Application of the Polynomial Maximization Method for Estimation Parameters in the Polynomial Regression with Non-Gaussian Residuals

Serhii Zabolotnii[1,2], Oleksandr Tkachenko[1], Zygmunt L. Warsza[3],

[1] Cherkasy State Business College, Ukraine,
[2] Kyiv National University of Technologies and Design, Ukraine,
email: zabolotniua@gmail.com
[3] Research Net Łukasiewicz-Industrial Research Institute of Automation and Measurements PIAP-Ł, Al. Jerozolimskie 202, 02-486 Warszawa, Poland,
email: zlw1936@gmail.com

Abstract. This paper considers the application of the polynomial maximization method to find estimates of the parameters of polynomial regression. It is shown that this method can be effective for the case when the distribution of the random component of the regression models differs significantly from the Gaussian distribution. This approach is adaptive and is based on the analysis of higher-order statistics of regression residuals. Analytical expressions that allow finding estimates and analyzing their uncertainty are obtained. Cases of asymmetry and symmetry of the distribution of regression errors are considered. It is shown that the variance of estimates of the polynomial maximization method can be significantly less than the variance of the estimates of the least squares method, which is a special case. The increase in accuracy depends on the values of the cumulant coefficients of higher orders of random errors of the regression model. The results of statistical modeling by the Monte Carlo method confirm the effectiveness of the proposed approach.

1. Introduction

Estimation of parameters of the polynomial regression is one of the earliest problems in statistical analysis. For more than two centuries, the polynomial regression model has become widespread in solving various applied problems of technical, geophysical, biomedical, economic and other aspects of human activity. In most cases, variations of the least square (LS) method are used to find the estimates of parameters.

The widespread adoption of the least square's method is due to two main factors. Firstly, the algorithmic simplicity of its implementation. Secondly, by the well-known fact that under a number of conditions, the estimates of the Ordinary Least Square (OLS) method are optimal in the sense of ensuring a minimum of their variance. One of these conditions is the normalization of the distribution law of the random component of the regression model. It is because the linear OLS estimator is equivalent to the Maximum Likelihood Estimator (MLE). In addition, the Gaussian probabilistic models a convenient idealization, as its application greatly simplifies the synthesis and analysis of statistical methods. However, such a model is not always adequate to real statistical data. In case of mismatch may be employed various approaches, in particular, the normalizing transform [1, 2] or modifiers robust least squares or maximum likelihood estimators [3].

When fulfilling the assumption that model errors are independent of predictors, an approach that is sometimes called adaptive can be applied [4]. Adaptability is understood in the sense that a series of successive refinement steps are applied. At the first stage, the parameters of the deterministic component of the model are estimated by a simple method that does not consider the specifics of the probabilistic distribution of statistical data. After removing the deterministic component, the type of the random component of the model is identified and estimates of its parameters are found. At the third stage, refined (adaptive) estimates of the parameters of the deterministic component of the model are found and are considered the probabilistic properties of the residual variable (errors). Various authors have shown that the application of this approach can significantly reduce the variance of parameter estimates [5-8].

The adaptive approach can be based not only on likelihood maximization, which requires determining the probability distribution law and finding estimates of its parameters. This is a separate and often computationally complex task. An alternative can be based on the use of a probabilistic description of regression models in the form of higher-order statistics, for example, moments or cumulants [9-13]. The use of this description in combination with the adaptive approach used in this work.

2. Purpose of the study

Polynomial Maximization Method (PMM) is a relatively new method of statistical estimation, which is an integral part of the stochastic polynomial apparatus developed by Kunchenko [14]. The use of this mathematical apparatus provides new opportunities for constructing semi-parametric methods for processing non-Gaussian data. In addition to estimating the parameters, these can be tasks of testing statistical hypotheses [15], pattern recognition ("template matching") [16], posteriori [17] and on-line [18] disorders detections (change point problems) etc.

In a sense, PMM is ideologically similar to MLE. It also uses the principle of maximizing statistics from sample data in the neighborhood of the true value of the parameter being estimated. However, to form such a functional, the probability density distribution is not used, but a partial description based on higher-order statistics.

This study is a continuation of works [19, 20], which considered the use of PMM for solving the problem of adaptive estimation of parameters of linear regression with asymmetric and symmetric non-Gaussian residuals. Its main purpose is to generalize the results of these works for a more complex polynomial regression model, as well as to carry out a comparative analysis of the estimation efficiency by means of the statistical modeling by the Monte Carlo method.

It should also be noted that the notion of "stochastic polynomials" and the presence of the term "maximization of polynomials" in the name of the estimation method used in this work are not related to the polynomial type of the deterministic component of the regression model. These are completely different types of polynomials in terms of their functionality and practical application.

3. Kinds of correlation of the measured ordinates of points

Let there be a regression model of observations that describes the dependence of values of the target variable $\boldsymbol{Y} = \{y_1, y_2, \dots, y_N\}$ from the predictor $\boldsymbol{X}\{x_1, x_2, \dots, x_N\}$:

$$y_v = \sum_{p=0}^{Q-1} a_p x_v^p + \xi_v,\ v = \overrightarrow{1,N}, \tag{1}$$

where $\boldsymbol{\theta} = \{a_0, a_1, \dots a_{Q-1}\}$ - a vector containing the parameters (coefficients of the power polynomial) of the deterministic component of the model; ξ_v - readings of the random component (error) of the model ($E\{\xi\} = 0$), which is a sequence of independent and equally distributed random variables.

The probabilistic properties of the error of the regression model differ significantly from the Gaussian law and can be described using a sequence of central moments μ_r or cumulants κ_r (cumulative coefficients γ_r). With this description, the second-order cumulant κ_2 coincides with variance μ_2 random component, and the cumulant coefficients of higher orders $\gamma_r = \kappa_r / \kappa_2^{r/2}$ numerically describe the degree of difference from a Gaussian distribution.

The general task is to find estimates of the components of the vector parameter $\boldsymbol{\theta} = \{a_0, a_1, \dots a_{Q-1}\}$ based on statistical analysis of multiple points (x_v, y_v), $v = \overrightarrow{1,N}$. In this case, the values of the parameters of the probabilistic component of the model are a priori unknown.

4. Results

To solve the posed problem using the we use the power transformations of sample values as basic functions, i.e. $\phi_i(y_v) = y_v^i$. In this case, the sequence of mathematical expectations is a set of initial moments of the corresponding order

$$\Psi_{iv} = E\{y_v^i\} = \alpha_{iv}, i = \overrightarrow{1,S},\ v = \overrightarrow{1,N},$$

and functions $F_{(i,j)v} = \alpha_{(i+j)v} - \alpha_{iv}\alpha_{jv}$ called the centered correlants [14].

4.1. PPM-estimates of regression parameters with polynomial degree *S*=1

When using a stochastic polynomial of order $S = 1$ PMM Estimates of vector parameter elements $\boldsymbol{\theta}$ deterministic component of the regression model (1) can be found from the solution of the system Q equations of the form:

$$\sum_{v=1}^{N} \left\{ k_{1,v}^{(p)} \left[y_v - \sum_{p=0}^{Q-1} a_p x_v^p \right] \right\} = 0,. \tag{2}$$

where $k_{1,v}^{(p)} = \frac{1}{\mu_2} \frac{\partial}{\partial a_p} \left[\sum_{p=0}^{Q-1} a_p x_v^p \right] = \frac{x_v^p}{\mu_2}$, $p = \overrightarrow{0,Q-1}$.

After certain transformations (2) can be represented in matrix form as a system of linear equations

$$(\boldsymbol{B}^T\boldsymbol{B})\boldsymbol{\theta}^T = \frac{1}{N}\sum_{v=1}^{N}\boldsymbol{B}^T\boldsymbol{Y}^T, \tag{3}$$

where $\boldsymbol{B}$ - Vandermonde matrix with elements $b_{v,p} = x_v^p$, $p = \overrightarrow{0,Q-1}$, $v = \overrightarrow{1,N}$.

Note that the resulting system (11) for finding PMM estimates $\hat{\boldsymbol{\theta}}_{(1)}$ parameters of polynomial regression (1) for S=1 is equivalent to the linear system of OLS-estimates. This means that the accuracy of both methods is the same. Using (2) and (3), we can write down the variational matrix of estimates, which for asymptotic case (for $v \to \infty$) allows you to calculate their variance

$$\boldsymbol{V}_{(1)} = \mu_2\left[\boldsymbol{B}\boldsymbol{B}^T\right]^{-1}. \tag{4}$$

As you know, such estimates are optimal (according to the criterion of minimum variance) in the situation when the errors of the regression model have a Gaussian distribution. In the case when the probabilistic nature of the errors differs from the Gaussian law, there are more accurate alternative estimation methods. Below we consider a new approach to nonlinear estimation of regression parameters (1), based on the use of higher-order power-law stochastic polynomials.

4.2. PPM-estimates of regression parameters with polynomial degree *S*=2

When using stochastic polynomials of order $S = 2$ PMM scores $\hat{\boldsymbol{\theta}}_{(2)}$ deterministic component of the regression model (9) can be found from the solution of the system Q equations of the form:

$$\sum_{v=1}^{N}\left\{k_{1,v}^{(p)}\left[y_v - \sum_{p=0}^{Q-1} a_p x_v^p\right] + k_{2,v}^{(p)}\left[(y_v)^2 - \left[\left(\sum_{p=0}^{Q-1} a_p x_v^p\right)^2 + \mu_2\right]\right]\right\} = 0,$$
$$p = \overrightarrow{0,Q-1} \tag{5}$$

Where: the optimal coefficients $k_{i,v}^{(p)}$, $i = \overrightarrow{1,2}$ ensure the minimization of the variance of the estimates of the components of the desired parameter when using the degree of the polynomial $S = 2$.

These coefficients can be represented as functions depending on the central moments of the random component of the regression model:

$$k_{1,v}^{(p)} = \frac{\mu_4 - \mu_2^2 + 2\mu_3\left(\sum_{p=0}^{Q-1} a_p x_v^p\right)}{\mu_2(\mu_4 - \mu_2^2) - \mu_3^2} x_v^p, k_{2,v}^{(p)} = -\frac{\mu_3^2}{\mu_2(\mu_4 - \mu_2^2) - \mu_3^2} x_v^p. \tag{6}$$

Substituting the coefficients (6) in (5), after certain transformations, the system of equations for finding estimates can be written in the form:

$$\sum_{v=1}^{N}\left\{x_v^p\left[A_2\left(\sum_{p=0}^{Q-1} a_p x_v^p\right)^2 + B_{2,v}\left(\sum_{p=0}^{Q-1} a_p x_v^p\right) + C_{2,v}\right]\right\} = 0, \quad p = \overrightarrow{0,Q-1} \tag{7}$$

where $A_2 = \mu_3, B_{2,v} = \mu_4 - \mu_2^2 - y_v\mu_3, C_{2,v} = y_v^2\mu_3 - y_v\left(\mu_4 - \mu_2^2\right) - \mu_2\mu_3$. (8)

Obviously, for the degree of the polynomial $S = 2$ PMM estimates can only be found numerically, for example, using the Newton-Raphson iterative procedure.

Using expressions of the optimal coefficients (6), it can be shown that the elements of the variation matrix of PMM estimates for $S = 2$ differ from the elements of the matrix of linear estimates (4) by a certain coefficient $\boldsymbol{V}_{(2)} = g_2 \boldsymbol{V}_{(1)}$, which can be represented as

$$g_2 = 1 - \frac{\gamma_3^2}{2+\gamma_4}. \qquad (9).$$

The transition in expression (9) from the instantaneous to the cumulant description is due to the fact that the deviations of the values of the cumulant coefficients of higher orders γ_r from zero indicates the degree of deviation from a Gaussian distribution. In addition, higher order cumulant coefficients cannot take arbitrary values. In particular, for the cumulant coefficients of skewness and kurtosis, the inequality $\gamma_4 + 2 \geq \gamma_3^2$. Taking into account this inequality, based on the analysis of (9), we can conclude that the coefficient of variance reduction g_2 is a dimensionless quantity that belongs to the range $(0;1]$. Thus, the relative decrease in the asymptotic variance of PMM estimates is the same for all components of the vector parameter. With an increase in the asymmetry of the distribution, the value of the relative decrease in variance can be quite significant and even asymptotically tend to zero as the absolute value of the asymmetry coefficient approaches the limits of the range of permissible values $|\gamma_3| \to \sqrt{2+\gamma_4}$.

4.3 PPM-estimates of regression parameters with polynomial degree *S*=3 (symmetric case)

Analysis (9) shows that in the case of symmetry of the distribution (or at least for $\gamma_3 = 0$) the random component of the errors of the regression model of the PMM application at $S = 2$ is impractical. But in the case of symmetry (provided that the distribution is significantly different from the Gaussian law), the distribution can be applied by a stochastic log of a higher degree $S = 3$. In this case, PMM estimates $\hat{\boldsymbol{\theta}}_{(3)}$ are found from the solution of the system Q equations of the form

$$f_3^{(p)}(y_v, x_v) = \sum_{v=1}^{N} \Big\{ k_{1,v}^{(p)} \big[y_v - \sum_{p=0}^{Q-1} a_p x_v^p\big] + k_{2,v}^{(p)} \Big[y_v^2 - \Big(\big[\sum_{p=0}^{Q-1} a_p x_v^p\big]^2 + \mu_2 \Big) \Big] +$$
$$+ k_{3,v}^{(p)} \Big[y_v^3 - \Big(\big[\sum_{p=0}^{Q-1} a_p x_v^p\big]^3 + 3\mu_2 \sum_{p=0}^{Q-1} a_p x_v^p \Big) \Big] \Big\} = 0, \quad p = \overrightarrow{0, Q-1} \quad (10)$$

Where: the optimal coefficients $k_{i,v}^{(p)}$, $i = \overrightarrow{1,3}$ ensuring the minimization of the variance of the estimates of components of the required parameter, and can be represented in the form

$$k_{1,v}^{(p)} = \frac{3\big(\sum_{p=0}^{Q-1} a_p x_v^p\big)^2 (\mu_4 - 3\mu_2^2) + 3\mu_4\mu_2 - \mu_6}{\mu_2^2(\mu_4^2 - \mu_2\mu_6)} x_v^p,$$
$$k_{2,v}^{(p)} = \frac{-3(\mu_4 - 3\mu_2)}{\mu_2^{-2}(\mu_4^2 - \mu_2\mu_6)} \big(\sum_{p=0}^{Q-1} a_p x_v^p\big) x_v^p$$
$$k_{3,v}^{(p)} = \frac{\mu_4 - 3\mu_2}{\mu_2^{-2}(\mu_4^2 - \mu_2\mu_6)} x_v^p, \quad p = \overrightarrow{0, Q-1}, \qquad (11a,b,c)$$

Substituting the coefficients (11) into (10), after certain transformations, the system of equations for finding estimates can be written in the form:

$$\sum_{v=1}^{N}\left\{x_v^p\left[A_3\left(\sum_{p=0}^{Q-1}a_p x_v^p\right)^3+B_{3,v}\left(\sum_{p=0}^{Q-1}a_p x_v^p\right)^2+C_{3,v}\left(\sum_{p=0}^{Q-1}a_p x_v^p\right)+D_{3,v}\right]\right\}=0,\quad p=\overrightarrow{0,Q-1},\quad (12)$$

where $A_3 = 1,\ B_{3,v} = -3y_v,\ C_{3,v} = 3y_v^2 - \frac{\mu_6-3\mu_4\mu_2}{\mu_4-3\mu_2^2},\ D_{3,v} = y_v\frac{\mu_6-3\mu_4\mu_2}{\mu_4-3\mu_2^2} - y_v^3.$

Obviously, as in the previous case, for the degree of the polynomial $S = 3$ PMM estimates can only be found by numerically solving systems of equations (21).

Using expressions for the optimal coefficients (11), it can be shown that the elements of the variation matrix of PMM estimates for $S = 3$ differ from the elements of the matrix of linear estimates (4) by the coefficient

$$g_3 = 1 - \frac{\gamma_4^2}{6+9\gamma_4+\gamma_6}, \quad (13)$$

which is a function of 4th and 6th order cumulant coefficients.

As already noted above, for the value of the cumulant coefficients of a higher order, there are certain ranges of acceptable values. In particular, for symmetrically distributed random variables, the constraints for cumulant coefficients of the 4th and 6th order are determined by the inequalities:$\gamma_4 > -2$ and $\gamma_6 + 9\gamma_4 + 6 > \gamma_4^2$. Taking into account these inequalities, we can conclude that the variance reduction coefficient g_3 is a dimensionless quantity and belongs to the range $(0;1]$. Analysis (13) shows that the accuracy of PMM estimates for $S = 3$ can be significantly higher than the accuracy of linear estimates. In particular, when the value of the cumulant coefficients approaches the boundary of the region of permissible values$|\gamma_4| \to \sqrt{\gamma_6 + 9\gamma_4 + 6}$ the variance of PMM estimates tends asymptotically to zero.

It is necessary to note the universality of expressions (9) and (13), which describe the corresponding coefficients for reducing the variance of PMM estimates both for the estimated scalar parameter [14] and for the estimation of linear regression parameters [19, 20].

4.4 Algorithm for adaptive PMM estimates

The results of the theoretical analysis of the effectiveness of PMM estimates indicate the advisability of applying this approach only if the distribution of the random component of the regression model is significantly different from the Gaussian law. However, it is known that the inadequacy of the Gaussian model is not a critical factor from the point of view that OLS estimates still remain unbiased and consistent, although they cease to be optimal. Since the least square's method is inherently linear, the probable properties of the regression residuals practically do not differ from the properties of the original random component of the regression model [21].

Let us recall the fact that is important from a practical point of view that to obtain PMM estimates, information is used not on the distribution of regression errors, but their moment-cumulative description. At the same time, the solution of the problem of overcoming the a priori uncertainty regarding the probabilistic characteristics of the model is significantly simplified in comparison with the approach based on maximum likelihood. In addition, the obtained posteriori estimates of the cumulant coefficients of skewness and kurtosis can be used to test the hypothesis about the Gaussian law of distribution and the symmetry of regression errors [22].

Considering the above, we generalize the algorithm (elements of which are proposed in [19,20]) for finding adaptive PMM estimates of the regression parameters:

Step 1 - finding linear OLS -estimates of regression parameters;

Step 2 - formation of regression residuals and finding posteriori estimates of moments / cumulants up to the 4th order;

Step 3 - testing the hypothesis about the Gaussian distribution of regression residuals (if confirmed, the algorithm ends);

Step 4 - testing the hypothesis about the symmetry of the distribution of regression residuals (if it is confirmed, go to step 6);

Step 5 - finding PMM estimates using a polynomial of degree $S = 2$ and completion of the algorithm;

Step 6 - finding estimates of the moments of the 6th order of regression residuals;

Step 7 - finding PMM estimates using a polynomial of degree $S = 3$ and the completion of the algorithm.

5. Statistical modeling

To verify the theoretical results, the set of MATLAB / OCTAVIO compatible statistical modeling functions was modified, which was used in [17-20]. This toolkit, based on repeated tests by the Monte Carlo method, allows for a comparative analysis of the accuracy of various methods of statistical estimation, as well as to investigate the properties of PMM estimates of parameters when the statistical data model is non-Gaussian.

Since the value of the coefficients of the ratio of the variance of estimates is used as a comparative criterion of efficiency, and the theoretical values of the variance themselves are calculated for the asymptotic case, the experimental values $\hat{g}_S$ will obviously differ from theoretical ones in a certain way. The degree of discrepancy depends on the size of the samples used to find parameter estimates and the number of experiments performed. M according to the Monte Carlo procedure.

To carry out statistical modeling of the estimation of the parameters of regression dependences of the form (1), a quadratic polynomial with a vector of parameters was used as a deterministic component of the regression model $\boldsymbol{\theta} = \{1,2,-4\}$that were assumed to be unknown and subject to assessment.

When using PMM of degree S=2 as an asymmetric random component (error) of the regression model, a sequence of independent and equally distributed random variables with exponential, lognormal, gamma, and Weibull distribution is used (see Table 1). When using PMM of degree S=3, random variables with uniform, trapezoidal, triangular and Laplace distribution are used as a symmetrically distributed random error (see Table 2). In this case, the values of the parameters of the random component of the regression model, which are necessary to find the adaptive PMM estimates, were considered a priori unknown, and in accordance with the adaptive algorithm, their posterior estimates were used.

Table 1. Results of statistical modeling of Monte Carlo with the degree of the polynomial S=2.

Distribution		Theoretical values			Results of statistical modeling Monte Carlo								
					$\hat{g}_2$								
		γ_3	γ_4	g_2	$n=50$			$n=50$			$n=200$		
					a_0	a_1	a_2	a_0	a_1	a_2	a_0	a_1	a_2
Gamma	$\alpha = 0.5$	2.83	12	0.42	0.44	0.38	0.37	0.43	0.35	0.35	0.43	0.36	0.36
Exponential (Gamma, $\alpha = 1$)		2	6	0.5	0.57	0.53	0.54	0.54	0.46	0.46	0.53	0.47	0.47
Gamma	$\alpha = 2$	1.41	3	0.6	0.64	0.64	0.63	0.63	0.59	0.6	0.63	0.59	0.59
	$\alpha = 4$	1	1.5	0.71	0.75	0.74	0.75	0.73	0.72	0.71	0.74	0.72	0.71
Lognormal $\sigma^2 = 0.1, \mu = 1$		1	1.86	0.74	0.77	0.76	0.76	0.75	0.74	0.74	0.74	0.75	0.75
Weibull $\alpha = 1, b = 2$		0.63	0.25	0.82	0.91	0.85	0.85	0.84	0.84	0.84	0.83	0.82	0.82

Table 2. Results of statistical modeling of Monte Carlo with the degree of the polynomial S=3.

Distribution		Theoretical values			Results of statistical modeling Monte Carlo								
					$\hat{g}_3$								
		γ_4	γ_6	g_3	$n=50$			$n=50$			$n=200$		
					a_0	a_1	a_2	a_0	a_1	a_2	a_0	a_1	a_2
Uniform		-1.2	6.9	0.3	0.55	0.53	0.54	0.41	0.4	0.4	0.35	0.35	0.35
Trapezoid	$\beta = 0.75$	-1.1	6.4	0.36	0.61	0.58	0.58	0.47	0.47	0.46	0.41	0.41	0.41
	$\beta = 0.5$	-1	5	0.55	0.75	0.73	0.74	0.66	0.66	0.66	0.59	0.58	0.58
	$\beta = 0.25$	-0.7	2.9	0.76	0.95	0.94	0.94	0.86	0.86	0.87	0.81	0.82	0.82
Triangular		-0.6	1.7	0.84	one	one	one	0.95	0.94	0.94	0.9	0.9	0.89
Laplace		3	30	0.86	1.4	1.66	1.72	0.86	0.86	0.85	0.83	0.82	0.82

The set of experimental values of the coefficients, the variance ratio was obtained at $M = 10^4$. To find the refined values of adaptive PMM estimates by solving systems of polynomial equations, we used the Newton-Raphson numerical procedure.

The results presented in Table 1 and Table 2 show that the experimental values of the ratio of PMM variances to the variance of OLS estimates differ to a certain extent from the theoretical ones. But as noted earlier, this is due to the fact that the theoretical values g_2 and g_3 calculated based on formulas (9) and (13) obtained for the asymptotic case. In addition, the accuracy of obtaining adaptive PMM estimates is affected by the uncertainty of the posterior estimates of parameters (moments of the corresponding order) of regression residuals, which also decreases with increasing sample size. However, in general, the analysis of the presented data indicates the adequacy of analytical calculations, since the maximum relative discrepancy between the theoretical and experimental values of the coefficients is the variance ratio at N=200 does not exceed 3%.

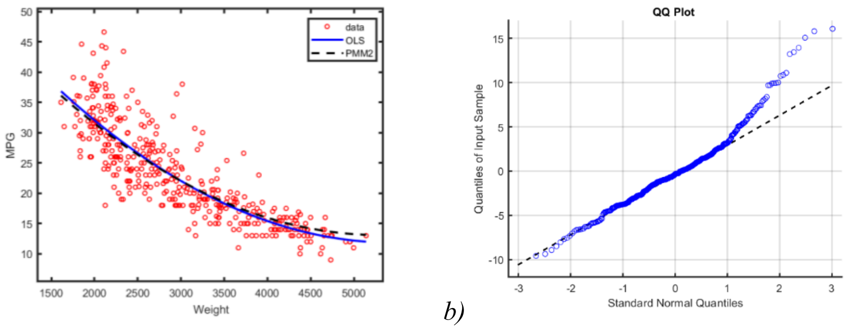
Another important result of statistical modeling is the confirmation of the theoretical property, which is that the relative increase in the accuracy of PMM estimates is the same for all components of the vector parameter (coefficients of the deterministic component of the polynomial regression). Moreover, it is practically independent of the order Q the deterministic component of the regression model (1). This property is also confirmed by the significant correlation between the results of this study and the results of statistical modeling of finding estimates of the parameters of linear one-factor regression, given in [19, 20].

6. Experiments with real data

As an example, for testing the proposed adaptive procedure for finding PMM-estimates parameters of polynomials regression, we used the Auto MPG data set from the UCI repository. This set of data represents the dependence of fuel consumption in the urban cycle on the different characteristics of cars [23].

6.1 Experiment 1 (asymmetrical case – PMM *S=2*)

In the first experiment, we solve the problem of estimating the parameters of the dependence of fuel consumption (parameter *MPG* - Mile Per Gal) on the weight of the car (parameter *Weight*). Fig.1, (a) presents experimental data for $N = 392$ cars (considering the missing data) of different. A visual analysis of this figure suggests the existence of a quadratic relationship $MPG = a_0 + a_1 Weight + a_2 Weight^2$.



a) *b)*

Fig. 1. Construction of the model fuel consumption depending on the weight of cars: *a)* Experimental data and regression estimates, based on OLS and adaptive PMM2*;* *b)* QQ plot of OLS- regression residuals

OLS estimates (with 95% confidence bounds) of this model parameters have values

$$\hat{a}_0{}^{(1)} = 62.3\ (56.4, 68.1);$$

$$\hat{a}_1{}^{(1)} = -18.5 * 10^{-3}(-22.4 * 10^{-3},\ -14.6 * 10^{-3});$$

$$\hat{a}_2{}^{(1)} = 1.7 * 10^{-6}(1.1 * 10^{-6}, 2.3 * 10^{-6}).$$

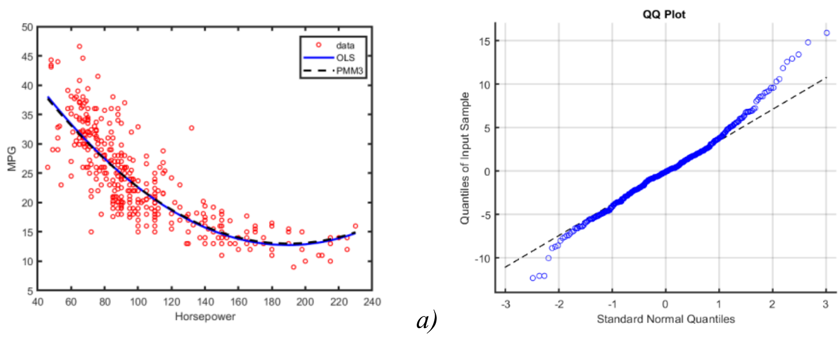
The asymmetry of the error distribution of the model is visually visible in Fig.1 (b), where presents a Q-Q plot of regression OLS-residuals. The hypothesis of the Gaussian of OLS-residuals is also refuted by the Yarki-Bera test ($JBSTAT = 93.2$ at the threshold value $CV = 5.8$ at a fixed significance level $\alpha lfa = 0.05$).

Estimated values $\hat{\gamma}_3 = 0.8$ and $\hat{\gamma}_4 = 1.8$ regression OLS-residuals (taking into account a sufficiently large volume of initial data N) according to (18), allow us to estimate the magnitude of the decrease in the variance estimates (PMM for $S = 2$ compared to OLS) at the level $\hat{g}_2 = 0.83$. As result PMM2 estimates (with 95% confidence bounds) of the parameters of this model are the values of:

$$\hat{a}_0{}^{(2)} = 60.7 \ (55.3, 66);$$
$$\hat{a}_1{}^{(2)} = -18 * 10^{-3}(-21.5 * 10^{-3}, \ -14.4 * 10^{-3});$$
$$\hat{a}_2{}^{(2)} = 1.7 * 10^{-6}(1.1 * 10^{-6}, 2.2 * 10^{-6}).$$

6.2 Experiment 2 (symmetrical case – PMM *S=3*)

In the second experiment, we solve the problem of estimating the parameters of the dependence of fuel consumption (parameter *MPG*) on the power of the car (parameter *Horsepower*). A visual analysis of Fig.2, (a) suggests the existence of a quadratic relationship $MPG = a_0 + a_1 Horsepower + a_2 Horsepower^2$.



a) *b)*

Fig. 2. Construction of the model fuel consumption depending on the horsepower of cars: *a)* Experimental data and regression estimates, based on OLS and adaptive PMM3*;* *b)* QQ plot of OLS- regression residuals

OLS estimates (with 95% confidence bounds) of the parameters of this model are the values of:

$$\hat{a}_0{}^{(1)} = 56.9 \ (53.36, 60.44);$$
$$\hat{a}_1{}^{(1)} = -46.6 * 10^{-2}(-52.7 * 10^{-2}, \ -40.5 * 10^{-2});$$
$$\hat{a}_2{}^{(1)} = 12.3 * 10^{-4}(9.9 * 10^{-4}, 14.7 * 10^{-4}).$$

The hypothesis of the Gaussian of OLS-residuals is refuted by the Yarki-Bera test ($JBSTAT = 30.6$ at the threshold value $CV = 5.8$ at a fixed significance level $\alpha lfa = 0.05$). However, visual analysis of the Q-Q plot of regression OLS-residuals presented in Fig.2 (b) indicates the symmetry of the distribution. This is numerically confirmed by the closeness to zero of the value of the skewness coefficient $\hat{\gamma}_3 \approx 0.2$. In this situation, for adaptive refinement of estimates, it is advisable to use PMM at degree $S = 3$.

Estimated values $\hat{\gamma}_4 = 1.3$ and $\hat{\gamma}_6 = -1.9$ regression OLS-residuals (taking into account a sufficiently large volume of initial data $N = 392$) according to (18), allow us to estimate the magnitude of the decrease in the variance estimates (PMM for $S = 3$ compared to OLS) at the level $\hat{g}_3 = 0.89$. As result PMM3-estimates (with 95% confidence bounds) of the parameters of this model are the values of:

$$\hat{a}_0^{(2)} = 58.2\ (54.9, 61.5);$$
$$\hat{a}_1^{(2)} = -48.9 * 10^{-2}(-54.7 * 10^{-2},\ -43.1 * 10^{-2});$$
$$\hat{a}_2^{(1)} = 13.2 * 10^{-4}(10.9 * 10^{-4}, 15.3 * 10^{-4}).$$

6.3. Bootstrap simulation

To verify the correctness of parameter estimation regression models, a statistical experiment based on the bootstrapping method. With the help of built-in MATLAB bootstrap resampling tools of the original sample was multiplied by 10^4 bootstrap samples with the return. For each of them, we found OLS and PMM estimates for quadratic regression parameters. The empirical distribution of these estimates is presented in Fig. 3 and Fig. 4 in the form of boxplot-graphs.

The presented results confirm the difference between the corresponding OLS and PMM parameter estimates, which consists in the shift of the centers of their confidence intervals and the change in the width.

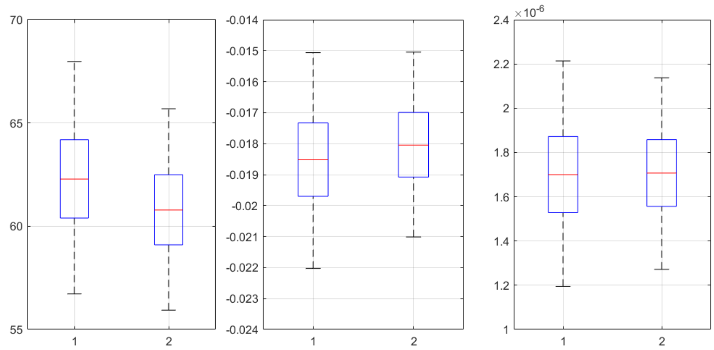


Fig. 3. Empirical distribution of bootstrap OLS (1) and PMM2 (2) estimates of regression $MPG = a_0 + a_1 Weight + a_2 Weight^2$

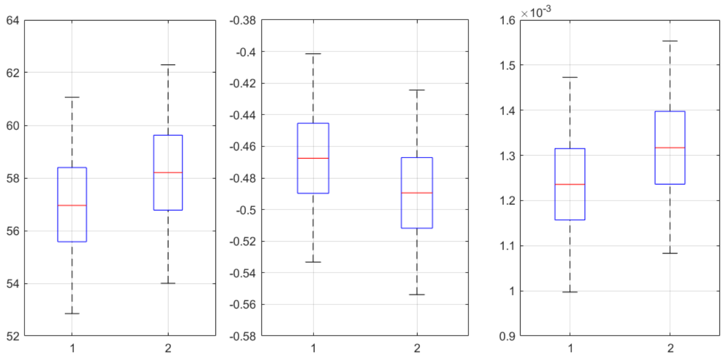


Fig. 4. Empirical distribution of bootstrap OLS (1) and PMM3 (2) estimates of regression

$$MPG = a_0 + a_1 Horsepower + a_2 Horsepower^2$$

7. Summary and conclusions

The totality of the results obtained confirms the possibility of effective application of the Polynomial Maximization Method for solving problems of regression analysis in situations where the regression residuals has a distribution that differs significantly from Gaussian idealization.

The presented results generalize the approaches to the estimation of linear in the parameter regression models, a particular case of which is polynomial regression. Note that the analytical expressions obtained in this work for finding estimates of the parameters of polynomial one-way regression can also be applied for the case of a multivariate linear or polynomial regression model.

The next stage of the study should be the distribution of the proposed adaptive approach based on the use of stochastic polynomials and a partial description by higher-order statistics for the case of nonlinear regression models. In addition, it is interesting to compare the efficiency of the adaptive estimates of the Polynomial Maximization Method and the Maximum Likelihood Estimation.

References

1. Johnson, N. L. (1949), Systems of Frequency Curves Generated by Methods of Translation, Biometrika, 36, 149-176.
2. Box, G. E. P., and Cox, D. R. (1964), An Analysis of Transformations, Journal of the Royal Statistical Society, Ser. B, 26, 211-246.
3. Huber, P. J., & RonchettiE. M. (2009). Robust Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470434697
4. Stone, C. J. (1975). Adaptive Maximum Likelihood Estimators of a Location Parameter. The Annals of Statistics, 3(2), 267–284. https://doi.org/10.1214/aos/1176343056
5. Schechtman, E., & Schechtman, G. (1986). Estimating the parameters in regression with uniformly distributed errors. Journal of Statistical Computation and Simulation, 26(3–4), 269–281. https://doi.org/10.1080/00949658608810965

6. Zeckhauser R, Thompson M (1970). "Linear Regression with Non-Normal Error Terms." The Review of Economics and Statistics, 52, 280–286.
7. Islam, M. Q., Tiku, M. L., & Yildirim, F. (2001). Nonnormal Regression. I. Skew Distributions. Communications in Statistics - Theory and Methods, 30(6), 993–1020. https://doi.org/10.1081/STA-100104347
8. Tiku, M. L., Islam, M. Q., & Selçuk, A. S. (2001). Nonnormal regression. II. Symmetric distributions. Communications in Statistics-Theory and Methods, 30(6), 1021–1045. https://doi.org/10.1081/STA-100104348
9. Van Montfort, K.; Mooijaart, A. and de Leeuw, J. (1987). Regression with errors in variables: estimators based on third order moments, *Statist. Neerlandica*, 41(4), 223–237.
10. Dagenais, M. G., & Dagenais, D. L. (1997). Higher Moment Estimators for Linear Regression Models with Errors in the Variables. *Journal of Econometrics*, *76*(1–2), 193–221. https://doi.org/10.1016/0304-4076(95)01789-5
11. Cragg, J. G. (1997). Using Higher Moments to Estimate the Simple Errors-in-Variables Model. *The RAND Journal of Economics*, *28*, S71. https://doi.org/10.2307/3087456
12. Gillard, J. (2014). Method of Moments Estimation in Linear Regression with Errors in both Variables. *Communications in Statistics - Theory and Methods*, *43*(15), 3208-3222. https://doi.org/10.1080/03610926.2012.698785
13. Giacalone, M. (2020). A combined method based on kurtosis indexes for estimating p in non-linear L p-norm regression. Sustainable Futures, 2, 100008. doi /10.1016/j.sftr.2020.100008
14. Kunchenko Y. (2002) Polynomial Parameter Estimations of Close to Gaussian Random variables. Germany, Aachen: Shaker Verlag.
15. Zabolotnii, S.W., Martynenko, S.S. & Salypa, S.V. Method of Verification of Hypothesis about Mean Value on a Basis of Expansion in a Space with Generating Element. Radioelectron. Comm. Syst. 61, 222–229 (2018). doi /10.3103/S0735272718050060
16. Chertov, O. and Slipets, T. (2014) 'Epileptic Seizures Diagnose Using Kunchenko's Polynomials Template Matching', in Fontes, M., Günther, M., and Marheineke, N. (eds) Progress in Industrial Mathematics at ECMI 2012. Cham: Springer International Publishing, pp. 245–248. doi: 10.1007/978-3-319-05365-3_33.
17. Zabolotnii, S.V., Warsza, Z. L. (2015). Semi-parametric polynomial method for retrospective estimation of the change-point of parameters of non-Gaussian sequences. Advanced Mathematical and Computational Tools in Metrology and Testing X, 400–408. doi:10.1142/9789814678629_0048
18. Zabolotnii, S.W., Warsza, Z.L. (2015). Semi-parametric polynomial modification of CUSUM algorithms for change-point detection of non-Gaussian sequences. In Electronic Proceedings of XXI IMEKO World Congress "Measurement in Research and Industry" September 2015 Prague Czech Republic. (Vol. 30, pp. 2088-2091).
19. Zabolotnii, S., Warsza, Z., Tkachenko, O.: Polynomial estimation of linear regression parameters for the asymmetric pdf of errors. Advances in Intelligent Systems and Computing. Vol. 743, pp. 758–772. Springer (2018). https://doi.org/10.1007/978-3-319-77179-3_75
20. Zabolotnii S.W., Warsza Z.L., Tkachenko O. (2020) Estimation of Linear Regression Parameters of Symmetric Non-Gaussian Errors by Polynomial Maximization Method. Advances in Intelligent Systems and Computing. Vol. 920. pp. 636–649 Springer, https://doi.org/10.1007/978-3-030-13273-6_59
21. Cook, R. D., & Weisberg, S. (1982). Residuals and Influence in Regression. Monographs on Statistics and Applied Probability. https://doi.org/10.2307/1269506
22. Jarque C.M., Bera A.K., (1987), A Tests of Observations and Regression Residuals. International Statistical Review, vo. 55, pp. 163-172.
23. Quinlan, J. R. (1993). Combining instance-based and model-based learning. In Proceedings of the Tenth International Conference on Machine Learning (pp. 236-243).