Polynomial Estimation of Linear Regression Parameters for the Asymmetric PDF of Errors

Chapter *in* Advances in Intelligent Systems and Computing · March 2018

DOI: 10.1007/978-3-319-77179-3_75

Serhii Zabolotnii
Cherkasy State Business College
46 PUBLICATIONS **99** CITATIONS

Zygmunt Lech Warsza
Łukasiewicz Research Network - Industrial Research Institute for Automation and…
130 PUBLICATIONS **325** CITATIONS

Polynomial Estimation of Linear Regression Parameters for the Asymmetric PDF of Errors

Serhii Zabolotnii[1], Zygmunt Lech Warsza[2](✉),
and Oleksandr Tkachenko[3]

[1] Cherkasy State Technological University, Cherkasy, Ukraine
zabolotni@ukr.net
[2] Industrial Research Institute for Automation and Measurements PIAP,
Al. Jerozolimskie 202, 02-486 Warsaw, Poland
zlw1936@gmail.com
[3] Cherkasy State Business College, Cherkasy, Ukraine
tkachenko.ck@gmail.com

Abstract. This paper presents a non-standard way of finding estimates of linear regression parameters for the case of asymmetrically distributed errors. This approach is based on the polynomial maximization method (PMM) and uses the moment and cumulant description of random variables. Analytic expressions are obtained that allow one to find estimates and analyze their accuracy for the degree of the polynomial $S = 1$ and $S = 2$. It is shown that the variance of polynomial estimates (for $S = 2$) in the general case is less than the variance of estimates of the ordinary least squares method, which is a particular case of the polynomial maximization method (for $S = 1$). The increase in accuracy depends on the values of cumulant coefficients of higher orders of random errors of regression. Statistical modeling (Monte Carlo & bootstrapping method) is performed, the results of which confirm the effectiveness of the proposed approach.

Keywords: Linear regression · Asymmetric errors · Stochastic polynomial
Variance · Skewness and kurtosis

1 Introduction

The use of regression models and methods is one of the most common statistical approaches for the analysis of data in various fields: economics, sociology, medicine, ecology, geology, astronomy, technical systems, etc. The main purpose of the regression analysis is to establish cause-effect relationships in the form of a certain deterministic function that describes the relation between the values of the goal (depended) variable and one or more independent variables (predictors). As a rule, if the initial data is stochastic, an obligatory component of the regression models is a random error. From the statistical point of view, the main task of regression analysis is to find estimates of the parameters of the deterministic function for which the extreme of the selected quality criterion is reached (usually this is the minimum of the root-mean-square deviations).

R. Szewczyk et al. (Eds.): AUTOMATION 2018, AISC 743, pp. 758–772, 2018.
https://doi.org/10.1007/978-3-319-77179-3_75

Under certain conditions of the Gauss-Markov theorem, the ordinary least-squares (OLS) method is optimal. In the case of a Gaussian character (normal distribution) of the random errors of the regression model, the OLS-estimations are characterized by the smallest variance. However, a significant part of modern research related to the processing of real data indicates that the Gaussian model is only a convenient idealization, which only makes it possible to significantly simplify the calculations [1].

There are various conceptual approaches to overtake "non-Gaussian" regression errors. One of them is based on the use of robust or non-parametric procedures, in particular, robust versions of least-squares method, the application of which is primarily aimed at ensuring stability against the influence of extreme deviations (outliers) [2].

However, there are a number of situations, for example, in biological [3] or technical [4] systems, when the difference from the Gaussian model is generated not by "erroneous" single outliers, but by the specifics of the data. In this case, the approach based on the normalizing transformations is used: Box-Cox [5], Johnson [6], etc. This approach allows transforming the original data to make the transition to the Gaussian error model, which, even with small sample sizes, makes it possible to test hypotheses and calculating confidence intervals using traditional Student t-statistics. In cases where one of the main criteria is the minimization of the uncertainty (variance) of estimates of parameters of regression models, a parametric approach based on the maximum likelihood method is used [7–9]. From a computational point of view, such an 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.

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 (moments and/or cumulants). Examples of its use for solving regression analysis problems are the papers [10–14]. It should be noted that the moment & cumulant description is generally used in combination with the method of moments (MM), which is characterized by computational simplicity, but relatively low statistical accuracy. Therefore, in this paper we consider a new way of finding estimates of regression model parameters for a moment & cumulant description of random non-Gaussian errors, based on the use of the original statistical estimation method – the Polynomial Maximization Method (PMM) [15].

2 The Purpose of Research

In this paper the application of a new unconventional approach in solving problems of finding estimates of regression's parameters for non-Gaussian errors is considered. This approach applied the Kunchenko's stochastic polynomials as the mathematical tool [16]. Examples of the use of mathematical apparatus in a variety of areas related to statistical data processing are pattern recognition based on template matching technology [17], probabilistic diagnosis of disorder (change-point problem) [18, 19], detection of signals against the non-Gaussian noise and parameter estimation [20, 21]. The results obtained show that the use of the apparatus of stochastic polynomials in conjunction with the construction of probabilistic models based on the statistics of higher orders (moment & cumulant description) greatly simplifies the adaptive methods for the synthesis process. The probability of these properties can by taking into account for improving the accuracy of processing (to reduce the probability of erroneous decisions, the variance of parameter estimates, etc.).

It should be mentioned that for the first time the possibility of using the modification of the PMM for solving the regression analysis problems developed for estimating the vector parameter for unequally distributed random values first time was declared in 1992 [15]. The author of this paper theoretically stated that for a non-Gaussian character of the statistical data, PMM estimates as a whole are characterized by a lower uncertainty (relative to the OLS), and the magnitude of the decrease of variance depends on the values of cumulant coefficients of higher orders.

The aim of below study is to synthesize computational algorithms for finding adaptive PMM-estimations, and also to analyze their properties by statistical Monte Carlo simulations using the model of simple linear regression with asymmetrically distributed errors.

3 Mathematical Formulation of the Problem

Let's have a one-factor linear regression observation model that describes the dependence of the goal variable y on the predictive variable x:

$$y_v = a_0 + a_1 x_v + \xi_v,\ v = \overrightarrow{1,N}, \tag{1}$$

where ξ_v – sequence of centered $(E\{\xi\} = 0)$ – independent and identically distributed random variables (model of errors).

The probabilistic properties of the random component of the model differ from the Gaussian law (they have an asymmetric character of the distribution) and are described by a sequence of cumulants κ_r and/or cumulant coefficients γ_r. On this description, the second-order cumulant $\kappa_r = m_2$ determines the variance of the random component of the errors, and the cumulant coefficients of higher-order $\gamma_r = \kappa_r / \kappa_2^{r/2}$ numerically describe the degree of difference from the Gaussian distribution.

The problem consists in finding the parameter estimates a_0 and 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 a Vector Parameter by the Polynomial Maximization Method (PMM)

Conceptually, the maximization method of the polynomial (PMM) is close to the maximum likelihood (MMP) method. The main analogy is that both methods are based on using some statistical functional from the sample values, which should have a maximum in the neighborhood of the true value of the estimated parameter. The principal difference is that the use of MMP requires a complete description of the random variables in the form of the probability distribution density, PMM – not, only few first moments m_v or cumulants κ_r.

To solve the problem, we use the variant of the maximization method of the polynomial, developed for the case of finding estimates of a vector parameter for unequally distributed statistical data [15]. According to the PMM, the estimate of a vector parameter $\boldsymbol{\theta}$ from dimension Q can be found as the solution of the set of Q stochastic power equations:

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

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

Coefficients $k_{i,v}^{(p)}$ (for each component 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 $\boldsymbol{\theta}$:

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

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

Set of equations (3) can be solved analytically using the Kramer method.

In [15] it was shown that PMM-estimations, which are the solutions of set 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 parameter $\boldsymbol{\theta}$, which generally are described by the equation:

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

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

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

obtained by inversion of a quadratic (dimension Q) extracted data amount matrix $\boldsymbol{J}_{(S)}$, consisting of elements $J_{SN}^{(p,q)}$ of the form (4).

5 Polynomial Estimation of Linear Regression Parameters

5.1. Using the general formula (2), we discover that for a polynomial degree $S = 1$ PMM-estimating the vector parameter $\boldsymbol{\theta} = \{a_0, a_1\}$ of the linear regression model (1) can be found from the solution of two equations of the form:

$$\sum_{v=1}^{N} \left\{ k_{1,v}^{(p)} [y_v - (a_0 + a_1 x_v)] \right\} = 0, \ p = \overrightarrow{0, 1}, \tag{6}$$

where, according to (3), index $k_{i,v}^{(p)} = \frac{(x_v)^{p-1}}{\kappa_2}, \ p = \overrightarrow{0, 1}$.

Solving the system (6) we obtain analytical expressions describing estimates of the required parameters in the form:

$$\hat{a}_0 = \frac{1}{N} \left[\sum_{v-1}^{N} y_i - \hat{a}_1 \sum_{v-1}^{N} x_i \right], \ \hat{a}_1 = \frac{N \sum\limits_{v-1}^{N} y_i x_i - \sum\limits_{v-1}^{N} y_i \sum\limits_{v-1}^{N} x_i}{N \sum\limits_{v-1}^{N} (x_i)^2 - \left(\sum\limits_{v-1}^{N} x_i \right)^2}. \tag{7}$$

It should be noted that obtained estimates (7) of the linear regression parameters completely coincide with the linear OLS estimates, which are optimal (by the criterion of the minimum dispersion) for the situation when the errors of the regression model have a Gaussian distribution [1]. If the error distribution differs from a Gaussian law, there are alternative methods of finding estimates, based on nonlinear transformations, which can provide a reduction of the resulting variance of estimates. Below we consider a new method for constructing the non-linear estimates of the parameter, which is based on the use of power polynomials [15, 16].

5.2. When stochastic polynomials the degrees $S = 2$ are used, PMM-estimating the sought vector parameter $\boldsymbol{\theta} = \{a_0, a_1\}$ can be found from the solution of a system of two equations formed on the basis of the general formula (2):

$$\sum_{v=1}^{N} \left\{ k_{1,v}^{(p)} [y_v - (a_0 + a_1 x_v)] + k_{2,v}^{(p)} \left[(y_v)^2 - (a_0 + a_1 x_v)^2 - \kappa_2 \right] \right\} = 0, \ p = \overrightarrow{0, 1}, \tag{8}$$

where two pairs of optimal coefficients $k_{i,v}^{(p)}, \ i = \overrightarrow{1, 2}$ ensure the minimization of the variance of the estimates of the sought parameter when this degree of polynomial is used [15, 16]. They are found by solving set of two linear equations of the form (3) and can be described by expressions:

$$k_{1,v}^{(p)} = \frac{2\gamma_3 (a_0 + a_1 x_v) + \kappa_2^{1/2} (2 + \gamma_4)}{\kappa_2^{3/2} \left(2 - \gamma_3^2 + \gamma_4 \right)} (x_v)^{p-1}, \quad k_{2,v}^{(p)} = -\frac{\gamma_3}{\kappa_2^{3/2} \left(2 - \gamma_3^2 + \gamma_4 \right)} (x_v)^{p-1} \tag{9}$$

where $p = \overrightarrow{0, 1}$.

Substituting the coefficients (9) in (8), after simple transformations, the system of equations for finding the parameter $\boldsymbol{\theta}$ estimates can be represented in the form:

$$\begin{aligned}\sum_{v=1}^{N}(x_v)^{p-1}\Big\{\gamma_3(a_0+a_1x_v)^2-\Big[2\gamma_3y_v-\kappa_2^{1/2}(2+\gamma_4)\Big](a_0+a_1x_v)\\+\Big[\gamma_3(y_v)^2-\kappa_2^{1/2}(2+\gamma_4)y_v-\kappa_2\gamma_3\Big]\Big\},\ p=\overrightarrow{0,\,1}\end{aligned}\tag{10}$$

Analysis of this expression shows that in the case of the symmetry of the distribution (or at least equality $\gamma_3 = 0$) of the random component of errors of the regression model, the system of equations (10) degenerates into a linear system analogous to (6), i.e. PMM-estimations coincide with OLS estimates.

Obviously, for a polynomial of degree, $S = 2$ PMM-estimates can only be found by means of a numerical solution of the systems of equations (10). As a first approximation, using the appropriate iterative procedures, it is logical to use least-squares estimators of the form (7).

6 Analysis of the Accuracy of Polynomial Estimates

To quantitatively calculation the values of changes in the accuracy of estimates, we use the notion of a coefficient for reducing the variance:

$$g_{(\theta_p)S}=\frac{\sigma^2_{(\theta_p)S}}{\sigma^2_{(\theta_p)OLS}}=\frac{\sigma^2_{(\theta_p)S}}{\sigma^2_{(\theta_p)1}},\ p=\overrightarrow{0,\,Q-1}.\tag{11}$$

These coefficients are formed as variance ratios of PMM parameter estimates $\boldsymbol{\theta}$, found using a polynomial of the S-th order for variances of the least-squares estimations of the corresponding components of this vector parameter. Since OLS estimators are equivalent to PMM estimates obtained for a power, their variances will also coincide, i.e. $\sigma^2_{(\theta_p)OLS}=\sigma^2_{(\theta_p)1}$.

Using the relations (4), which determine the amount of extracted information about the components of the estimated parameter $\boldsymbol{\theta}$, taking into account the coefficients $k_{1,v}^{(p)}$ from the system of equations (6), on the basis of the general expression (5), we write the variation matrix estimations for $S = 1$:

$$\boldsymbol{V}_{(1)}=\kappa_2\left[\boldsymbol{B}\boldsymbol{B}^T\right]^{-1},\tag{12}$$

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

Similarly, using expressions (9) describing the optimal coefficients $k_{i,v}^{(p)}$, $i=\overrightarrow{1,\,2}$, we can write the variation matrix estimations for $S = 2$:

$$\boldsymbol{V}_{(2)} = \kappa_2 \left(1 - \frac{\gamma_3^2}{2 + \gamma_4}\right) \left[\boldsymbol{B}\boldsymbol{B}^T\right]^{-1}, \tag{13}$$

Since the variances of PMM-estimators lie on the main diagonal of the matrix $\boldsymbol{V}_{(S)}$, the value of the dispersion $g_{(\theta_p)S}$ does not depend on the index p of the component of the vector parameter $\boldsymbol{\theta}$. Thus, in the case of $S = 2$ a potential reduction in the variance of estimates, is determined exclusively by the degree of non-Gaussness, expressed numerically by the values of the cumulant coefficients of skewness and kurtosis, i.e.

$$g_{(\theta_p)2} = 1 - \frac{\gamma_3^2}{2 + \gamma_4}, \; p = \overrightarrow{0, \; Q-1}. \tag{14}$$

It is necessary to note the universality of formula (14), since it also describes the coefficient of reduction in the variances of PMM-estimates (at a power $S = 2$) of the scalar parameter under the conditions of asymmetric measurement errors [16].

It should be noted that the coefficients of higher order cumulant of random variables are not arbitrary values, because their combination has the domain of admissible values [22]. For example, for random variables, probability properties of which are given by cumulant coefficients of the 3-rd and 4-th order, the domain of admissible values of this parameters are limited inequality: $\gamma_4 + 2 \geq \gamma_3^2$. Taken into account this inequality, from the analysis of (14), we can conclude that the coefficient of variance reduction $g_{(\theta)2}$ is dimensionless and belongs to the range $(0; 1]$.

Figure 1 shows the graphs built with this inequality. These graphs visualize dependences of coefficient $g_{(\theta_p)2}$. They are constructed as sections of two variables function of the form (14) and are dependent on one of the parameters (skewness γ_3) with fixed values of the other (kurtosis γ_4).

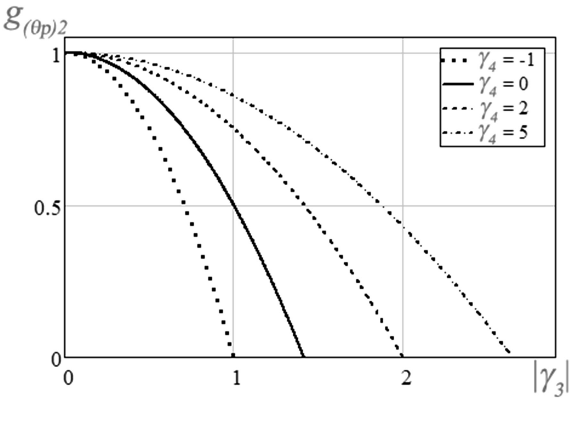

Fig. 1. Coefficient of reduction variance (for $S = 2$) dependency on cumulative coefficients γ_3, γ_4.

7 Adaptability of Procedures for Polynomial Estimation

Recall that when calculating PMM-estimates, a priori information is taken into account not about the type of error distribution of the regression model, but about the values of cumulants (cumulant coefficients). In the model experiment, these values can be easily obtained on the basis of analytical expressions connecting the parameters of the distribution densities with their moments and the corresponding cumulants. However, in real situations, information about the probabilistic properties of errors of the regression model, as a rule, is absent. Therefore, in this paper we propose an approach that makes it possible to implement an adaptive learning procedure quite simply. It is based on the fact that the inadequacy of the Gaussian hypothesis with respect to the error model is not critical from the point of view that OLS-estimates remain unbiased and consistent (although they cease to be effective). And since least-square method is inherently a linear method, the probabilistic properties of regression residues after its application do not actually differ from the properties of the original random component of the regression model [23].

Thus, the adaptive procedure for obtaining PMM-estimates will consist of the following steps:

Step 1: finding OLS estimates $\hat{\boldsymbol{\theta}}^{(1)}$ and forming a sequence of regression residues ξ'_v, $v = \overrightarrow{1,N}$;

Step 2: finding estimates of necessary moments and cumulants (up to the $2S$-th order) regression OLS-residuals;

Step 3: finding updated PMM-estimates $\hat{\boldsymbol{\theta}}^{(S)}$ using numerical methods for solving systems of equations.

When using stochastic polynomials, the degree $S = 2$ estimating the necessary parameters is easy enough to obtain on the basis of relations:

$$\hat{\kappa}_2 = \hat{m}_2,\ \hat{\gamma}_3 = \hat{m}_3 \Big/ \hat{m}_2^{3/2},\ \hat{\gamma}_4 = \hat{m}_4 \big/ \hat{m}_2^2 - 3 \tag{15}$$

where: $\hat{m}_i$ – selective central moment of i-th order:

$$\hat{m}_i = \frac{1}{N}\sum_{v=1}^{N}\left(\xi'_v - \overline{\xi'}\right)^i. \tag{16}$$

8 Statistical Modeling of Polynomial Estimation

Based on the results obtained, in the MATLAB software environment, the software package described in [21]. Now it additionally allows for the statistical modeling of algorithms for estimating the parameters of linear regression using the non-Gaussian error model. This complex is based on multiple tests (Monte Carlo and bootstrapping method), allows for comparative analysis of the accuracy of various algorithms for statistical estimation, and also investigates the properties of polynomial estimates.

Since the value of the coefficient of reduction of the variance, described by the expression (14), which is obtained for the limiting case (for $N \to \infty$), the estimates $\hat{g}_{(\theta_p)2}$, is used as a comparative efficiency criterion, will differ from the theoretical values. Obviously, the degree of divergence will depend on the volume of the original sample. The set of experimental values of the variance reduction coefficients, obtained on the basis of the Monte Carlo method, are presented in Table 1. These data were obtained with $M = 10^4$ multiple tests for various types of asymmetric random error distributions of the regression model. In this case, the values of the parameters necessary for finding the adaptive PMM-estimates were considered a priori unknown, and a posteriori estimates of the form (15) were used. To find the refined values of adaptive PMM-estimations by solving systems of equations of the form (10), a numerical Newton-Raphson procedure is used [24].

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

Distribution		Theoretical values			Monte-Carlo simulation					
		γ_3	γ_4	$g_{(\theta_p)2}$	$\hat{g}_{(\theta_p)2}$					
					$n = 20$		$n = 50$		$n = 200$	
					a_0	a_1	a_0	a_1	a_0	a_1
Gamma	$\alpha = 0.5$	2.83	12	0.43	0.53	0.38	0.51	0.35	0.48	0.33
Exponential (Gamma, $\alpha = 1$)		2	6	0.5	0.66	0.56	0.59	0.46	0.56	0.42
Gamma	$\alpha = 2$	1.41	3	0.6	0.75	0.68	0.68	0.59	0.67	0.57
	$\alpha = 4$	1	1.5	0.71	0.88	0.85	0.81	0.74	0.78	0.71
Lognormal $\sigma^2 = 0.1$, $\mu = 1$		1	1.86	0.74	0.89	0.85	0.82	0.76	0.79	0.73
Weibull $a = 1$, $b = 2$		0.63	0.25	0.82	0.97	0.96	0.91	0.88	0.87	0.83

Analysis of the data given in Table 1 shows a significant correlation between the results of analytical calculations and statistical modelling. It is obvious that with an increase of the initial sample size N, the discrepancy between the theoretical and experimental values of variance reduction coefficient (even for small values $N = 20$ difference does not exceed 20%) is decreasing.

Another important result of the statistical modeling is the test of assumption that with increasing N the distribution of PMM-estimates asymptotically approaching to the Gaussian law. The validity of the Gaussian distribution of PMM-estimates of hypotheses was investigated by using Lilliefors test based on Kolmogorov-Smirnov statistics [25] which is built-in MATLAB. Table 2 presents the test results as a series output parameters of the Lilliefors test. $LSTAT$ – sample value of the statistic test; CV – the critical value of the test statistic. If $LSTAT < CV$, the null hypothesis for a given critical level is valid. The results in Table 2 are obtained for different types of error distributions of the regression model and size N of the sample data at a fixed significance level $\alpha_0 = 0.05$ zero (Gaussian) hypothesis and $M = 10^4$ experiments.

Table 2. The result of testing the adequacy of the Gaussian distribution PMM-estimates (for $S = 2$) on the basis of Lilliefors test.

Distribution		Output parameters of Lilliefors test						
		LSTAT						*CV*
		$n = 20$		$n = 50$		$n = 200$		
		a_0	a_1	a_0	a_1	a_0	a_1	
Gamma	$\alpha = 0.5$	0.062	0.057	0.041	0.019	0.023	0.009	0.009
Exponential (Gamma, $\alpha = 1$)		0.051	0.039	0.026	0.014	0.021	0.008	
Gamma	$\alpha = 2$	0.031	0.022	0.021	0.009	0.009	0.007	
	$\alpha = 4$	0.028	0.016	0.016	0.007	0.008	0.006	
Lognormal $\sigma^2 = 0.1, \mu = 1$		0.021	0.016	0.014	0.006	0.008	0.005	
Weibull $a = 1$, $b = 2$		0.019	0.013	0.013	0.006	0.007	0.004	

In addition, as of the simulation results in Figs. 2 and 3 are examples showing the distribution of the experimental values of OLS and PMM-estimates (at $S = 2$) of components of vector parameters $\boldsymbol{\theta} = \{a_0, a_1\}$. 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 $a_1 = 3$ for model of errors based the random variable with an Gamma distribution ($\alpha = 2$).

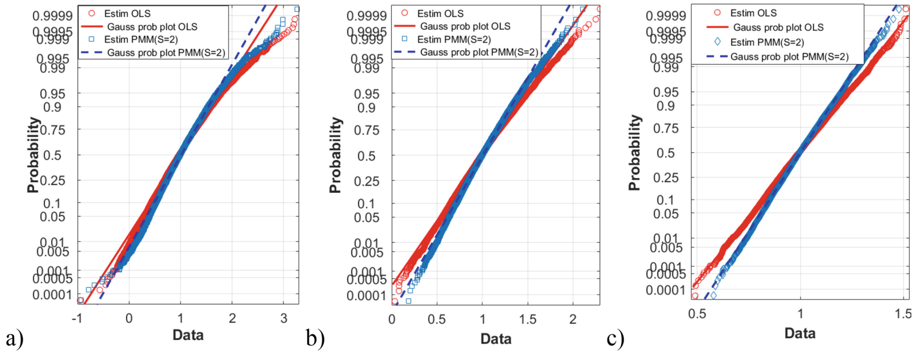

Fig. 2. Gaussian probabilistic graphs approximating the experimental values of the linear regression parameter a_0: (a) $n = 20$; (b) $n = 50$; (c) $n = 200$.

Analysis of these and other results of experimental research shows that for the regression error, the distribution of which is a significantly non-Gaussian model (large absolute values of skewness and kurtosis), the normalization of the distribution of estimates is observed only for a sufficiently large size of initial sample elements, i.e. $N > 200$. An important is that, despite the presence of nonlinear transformations

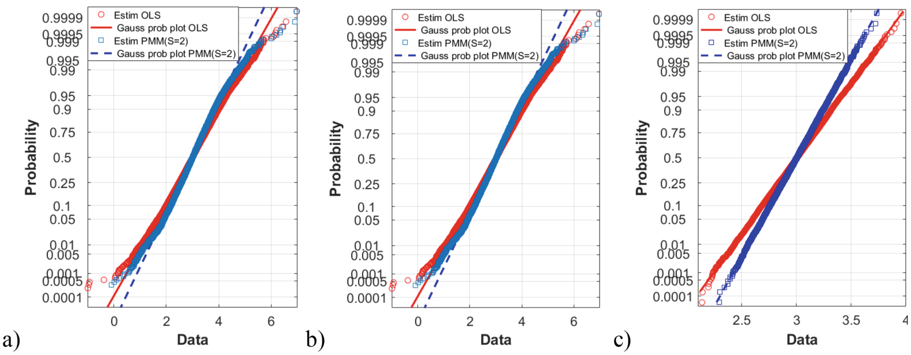

Fig. 3. Gaussian probabilistic graphs approximating the experimental values of the linear regression parameter a_1: (a) n = 20; (b) n = 50; (c) n = 200.

(calculation of quadratic statistics for finding PMM-estimates for S = 2), rate of normalization of the empirical distribution of PMM-estimates is not inferior, and in some cases, exceeds the dynamics of the normalization of OLS-estimates.

9 Experiment with Real Data

As an example for testing the proposed adaptive procedure for finding PMM-estimates of linear regression parameters, we used the Auto MPG data set from the UCI repository [26]. This set of data represents the dependence of fuel consumption (galon per mile) in the urban cycle on the different characteristics of cars [27].

Figure 4a shows the dependence of fuel consumption (variable y) on one of the parameters (acceleration as predictor x) for N = 392 (taking into account the missing data) types of cars. A visual analysis of this figure suggests the existence of a linear relationship between the parameters y and x, which can be described by a regression model of the form (1). OLS estimates of the parameters of this model are the values of: $\hat{a}_0^{(1)} = 4.83$, $\hat{a}_1^{(1)} = 1.2$. Based on the Broyush-Pagan test [28], the value of *p-value* at 0.01, actually confirms the hypothesis of hetero-scedasticity of regression OLS residuals and the correctness of the linear dependence.

The asymmetry of the error distribution of the model is visually visible in Fig. 4b, where a histogram of regression OLS residuals. The hypothesis of the Gaussianity of OLS-residuals is also refuted by the Yarki-Bera test embedded in the MATLAB ($JBSTAT$ = 17.4 at the threshold value CV = 5.8 at a fixed significance level $\alpha_0 = 0.05$) based on an analysis of the value of cumulant skewness and kurtosis [29].

Estimated values $\hat{\kappa}_2 = 50$, $\hat{\gamma}_3 = 0.49$ and $\hat{\gamma}_4 = -0.32$ of regression OLS-residuals (taking into account a sufficiently large volume of initial data N = 392) according to (14), allow us to estimate the magnitude of the decrease in the dispersion of the obtained PMM-estimates of the regression parameters ($\hat{a}_0^{(2)} = 6.5$; $\hat{a}_1^{(2)} = 1.09$) at the level $\hat{g}_{(\theta_p)2} = 0.86$.

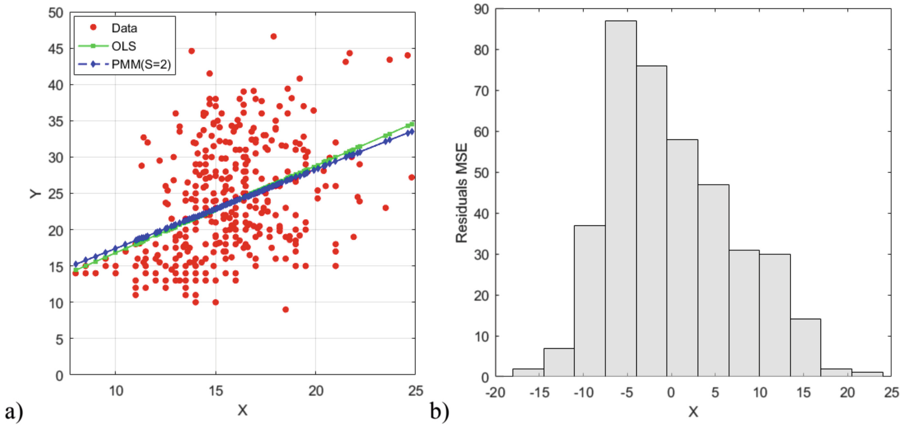

Fig. 4. Construction of the model fuel consumption depending on the vehicle acceleration: (a) experimental data and regression estimates, based on OLS and adaptive PMM; (b) histogram of regression OLS-residuals.

It is known that interval, rather than point, estimates of parameters are more preferable for regression analysis problems. Taking into account the observance of the condition for normalizing the distribution of OLS-estimates (for N – large ones), we get that at a confidence level of 0.95, the interval (0.82; 8.84) covers the value of the regression parameter a_0, and the interval (0.94; 1.45) – the parameter a_1. At the same time, taking into account the value of the estimation of the dispersion reduction coefficient, as well as the obtained point values of the PMM-estimations, allows for a given level of confidence probability, it is justified to correct and reduce by $1 - \sqrt{0.86} \approx 7\%$ the width of the confidence interval, obtaining its limits (2.8; 10.22) for parameter a_0, and (0.85; 1.33) for parameter a_1.

To verify the correctness of results, a statistical experiment based on the bootstrapping method [30] was provided. 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 is shown in Fig. 5. The boundaries of the 95% confidence intervals obtained as a result of statistical bootstrap modeling, and comparison with the calculated values of the interval estimates obtained above indicate their significant closeness. This generally confirms the reliability of the analytical calculations.

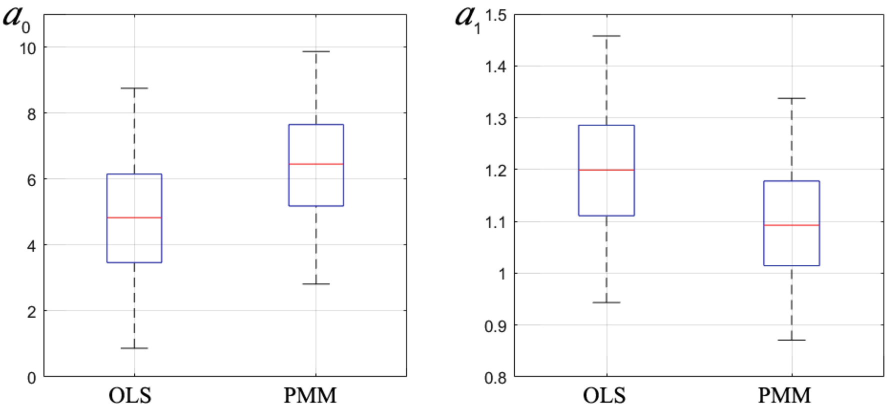

Fig. 5. Empirical distribution of the linear regression parameter estimates

10 Conclusions

Analysis of the results confirmed the possibility and expediency of applying the polynomial maximization method to find estimates of single-factor linear regression parameters under the condition of the asymmetry of error distribution, which are described with the help of higher-order statistics.

Theoretical studies have shown that OLS-estimators are a particular case of PMM-estimates obtained for the degree of a polynomial $S = 1$. Nonlinear PMM-estimates synthesized at the degree of a polynomial $S = 2$, mostly characterized by greater accuracy in comparison with OLS estimates. The coefficient of dispersion decrease is determined by the degree of non-Gaussianity of the random component of the regression model, expressed numerically by absolute values of cumulative coefficients of skewness and kurtosis. In the example of real data, the theoretical value of the variance reduction of the regression parameters estimates was about 14%. It was confirmed by the bootstrap analysis.

The results obtained in statistical modeling by the Monte Carlo method mostly confirm the effectiveness of the use of PMM (for $S = 2$) for situations in which the errors of the regression model have an asymmetric character of the distribution. With the increase in the amount of sample data, the experimental values of the variance reduction coefficients tend to theoretically calculated values, and the distribution of PMM estimates is normalized. An important factor is that the operability of the proposed polynomial regression analysis procedures is also observed in practically important situations of the absence of a priori information about the parameters of non-Gaussian errors. In this case, a posteriori estimates of the necessary error parameters can be used, calculated on basis of a statistical analysis of regression OLS residuals.

Thus, a new approach to finding estimates of the parameters of regression models is proposed, which can be interpreted as adaptive and compromise from the point of view of practical implementation. The proposed procedures potentially have less analytical and computational complexity than the parametric approach based on PMMs and provide an increase in accuracy in comparison with OLS, which does not take into account the difference between the probability distribution and statistical data from the Gaussian law.

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

- The extension of the proposed approach and the study of the peculiarities of its application for nonlinear single-factor regression models, as well as consideration of more general multi-factorial cases.
- Consideration of cases of PMM application for finding estimates of parameters of regression models having non-Gaussian, but symmetrically distributed errors.

References

1. Ryan, T.P.: Modern Regression Methods, vol. 655. Wiley, Hoboken (2008)
2. Huber, P.J., Ronchetti, E.M.: Robust Statistics. Wiley, Hoboken (2009). https://doi.org/10.1002/9780470434697
3. Williams, M.S.: A regression technique accounting for heteroscedastic and asymmetric errors. J. Agric. Biol. Environ. Stat. **2**(1), 108–129 (1997). https://doi.org/10.2307/1400643
4. Prykhodko, S., Makarova, L.: Confidence interval of nonlinear regression of restoration time of network terminal devices. Eastern-Eur. J. Enterp. Technol. **3**(4(69)), 26–31 (2014). https://doi.org/10.15587/1729-4061.2014.24663
5. Box, G.E.P., Cox, D.R.: An analysis of transformations. J. Roy. Stat. Soc. Ser. B **26**, 211–246 (1964)
6. Johnson, N.L.: Systems of frequency curves generated by methods of translation. Biometrika **36**, 149–176 (1949)
7. Zeckhauser, R., Thompson, M.: Linear regression with non-normal error terms. Rev. Econ. Stat. **52**(3), 280–286 (1970)
8. Marazzi, A., Yohai, V.J.: Adaptively truncated maximum likelihood regression with asymmetric errors. J. Stat. Plan. Inference **122**(1–2), 271–291 (2004). https://doi.org/10.1016/j.jspi.2003.06.011
9. Bianco, A.M., Garcia Ben, M., Yohai, V.J.: Robust estimation for linear regression with asymmetric errors. Can. J. Stat. **33**(4), 511–528 (2005)
10. Pal, M.: Consistent moment estimators of regression coefficients in the presence of errors in variables. J. Econometrics **14**, 349–364 (1980)
11. Van Montfort, K., Mooijaart, A., de Leeuw, J.: Regression with errors in variables: estimators based on third order moments. Stat. Neerlandica **41**(4), 223–237 (1987)
12. Dagenais, M.G., Dagenais, D.L.: Higher moment estimators for linear regression models with errors in the variables. J. Econometrics **76**(1–2), 193–221 (1997). https://doi.org/10.1016/0304-4076(95)01789-5
13. 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
14. Gillard, J.: Method of moments estimation in linear regression with errors in both variables. Commun. Stat. Theory Methods **43**(15), 3208–3222 (2014). https://doi.org/10.1080/03610926.2012.698785

15. Kunchenko, Y.P., Lega, Y.G.: Estimation of the parameters of random variables by the polynomial maximization method. Naukova dumka, Kiev (1991). (in Russian)
16. Kunchenko, Y.: Polynomial Parameter Estimations of Close to Gaussian Random Variables. Shaker Verlag, Aachen (2002)
17. Chertov, O., Slipets, T.: Epileptic seizures diagnose using Kunchenko's polynomials template matching. In: Fontes, M., Günther, M., Marheineke, N. (eds.) Progress in Industrial Mathematics at ECMI 2012, pp. 245–248. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-05365-3_33
18. Zabolotnii, S.W., Warsza, 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. Springer (2016). https://doi.org/10.1007/978-3-319-29357-8_80
19. Zabolotnii, S.W., Warsza, Z.L.: Semi-parametric polynomial modification of CUSUM algorithms for change-point detection of non-Gaussian sequences. In: XXI IMEKO World Congress Measurement in Research and Industry (2015)
20. Palahin, V., Juhár, J.: Joint signal parameter estimation in non-Gaussian noise by the method of polynomial maximization. J. Electr. Eng. **67**(3), 217–221 (2016). https://doi.org/10.1515/jee-2016-0031
21. 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. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54042-9_45
22. Cramér, H.: Mathematical Methods of Statistics (PMS-9). Princeton University Press, Princeton (2016)
23. Cook, R.D., Weisberg, S.: Residuals and influence in regression. Monographs on Statistics and Applied Probability (1982). https://doi.org/10.2307/1269506
24. Mathews, J.H., Fink, K.D.: Numerical Methods Using MATLAB, vol. 4. Pearson, London (2004)
25. Lilliefors, H.W.: On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc. **62**(318), 399–402 (1967)
26. Lichman, M.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2013). http://archive.ics.uci.edu/ml
27. Quinlan, J.R.: Combining instance-based and model-based learning. In: Proceedings of the Tenth International Conference on Machine Learning, pp. 236–243 (1993)
28. Breusch, T.S., Pagan, A.R.: A simple test for heteroscedasticity and random coefficient variation. Econometrica J. Econometric Soc. **47**, 1287–1294 (1979)
29. Jarque, C.M., Bera, A.K.: A tests of observations and regression residuals. Int. Stat. Rev. **55**, 163–172 (1987)
30. Efron, B.: Better bootstrap confidence intervals. J. Am. Stat. Assoc. **82**(397), 171–185 (1987). https://doi.org/10.1080/01621459.1987.10478410