Metadata of the chapter that will be visualized in SpringerLink

Book Title	Automation 2019	
Series Title		
Chapter Title	Estimation of Linear Regression Parameters of Symmetric Non-Gaussian Errors by Polynomial Maximization Method	
Copyright Year	2020	
Copyright HolderName	Springer Nature Switzerland AG	
Author	Family Name	**Zabolotnii**
	Particle	
	Given Name	**Serhii W.**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Cherkasy State Technological University
	Address	Cherkasy, Ukraine
	Email	s.zabolotnii@chdtu.edu.ua
Corresponding Author	Family Name	**Warsza**
	Particle	
	Given Name	**Zygmunt L.**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Industrial Research Institute for Automation and Measurements PIAP
	Address	Al. Jerozolimskie 202, 02-486, Warsaw, Poland
	Email	zlw1936@gmail.com
Author	Family Name	**Tkachenko**
	Particle	
	Given Name	**Oleksandr**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Cherkasy State Technological University
	Address	Cherkasy, Ukraine
	Email	

Abstract

In this paper, a new way of estimation of single-factor linear regression parameters of symmetrically distributed non-Gaussian errors is proposed. This new approach is based on the method of maximizing polynomials (PMM) and uses the description of random variables by higher order statistics (moments and cumulants). Analytic expressions that allow to find estimates and analyze their asymptotic accuracy are obtained for the degree of polynomial $S = 3$. It is shown that the variance of polynomial estimates can be

less than the variance of estimates of the ordinary least squares' method. The increase of accuracy depends on the values of cumulant coefficients of higher order of the random regression errors. The statistical modeling of the Monte Carlo method has been performed. The results confirm the effectiveness of the proposed approach.

Keywords (separated by '-') Linear regression - Symmetrical distribution of errors - Stochastic polynomial - Variance - High order statistics - Cumulants

Estimation of Linear Regression Parameters of Symmetric Non-Gaussian Errors by Polynomial Maximization Method

Serhii W. Zabolotnii[1], Zygmunt L. Warsza[2(✉)], and Oleksandr Tkachenko[1]

[1] Cherkasy State Technological University, Cherkasy, Ukraine
s.zabolotnii@chdtu.edu.ua
[2] Industrial Research Institute for Automation and Measurements PIAP, Al. Jerozolimskie 202, 02-486 Warsaw, Poland
zlwl936@gmail.com

Abstract. In this paper, a new way of estimation of single-factor linear regression parameters of symmetrically distributed non-Gaussian errors is proposed. This new approach is based on the method of maximizing polynomials (PMM) and uses the description of random variables by higher order statistics (moments and cumulants). Analytic expressions that allow to find estimates and analyze their asymptotic accuracy are obtained for the degree of polynomial $S = 3$. It is shown that the variance of polynomial estimates can be less than the variance of estimates of the ordinary least squares' method. The increase of accuracy depends on the values of cumulant coefficients of higher order of the random regression errors. The statistical modeling of the Monte Carlo method has been performed. The results confirm the effectiveness of the proposed approach.

AQ1

AQ2

Keywords: Linear regression · Symmetrical distribution of errors · Stochastic polynomial · Variance · High order statistics · Cumulants

1 Introduction

A linear one-factor relationship between two variables is very simple and is the most common type of regression models. In addition, many non-linear dependencies (power, exponential, logarithmic, etc.) can also be reduced to a linear model by corresponding transformations. The consequence of this is the widespread use of linear regression in technical, ecological, medical economic, and other applications.

The most common way of finding estimates of the parameters of linear regression is to use the method OLS (Ordinary Least Squares) [1, 12]. The OLS-estimations are linear, unmixed and have minimal variance, provided that the errors are homogeneous, uncorrelated and normally distributed. However, many researchers note that quite often the normal (Gaussian) law is not the only possible probabilistic model of regression errors. In earlier publications it was shown that regression errors can have larger tails [1, 2] in comparison with the Gaussian distribution or may be limited [3].

R. Szewczyk et al. (Eds.): AUTOMATION 2019, AISC 920, pp. 1–14, 2020.
https://doi.org/10.1007/978-3-030-13273-6_59

For the non-Gaussian nature of statistical data, an increase in accuracy can be achieved using a parametric approach based on the Maximum Likelihood Estimator (MLE) method. The use of MLE requires an adequate description of error models based on probability density (PDF). To solve the problems of regression analysis, many different types of symmetric PDFs are used: elliptic laws (Logistic, Cauchy, Student t) [4–6], the family of Exponential Power Family (EPF) exponential laws [7] and mixtures of Gaussian distributions [8, 9]. In addition, models of symmetric distributions, that allow changing the magnitude of the kurtosis coefficient and the severity of the tails of regression errors are specially developed [10–12].

From a computational point of view, the parametric approach is characterized by a significantly greater complexity (relative to the least-squares method), as well as a significant increase in the volume of a priori information. This is due to the necessity of preliminary specification (selection) of the probability distribution law for the error model and evaluation of non-informative parameters. Therefore, in practice, non-parametric methods as simpler from the implementation point of view are often used. These may be robust versions of the least-squares method [13], whose application is primarily aimed at ensuring stability against the influence of extreme deviations (emissions) [2], estimating the Least Absolute Deviation (LAD) method [14], quantile [15] and signed [16] methods of estimation.

A compromise from the point of view of the complexity and completeness of the probabilistic description of non-Gaussian random variables is the approach based on Higher-Order Statistics (HOS). Examples of its use for solving regression analysis problems are given in the papers [17–19].

2 The Aim of Research

This paper is a direct continuation of the paper [20], where the application of the Polynomial Maximization Method (PMM) [21] for the solution of the regression analysis (estimation of linear regression parameters for asymmetrically distributed errors) was firstly considered. Conceptually PMM is nearly like MLE, since it also uses the principle of maximizing statistics from sample data near the true value of the estimated parameter. However, for the formation of such statistics, descriptions of random variables by PDFs are not used, but by higher-order statisticians, for example, by moments or cumulants. We note the functionality of PMM, which can be used not only to find estimates of the scalar parameter [22–24], but also for finding the change-point problem of the properties of a random sequence in a posteriori formulation of the problem [25], as well as used in joint detection signals against the background of non-Gaussian noise [26]. The totality of the results obtained in these papers shows that for non-Gaussian statistics the MMP-estimations can be much more effective (have a smaller variance) compared with linear estimates.

The purpose of this study is to synthesize algorithms for finding adaptive PMM-estimators, and to estimate their accuracy by statistical Monte Carlo simulation using the model of a single-factor linear regression with symmetrically distributed non-Gaussian errors.

3 Mathematical Formulation of the Problem

Let's have a one-factor regression model of observations describing the values dependence of the target variable y on its predictor x:

$$y_v = f(x_v) + \xi_v, v = \overrightarrow{1,N}, \tag{1}$$

where $f(x_v) = a_0 + a_1 x_v$ - determined linear dependence component and ξ_v - random component (error) of the model $(E\{\xi\} = 0)$.

Regression errors are a sequence of independent and identically distributed random variables that have a symmetric distribution, which is different from the Gaussian law. Probabilistic properties of errors can be described with the help of higher-order statistics (there are moments μ_r and cumulants κ_r, $r = \overrightarrow{2,6}$), the values of which are a priori unknown. The problem consists in finding estimates of the vector parameter $\boldsymbol{\theta} = \{a_0, a_1\}$ on the basis of a statistical analysis of the set of points (x_v, y_v), $v = \overrightarrow{1,N}$.

4 Estimation of the Vector Parameter by the Polynomial Maximization Method

To solve the problem, we use the variant of the polynomial maximization method (PMM), which is developed for the case of estimating the vector parameter for unequally distributed statistical data [20, 21]. According to the PMM, the estimate of a vector parameter θ of dimension Q can be found as the solution of Q stochastic power equation system:

$$\sum_{v=1}^{N} \sum_{i=1}^{S} k_{i,v}^{(p)} \left[(y_v)^i - \alpha_{i,v} \right] \Bigg|_{\theta_p = \hat{\theta}_p} = 0, \; p = \overrightarrow{0, Q-1} \tag{2}$$

where: S – is the order of the polynomial used for parameter estimation, $\alpha_{i,v} = E\{(y_v)^i\}$ – are theoretical initial moments of the i-th order from a sequence y_v.

Coefficients $k_{i,v}^{(p)}$ (for each component of vector parameter $p = \overrightarrow{0, Q-1}$) can be found by solving the system of S linear algebraic equations, given by conditions of minimization of variance (with the appropriate order S) of the estimate of the parameter $\boldsymbol{\theta}$, namely:

$$\sum_{i=1}^{S} k_{i,v}^{(p)} F_{(i,j)v} = \frac{\partial}{\partial \theta_p} \alpha_{j,v}, \; j = \overrightarrow{1,S}, \; p = \overrightarrow{0, Q-1}, \tag{3}$$

where $F_{(i,j)v} = \alpha_{(i+j),v} - \alpha_{i,v}\alpha_{j,v}$.

Systems of Eq. (3) can be solved analytically using the Kramer method.

It was shown that PMM-estimations, which are the solutions of system of stochastic equations of the form (2), are consistent and asymptotically unbiased. To calculate the

variance of parameter estimates is necessary to find the volume of extracted information on the estimated parameters θ, which generally are described by the equation:

$$J_{SN}^{(p,q)} = \sum_{v=1}^{N}\sum_{i=1}^{S}\sum_{j=1}^{S} k_{i,v}^{(p)} k_{j,v}^{(q)} F_{(i,j)v} = \sum_{v=1}^{N}\sum_{i=1}^{S} k_{i,v}^{(p)} \frac{\partial}{\partial\theta_q}\alpha_{i,v},\ p,q = \overrightarrow{0,\ Q-1}. \quad (4)$$

The statistical sense of function $J_{SN}^{(p,q)}$ is like the classical Fisher concept of information quantity. The asymptotic values (for $N \to \infty$) of the variances $\sigma^2_{(\theta_p)S}$ of PMM-estimates - components the vector parameter $\boldsymbol{\theta}$ lie on the main diagonal of the variational matrix of estimates:

$$\boldsymbol{V}_{(S)} = \left[\boldsymbol{J}_{(S)}\right]^{-1}. \quad (5)$$

Obtained by inversion of the quadratic (dimension Q) matrix of the amount of information retrieved $\boldsymbol{J}_{(S)}$, consisting of elements $J_{SN}^{(p,q)}$ like in (4).

5 Polynomial Estimation of Linear Regression Parameters

It was shown in [20] that when the degree of the polynomial of the PMM-estimator of the vector parameter $\boldsymbol{\theta} = \{a_0, a_1\}$ of the linear regression model (1) is used, they completely coincide with linear OLS estimates. Such estimates are optimal (by the criterion of minimum variance) in a situation where the errors of the regression model have a Gaussian distribution. In addition, it was shown in [20–23] that when the statistical data are symmetrically distributed (which corresponds to zero values of unpaired order cumulant coefficients), PMM-estimator with the degree of polynomial $S = 2$ degenerate into linear estimates obtained for $S = 1$. Therefore, the case of finding PMM-estimators, which is based on the use of polynomials of degree, it is considered below. When PMM-estimates of two components from the sought vector parameter are found from the solution of a system of two equations formed on the general formula (2):

$$\begin{aligned}&\sum_{v=1}^{N}\left\{k_{1,v}^{(p)}[y_v - f(x_v)] + k_{2,v}^{(p)}\left[(y_v)^2 - \left([f(x_v)]^2 + \mu_2\right)\right]\right.\\ &\left. +\ k_{3,v}^{(p)}\left[(y_v)^3 - \left([f(x_v)]^3 + 3f(x_v)\mu_2\right)\right]\right\} = 0, \quad p = \overrightarrow{0,1}\end{aligned} \quad (6)$$

where the optimal coefficients $k_{i,v}^{(p)},\ i = \overrightarrow{1,\ 3}$ ensure the minimization of the variance of the estimates (for the corresponding degree of the polynomial). They can be found by solving systems of linear equations of the form (3) described by expressions:

$$\begin{aligned}k_{1,v}^{(p)} &= \tfrac{1}{\Delta_3}\left[3[f(x_v)]^2\left(\mu_4 - 3\mu_2^2\right) + 3\mu_4\mu_2 - \mu_6\right]\tfrac{\partial}{\partial a_p}f(x_v),\\ k_{2,v}^{(p)} = \tfrac{-3}{\Delta_3}f(x_v)[\mu_4 - 3\mu_2]\tfrac{\partial}{\partial a_p}f(x_v),\ k_{3,v}^{(p)} &= \tfrac{1}{\Delta_3}[\mu_4 - 3\mu_2]\tfrac{\partial}{\partial a_p}f(x_v),\ p = \overrightarrow{0,1},\end{aligned} \quad (7)$$

where $\Delta_3 = \mu_2^{-2}\left(\mu_4^2 - \mu_2\mu_6\right)$.

Substituting the coefficients (7) in (6), after some transformations, the system of equations for finding the estimates of the components of the parameter can be represented in the form:

$$\sum_{v=1}^{N} (x_v)^{p-1}\left[A(a_0+a_1x_v)^3+B(a_0+a_1x_v)^2+C(a_0+a_1x_v)+D\right]=0,\ p=\overrightarrow{0,1} \quad (8)$$

where $A=1$, $B=-3\hat{\alpha}_1$, $C=3\hat{\alpha}_2-\frac{\mu_6-3\mu_4\mu_2}{\mu_4-3\mu_2^2}$, $D=\hat{\alpha}_1\frac{\mu_6-3\mu_4\mu_2}{\mu_4-3\mu_2^2}-\hat{\alpha}_3$.

Note that in (8) the statistics $\hat{\alpha}_i=\frac{1}{N}\sum_{v=1}^{N}(y_v)^i$, $i=\overrightarrow{1,3}$ are sample initial moments, and μ_2, μ_4 and μ_6 - theoretical central moments of the regression errors.

An analysis of expression (8) shows that, using polynomials of degree, PMM estimates can only be found by means of a numerical solution of systems of nonlinear equations. As a first approximation, using the appropriate iterative procedures (for example, based on the Newton-Raphson method) it is logical to use OLS estimators to the required parameters.

6 Analysis the Accuracy of Polynomial Estimates

To quantify the magnitude of changes in the accuracy of estimates, we use the notion of a coefficient for reducing the variance of estimates of linear regression parameters:

$$g_{(\theta_p)S}=\frac{\sigma^2_{(\theta_p)S}}{\sigma^2_{(\theta_p)OLS}}=\frac{\sigma^2_{(\theta_p)S}}{\sigma^2_{(\theta_p)1}},\ p=\overrightarrow{0,1}. \quad (9)$$

These coefficients are formed as the variance ratios of PMM parameter θ, estimates found using the polynomial of the first order to the variances of the OLS estimates of the corresponding components of this vector parameter. And since OLS estimates are equivalent to PMM estimates obtained for a $S = 1°$, their variances will also coincide, i.e. $\sigma^2_{(\theta_p)OLS}=\sigma^2_{(\theta_p)1}$ [20].

It is shown [20] that using the apparatus of the amount of extracted information, the variance of PMM-estimates (for $S=1$) linear regression parameters can be found as elements of the main diagonal of the variance estimate matrix of the form:

$$\boldsymbol{V}_{(1)}=\mu_2\left[\boldsymbol{B}\boldsymbol{B}^T\right]^{-1}, \quad (10)$$

where $\boldsymbol{B}$ - matrix of $2\times N$ with elements $b_{v,p}=(x_v)^{p-1}$, $p=\overrightarrow{0,1}$, $v=\overrightarrow{1,N}$.

Using expressions (7) describing the optimal coefficients $k_{i,v}^{(p)}$, $i = \overrightarrow{1,\ 3}$ based on (4) and (5) we can form the variational matrix of PMM-estimators for

$$\boldsymbol{V}_{(3)} = \frac{\mu_2\mu_6 - \mu_4^2}{9\mu_2^3 - 6\mu_2\mu_4 + \mu_6}\left[\boldsymbol{B}\boldsymbol{B}^T\right]^{-1} = \kappa_2\left(1 - \frac{\gamma_4^2}{6 + 9\gamma_4 + \gamma_6}\right)\left[\boldsymbol{B}\boldsymbol{B}^T\right]^{-1} \tag{11}$$

where $\gamma_4 = \frac{\mu_4}{\mu_2^2} - 3$, $\gamma_6 = \frac{\mu_6}{\mu_2^3} - 15\frac{\mu_4}{\mu_2^2} + 30$ – dimensionless cumulant coefficients.

The transition in the expressions (11) from the moment description to the cumulant one is caused not only by the greater compactness of the latter, but also by the fact that the deviation of the values of higher-order cumulant coefficients $\gamma_r = \kappa_r/\kappa_2^{r/2}$ from zero describes the degree of difference from the Gaussian model. This allows us to interpret the results more clearly. It is obvious from (11) that the values of the variance reduction coefficients $g_{(\theta_p)3}$ do not depend on the index p of the vector parameter component θ, and the potential decrease in the estimates variance is determined precisely by the degree of non-gaussness, expressed numerically by the cumulant coefficients of the 4th and 6th orders, e.g.:

$$g_{(\theta_p)3} = 1 - \frac{\gamma_4^2}{6 + 9\gamma_4 + \gamma_6},\ p = \overrightarrow{0,\ 1}. \tag{12}$$

It is necessary to mention the universality of formula (12), since it also describes the coefficient of decrease in the variances of PMM-estimators (at the degree $S = 3$) of the scalar parameter under the conditions of symmetric measurement errors [22, 23]. It should be noted that the coefficients of higher order cumulants of random variables are not arbitrary values, because their combination has the domain of admissible values [27]. For example, for symmetrically distributed random variables, probability properties of which are given by cumulant coefficients of the 4th and 6th order, the domain of admissible values of these parameters are limited to two inequalities: $\gamma_4 > -2$ and $\gamma_6 + 9\gamma_4 + 6 > \gamma_4^2$. Considered the last inequalities, from the analysis of (12), we can conclude that the coefficient $g_{(\theta_p)3}$ of variance reduction has the range $(0;\ 1]$ and is dimensionless. Figure 1 shows the graphs which are built with these limitations.

Since the reduction of variance coefficient $g_{(\theta_p)3}$ is a function of two variables γ_4 and γ_6, the set of their values will be a surface (Fig. 1a). For greater clarity, in addition to the 3D-graphic, projections of the contour plots on the plane are also presented. The darker regions in Fig. 1b correspond to the projections of large values of the function (12).

From these graphs, you can see that the variance of PMM-estimates greatly increases and tends asymptotically to zero when value of cumulant coefficients approaching to the border region bounded by a parabola $\gamma_6 = \gamma_4^2 - 9\gamma_4 - 6$. The reduction of estimate variances is not observed only in the case when the kurtosis coefficient is zero. In other cases (with $\gamma_4 \neq 0$) the PMM-estimates has the less variance than the OLS-estimates.

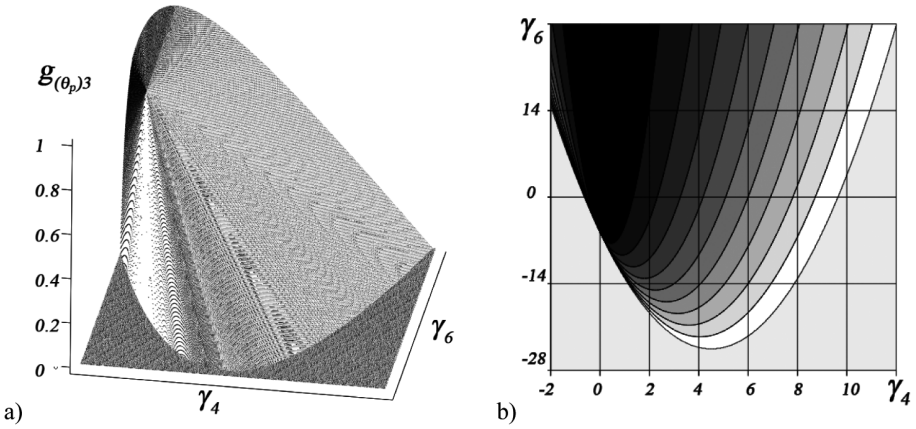

Fig. 1. Coefficient of reduction variance $g_{(\theta p)3}$ dependency on cumulative coefficients γ_4, γ_6: (a) 3D- graphic; (b) Contour plots.

7 Features of the Algorithm for Adaptive Polynomial Estimates

The results of the theoretical analysis of the effectiveness of polynomial estimates presented in the previous section, as well as the corresponding results of [20], testify to the advisability of using PMM only if the distribution of regression errors is not of the Gaussian law. It was also noted that the inadequacy of the Gaussian model of regression errors is not critical from the point of view that OLS estimates remain unbiased and consistent, although they cease to be optimal. In this case, the OLS is inherently linear, and, consequently, the probabilistic properties of the regression residues after its application do not actually differ from the properties of the original random component of the regression model [28]. Such a factor is often used to obtain adaptive estimates using the maximum likelihood method, which additionally requires solving the problems of estimating the error distribution density [7, 29].

Once again, we note that it is important from the practical point of view that to obtain PMM-estimates, information is used not about the distribution of errors, but their moment-cumulant description. Thus, a fairly simple way of overcoming a priori uncertainty about the probabilistic properties of the error model arises by finding estimates of a limited number of their parameters: moments or cumulants. Moreover, the obtained estimates of the cumulative coefficients of the 3rd (asymmetry) and the 4th (kurtosis) orders can also be used to test the hypothesis of Gaussianity and the symmetry of regression errors [30, 31]. On the above, we modify the approach proposed in [20], which allows us to find adaptive PMM estimates of the regression parameters in accordance with the following algorithm:

Step 1 – finding OLS estimates of regression parameters;
Step 2 – formation of regression residues and finding estimates of their moments and cumulants up to the 4th order;

Step 3 – testing the Gaussian regression residues distribution hypothesis (in the case of its non-refutation, the algorithm ends);
Step 4 – testing hypothesis about the regression residues distribution symmetry (in case of its non-refutation, the transition to step 6);
Step 5 – finding PMM estimates using the polynomial of degree $S = 2$ ((according to [20]) and completing the algorithm;
Step 6 – estimates moments of 6th orders of regression residuals of OLS;
Step 7 – finding PMM-estimates using the degree polynomial $S = 3$ (by a numerical solution of the system (8)) and completing the algorithm.

8 Monte-Carlo Simulation

Based on the results obtained, the set program (for MATLAB/OCTAVE), used in [20, 22–24], was modernized. This set of *m*-scripts and *m*-functions based on the Monte Carlo statistical simulation compare the accuracy of OLS and PMM-estimates at different models of the regression of non-Gaussian errors.

Table 1 presents the Monte Carlo simulation results.

Table 1. The results of Monte-Carlo parameters estimation simulation.

Distribution		Theoretical values			Monte-Carlo simulation results					
		γ_4	γ_6	$g_{(\theta_p)3}$	$\hat{g}_{(\theta_p)3}$					
					$N = 20$		$N = 50$		$N = 200$	
					a_0	a_1	a_0	a_1	a_0	a_1
Arcsines		−1.3	8.2	0.2	0.51	0.53	0.31	0.32	0.22	0.23
Uniform		−1.2	6.9	0.3	0.67	0.67	0.44	0.45	0.33	0.33
Trapezoidal	$\beta = 0.75$	−1.1	6.4	0.36	0.70	0.70	0.50	0.50	0.40	0.40
	$\beta = 0.5$	−1	5	0.55	0.85	0.84	0.68	0.69	0.57	0.58
	$\beta = 0.25$	−0.7	2.9	0.76	**1.01**	**1.00**	0.89	0.90	0.80	0.81
Triangular		−0.6	1.7	0.84	**1.08**	**1.09**	0.97	0.97	0.88	0.89
Laplace		3	30	0.86	**1.47**	**1.50**	0.86	0.87	0.85	0.85

As a comparative criterion of effectiveness, estimates of the magnitude of variance reduction coefficients $\hat{g}_{(\theta_p)3}$ are used, whose are determined according to (9). Analysis of the results of Table 1 shows that experimental values $\hat{g}_{(\theta_p)3}$ differ from theoretical values, since the analytical formula (12) was obtained for the asymptotic case (for $N \to \infty$). In addition, the uncertainty of a posteriori estimates of regression error parameters (paired moments up to the 6th order), which also depends significantly on the sample size N, affects the accuracy of obtaining adaptive PMM-estimates of the desired components of the parameter θ. But to a much greater extent, the relative effectiveness of PMM depends on the probabilistic properties of regression errors and

is more pronounced for distributions having a flat-topped character and negative values of the kurtosis coefficient. For example, for error models in which $\gamma_4 < -1$ the variance decreases from 15% (for small samples $N = 20$) to several times (with growth N or decrease γ_4).

In addition, in Figs. 2 and 3 examples of simulation results are given showing distributions of the experimental values of OLS and PMM-estimates (at $S = 3$) of components of vector parameters $\boldsymbol{\theta} = \{a_0, a_1\}$.

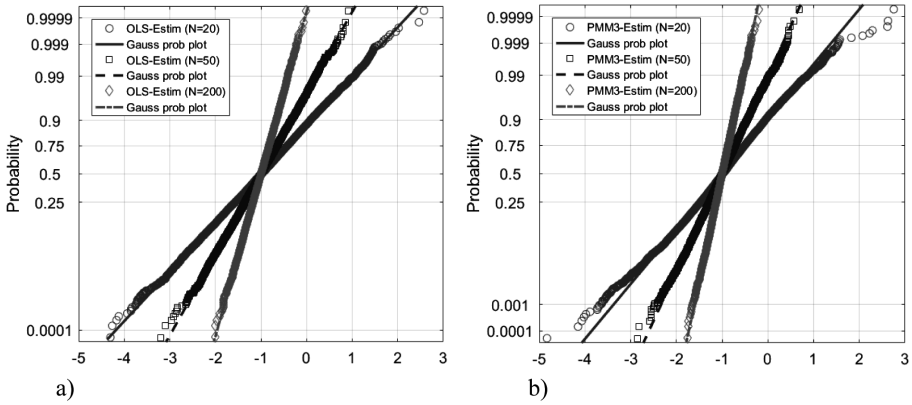

Fig. 2. Gaussian probability graphs approximating experimental values of the linear regression parameter $a_0 = -1$ of trapezoid error distribution: (a) OLS estimates; (b) PMM estimates.

In these examples the input data have the $M = 10^4$ samples and size ($N = 20, 50, 200$), containing the results estimation of the parameter $a_0 = -1$ and $a_1 = 2$ for model of errors based the random variable with Trapeze PDF ($\beta = 0.5$).

Analysis of these data and other results of experimental research shows that for the regression error, the distribution of which is a significantly different from the Gaussian model, the normalization of the distribution of estimates is observed only for a sufficiently large size of initial sample elements. This is explained by the presence of additional nonlinear transformations (calculation of quadratic and cubic statistics when PMM estimates are found for $S = 3$), which slows down the dynamics of the normalization of the empirical distribution of PMM-estimations relatively in OLS estimates.

9 Real Data Experiment

Let's test the proposed algorithm for finding the adaptive PMM estimations of linear regression parameters using a set of real data. This data set is atmospheric concentrations of CO_2 expressed in parts per million (ppm) reported in the preliminary 1997 SIO manometric mole derived from in situ air samples collected at Mauna Loa Observatory, Hawaii. [32]. For the experiment, part of this data was used by volume

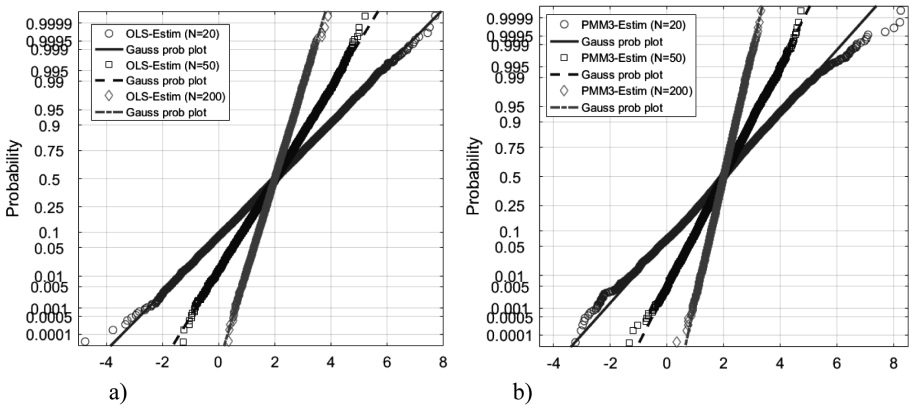

Fig. 3. Gaussian probability graphs approximating the experimental values of the linear regression parameter $a_1 = 2$ for the trapezoid error distribution: (a) OLS estimates; (b) PMM estimates.

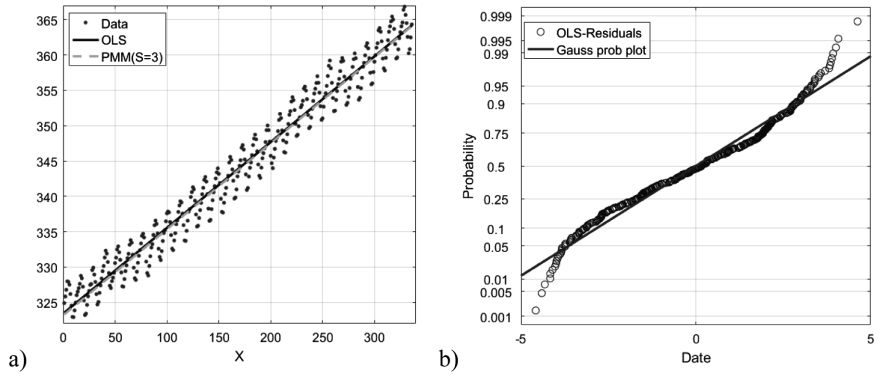

Fig. 4. Linear regression model (a) Experimental data and regression estimates; (b) Q-Q plot OLS-residuals.

$N = 336$ (monthly measurements from January 1970 to September 1997), which are described adequately (see Fig. 4a) by a linear regression model of the form (1).

The least-squares estimations of the parameters of such model are the values: $\hat{a}_0^{(1)} = 323.411$; $\hat{a}_1^{(1)} = 0.121$. S residuals is refuted by the Yarki-Ber test embedded in the MATLAB ($JBSTAT = 15.6$ at the threshold value $CV = 5.8$ at a fixed significance degree 0.05) [31]. The symmetry of the error distribution of the model is visually visible in Fig. 4(b), where the probabilistic graph of the Gaussian approximation (Q-Q plot) of the regression OLS residuals is presented and is confirmed by the small absolute value of the estimation of their asymmetry coefficient $\hat{\gamma}_3 = -0.15$.

Estimated values of even cumulant coefficients: $\hat{\gamma}_4 = -1$ and $\hat{\gamma}_6 = 5.2$ regression OLS residuals make it possible to determine (12) sufficiently accurately (taking into

account a sufficiently large volume of initial data N) to determine the value of the coefficient of variance reduction $\hat{g}_{(\theta_p)3} = 0.5$.

We note that the refined values of PMM estimates themselves ($\hat{a}_0^{(3)} = 323.156$; $\hat{a}_1^{(3)} = 0.122$) differ from the values of OLS estimators not significantly. However, a significant decrease in the variance allows us to build more narrow confidence integrals, which are usually used in regression analysis. Considering the observance of the condition for the normalization of the distribution of OLS estimates (for large N), we get that at a confidence level of 0.95, the interval $(322.936;\ 323.886)$ covers the value of the regression parameter a_0, and the interval $(0.119;\ 0.124)$ – covers the value parameter a_1. At the same time, taking into account the value of the estimation of the coefficient of variance reduction, as well as the obtained point values of PMM- estimations, allows for a given level of confidence probability, it is justified to correct and reduce $\left(1 - \sqrt{0.5}\right)\ 100 \approx 30\ \%$ the width of the confidence interval, having obtained its limits $(322.82;\ 323.491)$ for the parameter a_0, and $(0.12;\ 0.123)$ for the parameter a_1.

To verify the correctness of the results, a statistical experiment based on the bootstrapping method. With the help of built-in MATLAB bootstrap-resampling tools the original sample was multiplied by 10^4 bootstrap samples with the return. For each of them, we found OLS and PMM estimates for linear regression parameters. The empirical distribution of these estimates is presented in Fig. 5 in the form of boxplot-graphs, the upper and lower bounds of which are respectively 2.5% and 97.5% percentile.

The visual analysis of Fig. 5 shows that boundaries of the 95% confidence intervals obtained as the result of statistical bootstrap modeling, and correlate with the calculated values of the interval estimates obtained above indicate their significant closeness. This generally confirms the reliability of the analytical calculations.

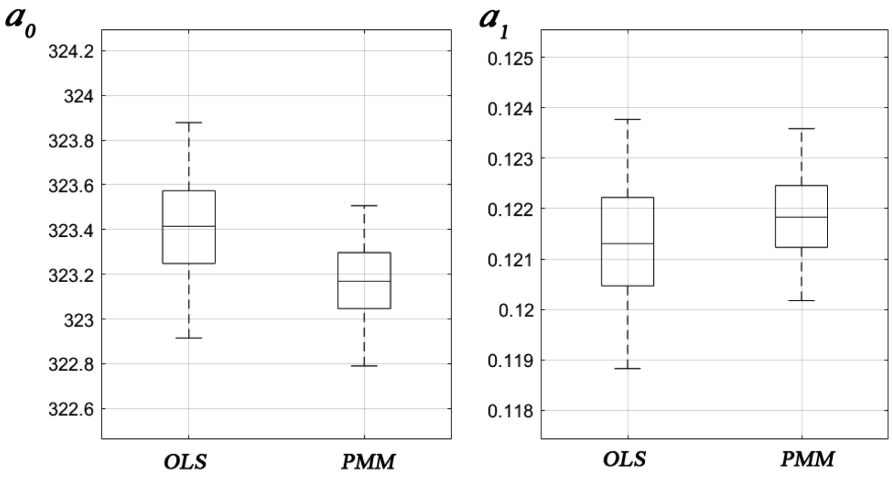

Fig. 5. Empirical (bootstrap) distribution of linear regression parameter estimates.

10 Conclusions

An analysis of the set of obtained results confirms the possibility and expediency of applying the method of polynomial maximization for finding estimates of single-factor linear regression parameters under the condition of a non-zero-symmetric distribution of errors, which are described using higher-order statistics.

Theoretical studies have shown that nonlinear PMM-estimates synthesized at a polynomial degree $S = 3$ are characterized by the greater accuracy than OLS-estimates. The coefficient of variance decrease is determined by the degree of non-Gaussianity of the random component of the regression model, expressed numerically by the absolute values of the cumulative coefficients of the 4th and 6th orders.

The results obtained in statistical modeling by the Monte Carlo method confirm the effectiveness of PMM (for $S = 3$) for situations where errors of the regression model have a symmetric distribution. The greatest efficiency (relative to OLS) of PMM is appears in situations where the kurtosis of regression errors has negative values, which is observed, for example in flat-top and double-modal distributions. For example, for model experimental data with the arcsine distribution, the variance of PMM estimates compared to OLS estimates is 5 times less. And for the considered example of real experimental data (CO_2 concentration in the atmosphere), the variance of estimates of the regression model parameters decreases by about 2 times. This made possible to reduce the width of their confidence intervals by 30%.

The asymptotic nature of PMM is confirmed in [20], since with the increase in the sample data the experimental values of the variance reduction coefficients tend to theoretically calculated values, and the distribution of the estimates is normalized.

In general, the proposed approach has less analytical and computational complexity than the parametric MLE and provides a reduction in uncertainty in comparison with OLS, which does not consider the difference in the probability distribution of statistical data from the Gaussian law.

Among many possible directions of further research, one should mention the following:

- carry out a comparative analysis of the efficiency of PMM with estimates of alternative nonparametric methods (least absolute deviation, sign, etc.);
- extend the proposed approach and explore the features of its application for non-linear single-factor regression models, and consider a more general multifactor case;
- investigate the possibility of approximating the distribution of PMM estimates, models based on Johnson curves for constructing confidence intervals for small sample sizes.

References

1. Anscombe, F.J.: Topics in the investigation of linear relations fitted by the method of least squares. J. R. Stat. Soc. Ser. B (Methodological) **29**, 1–52 (1967)
2. Cox, D.R., Hinkley, D.V.: A note on the efficiency of least-squares estimates. J. R. Stat. Soc. Ser. B (Methodological) **30**, 284–289 (1968)

3. Schechtman, E., Schechtman, G.: Estimating the parameters in regression with uniformly distributed errors. J. Stat. Comput. Simul. **26**(3–4), 269–281 (1986). https://doi.org/10.1080/00949658608810965
4. Galea, M., Paula, G.A., Bolfarine, H.: Local influence in elliptical linear regression models. J. R. Stat. Soc. Ser. D: Stat. **46**(1), 71–79 (1997)
5. Liu, S.: Local influence in multivariate elliptical linear regression models. Linear Algebra Appl. **354**(1–3), 159–174 (2002). https://doi.org/10.1016/S0024-3795(01)00585-7
6. Ganguly, S.S.: Robust regression analysis for non-normal situations under symmetric distributions arising in medical research. J. Modern Appl. Stat. Meth. **13**(1), 446–462 (2014). https://doi.org/10.22237/jmasm/1398918480
7. Zeckhauser, R., Thompson, M.: Linear regression with non-normal error terms. Rev. Econ. Stat. **52**(3), 280–286 (1970)
8. Bartolucci, F., Scaccia, L.: The use of mixtures for dealing with non-normal regression errors. Comput. Stat. Data Anal. **48**(4), 821–834 (2005). https://doi.org/10.1016/j.csda.2004.04.005
9. Seo, B., Noh, J., Lee, T., Yoon, Y.J.: Adaptive robust regression with continuous Gaussian scale mixture errors. J. Korean Stat. Soc. **46**(1), 113–125 (2017). https://doi.org/10.1016/j.jkss.2016.08.002
10. Tiku, M.L., Islam, M.Q., Selçuk, A.S.: Non-normal regression II. Symmetric distributions. Commun. Stat. Theory Meth. **30**(6), 1021–1045 (2001). https://doi.org/10.1081/STA-100104348
11. Andargie, A.A., Rao, K.S.: Estimation of a linear model with two-parameter symmetric platykurtic distributed errors. J. Uncertaint. Anal. Appl. **1**(1), 1–19 (2013)
12. Atsedeweyn, A.A., Srinivasa Rao, K.: Linear regression model with generalized new symmetric error distribution. Math. Theory Model. **4**(2), 48–73 (2014). https://doi.org/10.1080/02664763.2013.839638
13. Huber, P.J., Ronchetti, E.M.: Robust Statistics. Wiley, Hoboken (2009). https://doi.org/10.1002/9780470434697
14. Narula, S.C., Wellington, J.F.: The minimum sum of absolute errors regression: a state of the art survey. Int. Stat. Rev. **50**(3), 317–326 (1982)
15. Koenker, R., Hallock, K.: Quantile regression: an introduction. J. Economic. Perspect. **15**(4), 43–56 (2001)
16. Tarassenko, P.F., Tarima, S.S., Zhuravlev, A.V., Singh, S.: On sign-based regression quantiles. J. Stat. Comput. Simul. **85**(7), 1420–1441 (2015). https://doi.org/10.1080/00949655.2013.875176
17. Dagenais, M.G., Dagenais, D.L.: Higher moment estimators for linear regression models with errors in the variables. J. Econom. **76**(1–2), 193–221 (1997). https://doi.org/10.1016/0304-4076(95)01789-5
18. Cragg, J.G.: Using higher moments to estimate the simple errors-in-variables model. RAND J. Econ. **28**, S71 (1997). https://doi.org/10.2307/3087456
19. Gillard, J.: Method of moments estimation in linear regression with errors in both variables. Commun. Stat. Theory Meth. **43**(15), 3208–3222 (2014)
20. Zabolotnii, S., Warsza, Z., Tkachenko, O.: Polynomial estimation of linear regression parameters for the asymmetric pdf of errors. In: Advances in Intelligent Systems and Computing. vol. 743, pp. 758–772. Springer (2018). https://doi.org/10.1007/978-3-319-77179-3_75
21. Kunchenko, Y.: Polynomial Parameter Estimations of Close to Gaussian Random variables. Shaker Verlag, Aachen (2002)

22. Warsza, Z.L., Zabolotnii, S.W.: A polynomial estimation of measurand parameters for samples of non-Gaussian symmetrically distributed data. In: Advances in Intelligent Systems and Computing, vol. 550, pp. 468–480. Springer (2017). http://doi.org/10.1007/978-3-319-54042-9_45
23. Warsza, Z.L., Zabolotnii, S.W.: Uncertainty of measuring data with trapeze distribution evaluated by the polynomial maximization method. Przemysł Chemiczny **1**(12), 68–71 (2017). https://doi.org/10.15199/62.2017.12.6. (in Polish)
24. Warsza, Z., Zabolotnii, S.: Estimation of measurand parameters for data from asymmetric distributions by polynomial maximization method. In: Advances in Intelligent Systems and Computing, vol. 743, pp. 746–757. Springer (2018). https://doi.org/10.1007/978-3-319-77179-3_74
25. Zabolotnii, S.W., Warszam, Z.L.: Semi-parametric estimation of the change-point of parameters of non-Gaussian sequences by polynomial maximization method. In: Advances in Intelligent Systems and Computing, vol. 440, pp. 903–919. Springer (2016). http://doi.org/10.1007/978-3-319-29357-8_80
26. Palahin, V., Juh, J.: Joint signal parameter estimation in non–Gaussian noise by the method of polynomial maximization. J. Electr. Eng. **67**, 217–221 (2016). https://doi.org/10.1515/jee-2016-0031
27. Cramér, H.: Mathematical Methods of Statistics, vol. 9. Princeton University Press, Princeton (2016)
28. Cook, R.D., Weisberg, S.: Residuals and Influence in Regression. Monographs on Statistics and Applied Probability. Chapman and Hall, New York (1982). https://doi.org/10.2307/1269506
29. Stone, C.J.: Adaptive maximum likelihood estimators of a location parameter. Annal. Stat. **3**(2), 267–284 (1975). https://doi.org/10.1214/aos/1176343056
30. Boos, D.D.: Detecting skewed errors from regression residuals. Technometrics **29**(1), 83–90 (1987). https://doi.org/10.1080/00401706.1987.10488185
31. Jarque, C.M., Bera, A.K.: A test for normality of observations and regression residuals. Int. Stat. Rev. **55**(2), 163–172 (2012)
32. Keeling, C.D., Whorf, T.P.: Scripps Institution of Oceanography (SIO). University of California, La Jolla, California USA 92093-0220. ftp://cdiac.esd.ornl.gov/pub/maunaloa-co2/maunaloa.co2

Author Query Form

Book ID : **471799_1_En**

Chapter No : **59**

Please ensure you fill out your response to the queries raised below and return this form along with your corrections.

Dear Author,
During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query Refs.	Details Required	Author's Response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	
AQ2	Per Springer style, both city and country names must be present in the affiliations. Accordingly, we have inserted the city name "Cherkasy" in affiliations 1. Please check and confirm if the inserted city name is correct. If not, please provide us with the correct city name.	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	··· under matter to remain	New matter followed by
Insert in text the matter indicated in the margin		or (x2)
Delete	through single character, rule or underline or through all characters to be deleted	or (x2)
Substitute character or substitute part of one or more word(s)	through letter or through characters	new character or new characters
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	under matter to be changed	
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	
Insert comma	(As above)	
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	through character or where required	
Reduce space between characters or words	between characters or words affected	