cases},$$

and $\varphi_1, ..., \varphi_m$ are designed linear independent on the support of measure μ functions.

If random values $C_1, ..., C_m$ are calculated as the solution of the linear equation system:

$$\sum_{j=1}^{m} C_j \varphi_j(x_i) = \varsigma(x_i), \qquad i = 1, ..., m,$$

where $(x_1, ..., x_m)$ are chosen at random in accordance to Δ^2-distribution, then the next equalities hold

$$E(C_j|\varepsilon_1, ..., \varepsilon_m) = \widehat{C}_j, \qquad EC_j = \widetilde{C}_j,$$

Here

$$(\widehat{C}_1,...,\widehat{C}_m) = \min_{d_1,...,d_m} \int \Big[\varsigma(x) - \sum_{l=1}^{m} d_l\varphi_l(x)\Big]^2 \mu(dx).$$

$$(\widetilde{C}_1,...,\widetilde{C}_m) = \min_{d_1,...,d_m} E\int \Big[\varsigma(x) - \sum_{l=1}^{m} d_l\varphi_l(x)\Big]^2 \mu(dx) =$$

$$= \min_{d_1,...,d_m} \int \Big[f(x) - \sum_{l=1}^{m} d_l\varphi_l(x)\Big]^2 \mu(dx).$$

Chapter 2 devoted to the randomisation of the least square estimators in the case of dependent observations, when the following equality takes place

$$E\varepsilon(x_i)\varepsilon(x_j) = b_{ij}. \tag{2}$$

and the matrix $B = \|b_{ij}\|$ is not diagonal generally speaking and is positive defined.

The sections 2.1 - 2.4 are devoted to auxiliary facts from linear algebra and least square method. In section 2.5 the analog of Δ^2-distribution under supposition (2) is constructed. The known decomposition $B = \Gamma\Gamma^T$ is used, where Γ-is a lower triangle matrix. In the sections 2.6-2.7 the algorithms of generalised Δ^2-distribution simulation are described and some examples of applications of the developed procedures to the experimental data analysis are discussed. As well as in the case of non-correlated observations, the results and algorithms can be useful for optimal design construction. The final section 2.8 of this chapter is devoted to the general problem of construction Δ^2-distribution analogue for the case when the error is a gaussian random process (field).

Correlation function $B(x,y) = E(\varepsilon(x),\varepsilon(y))$ of such process is a bilinear positive functional in the Hilbert space and consequently has representation

$$\mathfrak{B}(\xi,\eta) = (\widetilde{\mathfrak{B}}\xi,\eta),$$

where $\widetilde{\mathfrak{B}}$ is a linear positive operator. In the supposition of strong positivity there also exists an inverse integral operator Γ. We shall denote the kernel of Γ as $\gamma(x,y)$.

Now let the constants $\widehat{C}_1,...,\widehat{C}_m$ and $\widetilde{C}_1,...,\widetilde{C}_m$ be found from the conditions $(\widehat{C}_1,...,\widehat{C}_m) = \min\limits_{d_1,...,d_m} \mathsf{F}(\varsigma,d_1,...,d_m)$ and $(\widetilde{C}_1,...,\widetilde{C}_m) = \min\limits_{d_1,...,d_m} \mathsf{F}(f,d_1,...,d_m)$ respectively. Here $\mathsf{F}(f,d_1,...,d_m) = \int\int[g(x) -$

$\sum d_l\varphi_l(x)][g(y) - \sum d_l\varphi_l(y)]\mu(dx)\mu(dy)$, and $\varphi_1, ..., \varphi_m$ are prescribed linear independent on the support of μ functions.

The following theorem is proved.

Theorem. 2.5. *Let the point $Q = (y_1, ..., y_m)$ be chosen at random with respect to the density*

$$\Delta_d^2(Q) = \frac{det\|\varphi_l(x_i)\|_{1,l=1}^m \cdot det\|\psi_s(x_j)\|_{1,j=1}^m}{\int \mu^m(dQ) \cdot det\|\varphi_l(x_i)\|_{1,l=1}^m \cdot det\|\psi_s(x_j)\|_{1,j=1}^m},$$

where $\psi_k(x) = \int \gamma(x,y)\varphi_k(y)\mu(dy)$.

Then, if C_l $(l = 1, ..., m)$ are determined from the system of equations

$$\sum_{l=1}^{m} C_l\varphi_l(y_i) = \varsigma(y_i),$$

then $E(C_l|\varepsilon(y_i), ..., \varepsilon(y_m)) = \widehat{C}_l$ and $EC_l = \widetilde{C}_l$.

The second part of the thesis is devoted to the construction and investigation of tolerance intervals for **Gamma** distribution.

Tolerance intervals are statistical intervals constructed for the purpose of encompassing a specified proportion of the population with a specified assurance probability. Tolerance interval can be also defined as an interval with random bounds that contains no less than the prescribed part of the probability mass of continuous distribution with probability γ, prescribed earlier. Moreover this distribution is the distribution of our sampling data.

The algorithms of numerical construction of tolerance intervals for the given distribution can be very complicated. Tolerance intervals of Normal distribution have been studied.

The tolerance intervals for Gamma distribution is not studied well enough in spite of its practical importance (application in the fields of biology, medicine, insurance etc.).

Chapter 3 contains the general concepts of prediction and tolerance intervals, as well as their classifications an types. The notions of β-expectation and β-content tolerance intervals are illustrated on the example of the Normal distribution. In the case of β-expectation interval we expect that 100% β of the population will be within this interval $(0 < \beta < 1)$.

In the case of β-content tolerance interval we expect that at least 100% β of the population will be in this interval with γ confidence level ($\gamma, \beta \in (0,1)$).

Section 3.7 is dedicated to a short review of the literature.

Chapter 4 contains Introduction, a comparatively simple case of Exponential distribution (section 4.2) and then a very detailed presentation of tolerance intervals for γ-distribution.

Let $(Q_1 v, Q_2 v)$, (Qv, ∞) and $(0, Qv)$ be a two sided, a lower bound and upper bound tolerance intervals respectively. In sections 4.3 and 4.4 their properties are studied and the algorithms for their calculations are constructed.

The theorems 4.1-4.5 formulated and proved in these sections are devoted to problems of the tolerance interval uniqueness. The expression for the constant Q, Q_1, and Q_2 is also done.

As an example we formulate the next theorem.

Theorem. 4.1. *The constant Q is the upper-bound β-expectation tolerance interval for the γ-distribution is unique and equal to*

$$Q = \frac{1 - Be(n\alpha_1, \alpha_2, 1-\beta)}{Be(n\alpha_1, \alpha_2, 1-\beta)},$$

where $Be(n\alpha_1, \alpha_2, 1-\beta)$ denotes the β quantile for $Be(n\alpha_1, \alpha_2)$ (Be is β-distribution).

Section 4.5 is devoted to problems of simulation methods for constructing of the tolerance intervals. is devoted to problems of simulation methods for constructing of the tolerance intervals in the cases of Normal and Gamma distributions. The detailed analysis of the computational results shows a big difference in these two cases (apart from the cases of equality of means under big sampling size).

The appendix to the thesis contains software in Pascal and Fortran programing languages with comments. The software may be used in both in experimental data analysis and in the investigations of model problems. They can be used also in teaching of the applied statistics and data analysis.

MAIN RESULTS AND CONCLUSIONS

1. A software system for Gamma distribution simulation for the case of independent errors of observation has been created. It served as basis the software that conducts the resampling procedure of dividing the systematic and random component of errors for estimates of least squares.

2. The results by S. M. Ermakov and V. G. Zolotukhin for the case of linear- independent (not orthonormalized) basic functions of linear regression have been generalised.

3. The generalization of Gamma distribution for the case of dependant observations has been shown. A software system that simulates the indicated distribution (Gamma distribution) has been developed.

4. Algorithms and a system of software for construction of different types of tolerance intervals in the case of Gamma distribution have been developed.

5. A detailed comparison of the tolerance intervals for the cases of the Normal and Gamma distributions has been carried out. It was demonstrated that but for some rare exceptions there is an essential difference between the tolerance intervals in these two cases.

Thus, in the thesis, a number of theoretical and computational problems of statistic simulation have been solved and a software system that can be a useful addition to the existing package of data processing software has been presented, especially in the field that relates to the analysis of prediction error with the help of the regression parameters estimator and the construction of confidence intervals.

PUBLICATIONS ON THE THESIS SUBJECT

1. S.M. Ermakov, Mahmoud S. Eid. β-expectation tolerance intervals for Gamma distribution, Проблемы оптимизации дискретных систем. Сб. Под редакцией док. физ.-мат. наук, проф. М.К. Чиркова. СПб. Изд-во НИИХ СПбГУ 2001: стр. 45-59.