DIRECT SOLUTION OF POLYNOMIAL REGRESSION OF ORDER UP TO 3

Mykhaylo Dorozhovets, Dr.Sc., Prof.
Rzeszow University of Technology, Poland; michdor@prz.edu.pl
Lviv Polytechnic National University, Ukraine; e-mail: mykhailo.m.dorozhovets@lpnu.ua

Abstract. This article presents results related to the direct solution of the polynomial regression parameters based on the analytical solving of regression equations. The analytical solution is based on the normalization of the values of independent quantity with equidistance steps. The proposed solution does not need to directly solve a system of polynomial regression equations. The direct expressions to calculate estimators of regression coefficients, their standard deviations, and also standard and expanded deviation of polynomial functions are given. For a given number of measurement points, the parameters of these expressions have the same values independently of the range of input quantity. The proposed solution is illustrated by a numerical example used from a literature source.
Key words: Regression, Function, Polynomial; Estimation; Solution, Standard Deviation, Uncertainty

1. Introduction

In measurement practice and statistics, polynomial regression of order k is widely used to determine the relationship [1] – [8]:

$$Y = F(X) = \beta_0 + \beta_1 \cdot X + .. + \beta_k X^k + \varepsilon = \sum_{j=1}^{k} \beta_j \cdot X^j + \varepsilon \qquad (1)$$

between input (X) and output (Y) quantities of various systems, where $\boldsymbol{\beta} = (\beta_0, \beta_1, .., \beta_k)^T$ is a vector of k+1 regression coefficients; ε is normally distributed random noise with zero expected value (μ=0) and unknown standard deviation σ [1 - 6]. For example, relationship (1) in measurement is typically used to describe the conversion functions of sensors and other measurement systems [9], to build calibration function of measurement instruments and systems [10], [11], and to determine parameters of time series functions, etc. Using an appropriate number of measurement points, the goal is to determine the estimators of these coefficients and evaluate the accuracy of both the coefficients and the function itself.

Before using a sensor in a measurement system, we need to know its conversion function [9], which very often is described by polynomial function (1), and also its uncertainty. In the same situation, while calibrating the measuring instrument [10], [11], the main goal relates to establishing the relation between quantity values provided by measurement standards and the corresponding indications of a measuring system or instrument. The calibration curve usually is presented in the polynomial form (1).

Most of the theoretical and practical issues have been developed by various authors. There are very extensive literature sources on deep analysis of both theoretical and different practical aspects of regression, but only some of the literature sources are listed in [1] – [8]. In classical regression analysis, only the effect of random effects on the results of measuring the output quantity is usually considered. However, one should take into account the correlation of random effects as well as such effects on both the input and output quantities [12], [13].

Based on the general algorithm for polynomial regression of order up to 3, we present a modification of it, which provides a simplification of all stages of polynomial regression. Such a possibility arises in the case with an equal distance of values of the independent quantity. Then, by shifting the independent variable, all average values of odd powers of the values of input quantity are equal to zero. Due to this, in the system of equations, describing the regression, the number of non-zero coefficients is reduced by twice. The next modification is to normalize the values of the independent variable to the value of half of its range of variation. The application of these modifications provides independence of the coefficients of the normal system of equations from the absolute values and the range of variation of the independent quantity. I.e., with the same number of measurement points for any values of this quantity, the coefficients of the normal system of equations remain unchanged. Due to these modifications, all regression issues can be reduced to simple analytical relationships instead of direct solving a system of regression equations.

2. Drawbacks

A certain problem of applying the classical approach in regression is the need to solve systems of equations of the appropriate order, including the calculation of the inverse matrix [1] – [8]. Another problem is that with the same number of measurement points, but with a different range of variation of the independent quantity, all the elements of the system of equations, including the corresponding matrices, must be recalculated. This results in solving the system of equations again and recalculating the inverse matrix.

3. Goal

The goal of this article consists in the derivation of the simple analytical formulas, that make it possible to calculate directly the estimators of the coefficients, their standard deviation, and the standard deviation of the function with its expanded uncertainty, without the need to solve the system of equations and calculate the inverse matrix.

4. Modification of Classical Regression Solution

4.1. Classical regression solution

Traditionally, most publications on polynomial regression [1] – [11] to determinate regression coefficients $\mathbf{b}=(b_0, b_1, .., b_k)^T$ usually maximal likelihood estimation (MLE) or the least squares method (LSM) used [1] –[11]. In general, these coefficients are determined by n measurement points, i.e. pairs (x_i; y_i, $i=1, 2, \ldots, n$) of non-random values x_i (vector $\mathbf{x}=(x_1, x_2, .., x_n)^T$) of the input (independent) quantity X that are matched by values y_i (vector $\mathbf{y}=(y_1, y_2, .., y_n)^T$) of the output quantity Y. With some simplification, the results of the classical solution of the polynomial regression problem are the dependencies for estimating the regression coefficients $\mathbf{b}=(b_0, b_1, .., b_k)^T$ and the standard deviations of these estimators and also the regression function.

Since the main matrix formulas related to the classical procedure of solving polynomial regression will be used with proposed modifications of input quantity, this procedure is presented briefly here. When the values y_i of the output quantity Y are mutually uncorrelated ($\mathrm{cov}(y_i;y_j)=0$, $i\neq j$) then using LSM, the estimated values $\mathbf{b}=(b_0, b_1, .., b_k)^T$ of the regression coefficients $\boldsymbol{\beta}=(\beta_0, \beta_1, .., \beta_k)^T$ can be determined by matrix equation [1]- [6]:

$$\mathbf{b}=\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\cdot\mathbf{y}=\mathbf{M}^{-1}\mathbf{\Phi}^T\cdot\mathbf{y}\,, \tag{2}$$

here

$$\mathbf{\Phi}=\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^k \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix}, \tag{3}$$

$$\mathbf{M}=\left(\mathbf{\Phi}^T\mathbf{\Phi}\right),\quad \mathbf{M}^{-1}=\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}. \tag{4}$$

Because for the normally distributed random effects ε with standard deviation σ the distribution $p_{b_j}\left(b_j\middle|\beta_j,\sigma\right)$ of estimate b_j is normal too, and the standard deviation of the estimated regression coefficients is given by the formula [1] – [6]:

$$\sigma\left(b_j\right)=\sigma\cdot\sqrt{\left[\mathbf{M}^{-1}\right]_{j,j}}\,. \tag{5}$$

After estimation of the regression coefficients (vector $\mathbf{b}=(b_0, b_1, .., b_k)^T$) the dependence $y(x)$ between input X and output Y quantities can be presented by function:

$$y(x)=b_0+b_1\cdot x+..+b_k x^k=\sum_{j=1}^{k} b_j\cdot x^j\,. \tag{6}$$

Therefore, using the standard deviation of coefficients (5) the standard deviation of the estimated regression function (6) is given by the formula:

$$\sigma\left[y(x)\right]=\sigma\sqrt{\sum_{j=0}^{k}\sum_{i=0}^{k}\left[\mathbf{M}^{-1}\right]_{j,i}x^{j+i}}\,. \tag{7}$$

4.2. Matrix after modification of the values of input quantity

The first step of modification of the classical algorithm (2) – (4) consists of the following. Using denotations $\overline{x^j}$ for the arithmetic mean values x_i^j of the corresponding power j of x_i and $\overline{x^j y}$ for the arithmetic mean values of products $x_i^j\cdot y_i$:

$$\overline{x^j}=\frac{1}{n}\sum_{i=1}^{n}x_i^j\ ,j=0,2,..,k;\ \overline{x^j y}=\frac{1}{n}\sum_{i=1}^{n}x_i^j\cdot y_i\,,\quad j=0,\ldots,k, \tag{8}$$

matrix solution (2) can be presented directly in another form:

$$\mathbf{b}=\mathbf{M}_x^{-1}\cdot\mathbf{Y}\,, \tag{9}$$

here

$$\mathbf{M}_\mathbf{x}=\frac{1}{n}\cdot\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)=\begin{bmatrix} 1 & \overline{x} & \overline{x^2} & \cdots & \overline{x^{k-1}} & \overline{x^k} \\ \overline{x} & \overline{x^2} & \iddots & \overline{x^{k-1}} & \overline{x^k} & \overline{x^{k+1}} \\ \overline{x^2} & \iddots & \iddots & \iddots & \iddots & \vdots \\ \vdots & \overline{x^{k-1}} & \iddots & \iddots & \iddots & \overline{x^{2k-2}} \\ \overline{x^{k-1}} & \overline{x^k} & \overline{x^{m+1}} & \cdots & \overline{x^{2k-2}} & \overline{x^{2k-1}} \\ \overline{x^k} & \overline{x^{k+1}} & \cdots & \overline{x^{2k-2}} & \overline{x^{2k-1}} & \overline{x^{2k}} \end{bmatrix},$$

$$\mathbf{Y}=\frac{1}{n}\mathbf{\Phi}^T\mathbf{y}=\begin{bmatrix} \overline{y} \\ \overline{xy} \\ \overline{x^2 y} \\ \vdots \\ \overline{x^{k-1}y} \\ \overline{x^k y} \end{bmatrix}, \tag{10}$$

and

$$\mathbf{M}_x^{-1} = (\mathbf{M}_x)^{-1} \quad (11)$$

is an inverse matrix of the modified normal equation.

The next step of modification is the normalization of equidistance values of input quantity:

$$\chi_i = \frac{x_i - \overline{x}}{V} = \frac{2i}{n-1} - 1,\ i = 0,\dots,n-1\,, \quad (12)$$

here $\overline{x}$ is a mean value and $V = R/2$ is half of the range ($R = x_{max} - x_{min}$).

From (12) it can be seen that the range of new modified by (12) quantity is: $-1 \le \chi \le +1$ and the values of the variable χ_i are evenly symmetric. Then the values of the corresponding to (8) arithmetic mean values χ_i^j of the corresponding power j of χ_i and $\overline{y\chi^j}$ for the arithmetic mean values of products $y_i \cdot \chi_i^j$:

$$\overline{\chi^j} = \frac{1}{n}\sum_{i=0}^{n-1}\chi_i^j\,,\ \overline{y\cdot\chi^j} = \frac{1}{n}\sum_{i=0}^{n-1} y_i\cdot\chi_i^j\,,$$

$$j=0,\dots,k\,. \quad (13)$$

Due to odd symmetries χ_i from origin all odd components in the matrix (10) are equals to zero:

$$\overline{\chi^{2j+1}} = \frac{1}{n}\sum_{i=0}^{n-1}\chi_i^{2j+1} = 0\,. \quad (14)$$

In measurement practice, polynomial relationships (1) of the first (k=1), second (k=2), and third degree (k=3) are most often used. Therefore, for the linear, quadratic, and cubic polynomials the matrix components a given by expressions:

$$\mathbf{M1}\chi = \begin{vmatrix} 1 & 0 \\ 0 & \overline{\chi^2} \end{vmatrix};\quad \mathbf{Y1}_\chi = \begin{vmatrix} \overline{y} \\ \overline{y\cdot\chi} \end{vmatrix}. \quad (15)$$

$$\mathbf{M2}\chi = \begin{vmatrix} 1 & 0 & \overline{\chi^2} \\ 0 & \overline{\chi^2} & \overline{0} \\ \overline{\chi^2} & 0 & \overline{\chi^4} \end{vmatrix};\quad \mathbf{Y2}_\chi = \begin{vmatrix} \overline{y} \\ \overline{y\cdot\chi} \\ \overline{y\cdot\chi^2} \end{vmatrix}, \quad (16)$$

$$\mathbf{M3}_\chi = \begin{bmatrix} 1 & 0 & \overline{\chi^2} & 0 \\ 0 & \overline{\chi^2} & 0 & \overline{\chi^4} \\ \overline{\chi^2} & 0 & \overline{\chi^4} & 0 \\ 0 & \overline{\chi^4} & 0 & \overline{\chi^6} \end{bmatrix};\quad \mathbf{Y3}_\chi = \begin{vmatrix} \overline{y} \\ \overline{y\cdot\chi} \\ \overline{y\cdot\chi^2} \\ \overline{y\cdot\chi^3} \end{vmatrix}, \quad (17)$$

Three different coefficient denotations to distinguish the three regressions are used:

$$\mathbf{a}_\chi^T = \left(a_{\chi,0}; a_{\chi,1}\right) \text{ - linear,} \quad (18)$$

$$\mathbf{b}_\chi^T = \left(b_{\chi,0}; b_{\chi,1}; b_{\chi,2}\right) \text{ – quadratic,} \quad (19)$$

$$\mathbf{c}_\chi^T = \left(c_{\chi,0}; c_{\chi,1}; c_{\chi,2}; c_{\chi,3}\right) \text{ - cubic.} \quad (20)$$

The general solutions of these three matrix components (15), (16), and (17) can be presented as:

$$\mathbf{a}_\chi = \mathbf{M1}_\chi^{-1}\cdot\mathbf{Y1}_\chi\,, \quad (21)$$

$$\mathbf{b}_\chi = \mathbf{M2}_\chi^{-1}\cdot\mathbf{Y2}_\chi\,, \quad (22)$$

$$\mathbf{c}_\chi = \mathbf{M3}_\chi^{-1}\cdot\mathbf{Y3}_\chi\,. \quad (23)$$

For matrices (15), (16), and (17) the inverse matrices are:

$$\mathbf{M1}_\chi^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{1}{\overline{\chi^2}} \end{bmatrix}, \quad (24)$$

$$\mathbf{M2}_\chi^{-1} = \begin{bmatrix} \dfrac{\overline{\chi^4}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} & 0 & \dfrac{-\overline{\chi^2}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} \\ 0 & \dfrac{1}{\overline{\chi^2}} & \overline{0} \\ \dfrac{-\overline{\chi^2}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} & \overline{0} & \dfrac{1}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} \end{bmatrix} \quad (25)$$

$$M3_\chi^{-1} = \begin{bmatrix} \dfrac{\overline{\chi^4}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} & 0 & \dfrac{-\overline{\chi^2}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} & 0 \\ 0 & \dfrac{\overline{\chi^6}}{\overline{\chi^2}\cdot\overline{\chi^6} - \left(\overline{\chi^4}\right)^2} & 0 & \dfrac{-\overline{\chi^4}}{\overline{\chi^2}\cdot\overline{\chi^6} - \left(\overline{\chi^4}\right)^2} \\ \dfrac{-\overline{\chi^2}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} & 0 & \dfrac{1}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} & 0 \\ 0 & \dfrac{-\overline{\chi^4}}{\overline{\chi^2}\cdot\overline{\chi^6} - \left(\overline{\chi^4}\right)^2} & 0 & \dfrac{\overline{\chi^2}}{\overline{\chi^2}\cdot\overline{\chi^6} - \left(\overline{\chi^4}\right)^2} \end{bmatrix} \quad (26)$$

4.3. Analytical expressions for the coefficients and function

Due to the form of inverse matrices (24), (25), (26) it is possible to derive the analytical solutions for the regression coefficients in the forms:

$$a_{\chi,0} = \overline{y};\ a_{\chi,1} = \frac{\overline{y\cdot\chi}}{\overline{\chi^2}}\,; \quad (27)$$

$$b_{\chi,0} = \frac{\overline{\chi^4}\cdot\overline{y} - \overline{\chi^2}\cdot\overline{y\cdot\chi^2}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2};\ b_{\chi,1} = \frac{\overline{y\cdot\chi}}{\overline{\chi^2}};$$

$$b_{\chi,2} = \frac{\overline{y\cdot\chi^2} - \overline{\chi^2}\cdot\overline{y}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2}\,; \quad (28)$$

$$c_{\chi,0} = \frac{\overline{\chi^4}\cdot\overline{y} - \overline{\chi^2}\cdot\overline{y\cdot\chi^2}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2};\ c_{\chi,1} = \frac{\overline{\chi^6}\cdot\overline{y\cdot\chi} - \overline{\chi^4}\cdot\overline{y\cdot\chi^3}}{\overline{\chi^6}\cdot\overline{\chi^2} - \left(\overline{\chi^4}\right)^2};$$

$$c_{\chi,2} = \frac{\overline{y\cdot\chi^2} - \overline{\chi^2}\cdot\overline{y}}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2};\ c_{\chi,3} = \frac{\overline{\chi^2}\cdot\overline{y\cdot\chi^3} - \overline{\chi^4}\cdot\overline{y\cdot\chi}}{\overline{\chi^6}\cdot\overline{\chi^2} - \left(\overline{\chi^4}\right)^2} \quad (29)$$

Using values of estimated regression coefficients (27), (28), (29) it is possible to determine functions directly functions dependently on previously determined matrices components (15), (16), and (17):

$$y1(\chi) = \overline{y} + \frac{\overline{y\cdot\chi}}{\overline{\chi^2}}\cdot\chi\,; \quad (30)$$

$$y2(\chi)=\frac{\overline{\chi^4}\cdot\overline{y}-\overline{\chi^2}\cdot\overline{y\cdot\chi^2}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}+\frac{\overline{y\cdot\chi}}{\overline{\chi^2}}\cdot\chi+$$
$$+\frac{\overline{y\cdot\chi^2}-\overline{\chi^2}\cdot\overline{y}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}\cdot\chi^2; \qquad (31)$$

$$y3(\chi)=\frac{\overline{\chi^4}\cdot\overline{y}-\overline{\chi^2}\cdot\overline{y\cdot\chi^2}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}+\frac{\overline{\chi^6}\cdot\overline{y\cdot\chi}-\overline{\chi^4}\cdot\overline{y\cdot\chi^3}}{\overline{\chi^6}\cdot\overline{\chi^2}-\left(\overline{\chi^4}\right)^2}\cdot$$
$$\cdot\chi+\frac{\overline{y\cdot\chi^2}-\overline{\chi^2}\cdot\overline{y}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}\cdot\chi^2+\frac{\overline{\chi^2}\cdot\overline{y\cdot\chi^3}-\overline{\chi^4}\cdot\overline{y\cdot\chi}}{\overline{\chi^6}\cdot\overline{\chi^2}-\left(\overline{\chi^4}\right)^2}\cdot\chi^3. \qquad (32)$$

4.4. Analytical expressions for the variations and covariations of coefficients and function

For the definition of the matrix $\mathbf{M}_x$ in (10) and inverse matrix (11) for a known variation σ^2 the variation of the estimated coefficient c_j can be presented by the general formula:

$$\sigma^2(c_j)=\frac{\sigma^2}{n}\cdot\left[\mathbf{M}_x^{-1}\right]_{j,j} \qquad (33)$$

I.e. variations and nonzero covariations of corresponding regression coefficients are:

$$\sigma^2(a_{\chi,0})=\frac{\sigma^2}{n};\ \sigma^2(a_{\chi,1})=\frac{\sigma^2}{n}\cdot\frac{1}{\overline{\chi^2}} \qquad (34)$$

$$\sigma^2(b_{\chi,0})=\frac{\sigma^2}{n}\cdot\frac{\overline{\chi^4}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2};\ \sigma^2(b_{\chi,1})=\frac{\sigma^2}{n}\cdot\frac{1}{\overline{\chi^2}};$$
$$\sigma^2(b_{\chi,1})=\frac{\sigma^2}{n}\cdot\frac{1}{\overline{\chi^2}};\sigma^2(b_{\chi,2})=\frac{\sigma^2}{n}\cdot\frac{1}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2};$$
$$\operatorname{cov}(b_{\chi,0},b_{\chi,2})=\frac{\sigma^2}{n}\cdot\frac{-\overline{\chi^2}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}; \qquad (35)$$

$$\sigma^2(c_{\chi,0})=\frac{\sigma^2}{n}\cdot\frac{\overline{\chi^4}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2};\ \sigma^2(c_{\chi,1})=\frac{\sigma^2}{n}\cdot\frac{\overline{\chi^6}}{\overline{\chi^2}\cdot\overline{\chi^6}-\left(\overline{\chi^4}\right)^2};$$
$$\sigma^2(c_{\chi,2})=\frac{\sigma^2}{n}\cdot\frac{1}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2};\ \sigma^2(c_{\chi,3})=\frac{\sigma^2}{n}\cdot\frac{\overline{\chi^2}}{\overline{\chi^2}\cdot\overline{\chi^6}-\left(\overline{\chi^4}\right)^2};$$
$$\operatorname{cov}(b_{\chi,0},b_{\chi,2})=\frac{\sigma^2}{n}\cdot\frac{-\overline{\chi^2}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2};$$
$$\qquad (36)$$
$$\operatorname{cov}(b_{\chi 1},b_{\chi 3})=\frac{\sigma^2}{n}\cdot\frac{-\overline{\chi^4}}{\overline{\chi^6}\cdot\overline{\chi^2}-\left(\overline{\chi^4}\right)^2}.$$

Because in most practical cases the variance σ^2 of the random deviations in results y_i $(i=1,..,n)$ of measurement of output quantity is not known in (34) – (36) the unbiased estimators are used [1] - [8]:

$$S^2=\frac{1}{n-k-1}\sum_{i=1}^{n}\left(\sum_{j=0}^{k}b_j\cdot x_i^j-y_i\right)^2. \qquad (37)$$

Applying the denoting definition

$$s^2=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=0}^{k}b_j\cdot x_i^j-y_i\right)^2 \qquad (38)$$

of the so-called biased estimator of variation σ^2, for which $S^2=\frac{n}{n-k-1}s^2$, the biased estimators of variances can be of can be presented directly, without coefficients values:

$$s1_y^2=\overline{y^2}-\left(\overline{y}\right)^2-\frac{\left(\overline{y\cdot\chi}\right)^2}{\overline{\chi^2}}; \qquad (39)$$

$$s2_y^2=\overline{y^2}-\left(\overline{y}\right)^2-\frac{\left(\overline{y\cdot\chi}\right)^2}{\overline{\chi^2}}-\frac{\left(\overline{y\cdot\chi^2}-\overline{\chi^2}\cdot\overline{y}\right)^2}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}; \qquad (40)$$

$$s3_y^2=\overline{y^2}-\left(\overline{y}\right)^2-\frac{\left(\overline{y\cdot\chi}\right)^2}{\overline{\chi^2}}-\frac{\left(\overline{y\cdot\chi^2}-\overline{\chi^2}\cdot\overline{y}\right)^2}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}-$$
$$+\frac{\overline{y\cdot\chi^2}-\overline{\chi^2}\cdot\overline{y}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}\cdot\left(\frac{x-\overline{x}}{V}\right)^2, \qquad (41)$$

where $\overline{y^2}=\frac{1}{n}\sum_{i=1}^{n}y_i^2$.

4.5. Inverse substitutions

The dependencies of the estimated regression functions directly on the value of the input variable are obtained by substituting the values $\chi=\frac{x-\overline{x}}{V}$ into the corresponding relations (30) - (32):

$$y1(\chi)=\overline{y}+\frac{\overline{y\cdot\chi}}{\overline{\chi^2}}\cdot\left(\frac{x-\overline{x}}{V}\right); \qquad (42)$$

$$y2(\chi)=\frac{\overline{\chi^4}\cdot\overline{y}-\overline{\chi^2}\cdot\overline{y\cdot\chi^2}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}+\frac{\overline{y\cdot\chi}}{\overline{\chi^2}}\cdot\left(\frac{x-\overline{x}}{V}\right)+;$$
$$+\frac{\overline{y\cdot\chi^2}-\overline{\chi^2}\cdot\overline{y}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}\cdot\left(\frac{x-\overline{x}}{V}\right)^2 \qquad (43)$$

$$y3(\chi)=\frac{\overline{\chi^4}\cdot\overline{y}-\overline{\chi^2}\cdot\overline{y\cdot\chi^2}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}+\frac{\overline{\chi^6}\cdot\overline{y\cdot\chi}-\overline{\chi^4}\cdot\overline{y\cdot\chi^3}}{\overline{\chi^6}\cdot\overline{\chi^2}-\left(\overline{\chi^4}\right)^2}\cdot$$
$$\cdot\left(\frac{x-\overline{x}}{V}\right)+\frac{\overline{y\cdot\chi^2}-\overline{\chi^2}\cdot\overline{y}}{\overline{\chi^4}-\left(\overline{\chi^2}\right)^2}\cdot\left(\frac{x-\overline{x}}{V}\right)^2+$$
$$+\frac{\overline{\chi^2}\cdot\overline{y\cdot\chi^3}-\overline{\chi^4}\cdot\overline{y\cdot\chi}}{\overline{\chi^6}\cdot\overline{\chi^2}-\left(\overline{\chi^4}\right)^2}\cdot\left(\frac{x-\overline{x}}{V}\right)^3. \qquad (44)$$

From (27) – (29) and (44)- (46) the initial values of estimated coefficients are given as:

$$a_0=a_{\chi,0}-a_{\chi,1}\cdot\frac{\overline{x}}{V};\quad a_1=\frac{a\chi_1}{V}. \qquad (45)$$

$$b_0 = b_{\chi,0} - b_{\chi,1} \cdot \frac{\bar{x}}{V} + \frac{b_{\chi,2} \cdot (\bar{x})^2}{V^2};$$

$$b_1 = \frac{b_{\chi,1}}{V} - \frac{2b_{\chi,2} \cdot \bar{x}}{V^2}; \cdot b_2 = \frac{b_{\chi,2}}{V^2}. \quad (46)$$

$$c_0 = c_{\chi,0} - c_{\chi,1} \cdot \frac{\bar{x}}{V} + \frac{c_{\chi,2}(\bar{x})^2}{V^2} - \frac{c_{\chi,3}(\bar{x})^3}{V^3};$$

$$c_1 = \frac{c_{\chi,1}}{V} - \frac{2c_{\chi,2} \cdot \bar{x}}{V^2} + \frac{3c_{\chi,3} \cdot (\bar{x})^2}{V^3};$$

$$c_2 = \frac{c_{\chi 2}}{V^2} - \frac{3c_{\chi 3} \cdot \bar{x}}{V^3}; \ c_3 = \frac{c_{\chi,3}}{V^3} \quad (47)$$

Estimated variances of coefficients can be determined by formulas:

$$s^2(a_0) = \frac{s1^2}{n-2}\left(1 + \frac{1}{\overline{\chi^2}} \cdot \left(\frac{\bar{x}}{V}\right)^2\right);$$

$$\sigma^2(a_1) = \frac{\sigma^2}{n} \cdot \frac{1}{\overline{\chi^2} \cdot V^2}; \quad (48)$$

$$s^2(b_0) = \frac{s2^2}{n-3} \cdot \left[1 + \frac{\left[(\bar{x}/V)^2 - \left(\overline{\chi^2}\right)\right]^2}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} + \frac{1}{\overline{\chi^2}} \cdot \left(\frac{\bar{x}}{V}\right)^2\right];$$

$$s^2(b_1) = \frac{s2^2}{n-3} \cdot \frac{1}{V^2}\left[\frac{1}{\overline{\chi^2}} + 4\left(\frac{\bar{x}}{V}\right)^2 \cdot \frac{1}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2}\right]; \quad (49.1)$$

$$s^2(b_2) = \frac{s2^2}{n-3} \cdot \frac{\bar{1}}{V^4\left(\overline{\chi^4} - \left(\overline{\chi^2}\right)^2\right)}; \quad (49,2)$$

$$s^2(c_0) = \frac{s3^2}{n-4} \cdot \left[1 + \frac{\left[(\bar{x}/V)^2 - \left(\overline{\chi^2}\right)\right]^2}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} + \right.$$

$$\left. + \frac{1}{\overline{\chi^2}} \cdot \left(\frac{\bar{x}}{V}\right)^2 \frac{\left[\overline{\chi^2} \cdot (\bar{x}/V)^2 - \left(\overline{\chi^4}\right)\right]^2}{\overline{\chi^6} \cdot \overline{\chi^2} - \left(\overline{\chi^4}\right)^2}\right]; \quad (50.1)$$

$$s^2(c_1) = \frac{s3^2}{n-4} \cdot \frac{1}{V^2} \cdot \left[1 + \frac{4}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2}\left(\frac{\bar{x}}{V}\right)^2 + \right.$$

$$\left. + \frac{\left(3\overline{\chi^2} \cdot (\bar{x}/V)^2 - \left(\overline{\chi^4}\right)\right)^2}{\overline{\chi^2} \cdot \left(\overline{\chi^6} \cdot \overline{\chi^2} - \left(\overline{\chi^4}\right)^2\right)}\right]; \quad (50.2)$$

$$s^2(c_2) = \frac{s3^2}{n-4} \cdot \frac{1}{V^4}\left(\frac{1}{\overline{\chi^4} - \left(\overline{\chi^2}\right)^2} + \frac{9\overline{\chi^2}}{\overline{\chi^2} \cdot \overline{\chi^6} - \left(\overline{\chi^4}\right)^2}\left(\frac{\bar{x}}{V}\right)^2\right),$$

$$s^2(c_3) = \frac{s3^2}{n-4} \cdot \cdot \frac{1}{V^6} \cdot \frac{\overline{\chi^2}}{\overline{\chi^2} \cdot \overline{\chi^6} - \left(\overline{\chi^4}\right)^2}, \quad (50.3)$$

Estimated variances of regression functions:

$$s_y^2(y1(x)) = \frac{s1^2}{n-2}\left[1 + \frac{(x-\bar{x})^2}{\overline{\chi^2} \cdot V^2}\right]; \quad (51)$$

$$s_y^2(y2(x)) = \frac{s2^2}{n-3}\left[1 + \frac{\left((x-\bar{x})^2 - \overline{\chi^2} \cdot V^2\right)^2}{\left(\overline{\chi^4} - \left(\overline{\chi^2}\right)^2\right) \cdot V^4} + \frac{(x-\bar{x})^2}{\overline{\chi^2} \cdot V^2}\right]; \quad (52)$$

$$s_y^2(y3(x)) = \frac{s3^2}{n-4}\left[1 + \frac{\left((x-\bar{x})^2 - \overline{\chi^2} \cdot V^2\right)^2}{\left(\overline{\chi^4} - \left(\overline{\chi^2}\right)^2\right) \cdot V^4} + \right.$$

$$\left. + \frac{(x-\bar{x})^2}{\overline{\chi^2} \cdot V^2} \cdot \left(1 + \frac{\left(\overline{\chi^2} \cdot (x-\bar{x})^2 - \overline{\chi^4} \cdot V^2\right)^2}{\left(\overline{\chi^6} \cdot \overline{\chi^2} - \left(\overline{\chi^4}\right)^2\right) \cdot V^4}\right)\right]. \quad (53)$$

Because for the polynomial regression of order k the ratios $\frac{\beta_j - b_j}{s} = t_j$ have t-Student distribution with $d = n - k - 1$ degree of freedom, therefore for a confidence level p expanded uncertainties of coefficients $U_p(\beta_j)$ and regression function $U_p(Y(x))$ can be calculated from (48) – (53) using coverage factor $k_p = t_p(n-k-1)$ from t-Student distribution:

$$U_p(\beta_j) = t_p(n-k-1) \cdot \sqrt{s^2(\beta_j)};$$

$$U_p(Y(x)) = t_p(n-k-1) \cdot \sqrt{s^2(y(x))}. \quad (54)$$

5. Example

In this example, the cubic polynomial model ($k = 3$) is built as example 8.1 in [3]. This example relates to the treatment of algae density measures over time (t). The $n = 14$ solutions were randomly assigned for measurement to one of each of 14 successive days (n=14) of the study,

i.e. t= *day* =1 2 3 4 56 7 8 9 10 11 12 13 14.

The dependent variable (Y=*algae density*) reported is a log-scale measurement of the increased absorbance of light by the solution:

In the first replicate: $Y1$ = 0.530 1.183 1.603 1.994 2.708 3.006 3.867 4.059 4.349 4.699 4.983 5.100 5.288 5.374.

In the second replicate, $Y2$ = 0.184 0.664 1.553 1.910 2.585 3.009 3.403 3.892 4.367 4.551 4.656 4.754 4.842 4.969.

For the data given above in [3] a cubic polynomial model (k = 3) is applied which results in:

$$y(t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3 .$$

The first step is the determination of normalized values χ_i of input quantity (days). For the given data the mean value $\bar{x} = 7.5$ and half range V = 6.5. Therefore, due to (12):

χ_i =-1 -11/13 -9/13 -7/13 -5/13 -3/13 -1/13 1/13 3/13 5/13 7/13 9/13 11/13 1.

Detailed results will be presented applying only to the first replicate. The final results will be presented to both replicates. Due to (13) parameters of corresponding mean values are:

$$\overline{\chi^2} = 0.38462;\ \overline{\chi^4} = 0.26445;\ \overline{\chi^6} = 0.21498;$$

$$\bar{y} = 3.48164;\ \overline{y \cdot \chi} = 0.95988;\ \overline{y \cdot \chi^2} = 1.23621;$$

$$\overline{y \cdot \chi^3} = 0.64913$$

Therefore matrix (XX) components are:

$$M_x = \begin{bmatrix} 1 & 0 & 0.38462 & 0 \\ 0 & 0.38462 & 0 & 0.26445 \\ 0.38462 & 0 & 0.26445 & 0 \\ 0 & 0.26445 & 0 & 0.21498 \end{bmatrix}$$

$$Y1_x = \begin{bmatrix} 3.48164 \\ 0.95988 \\ 1.23621 \\ 0.64913 \end{bmatrix}.$$

In usual regression solution matrix $\mathbf{M} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)$ (4) is:

$$\mathbf{M} = \begin{bmatrix} 14 & 105 & 1.015 \cdot 10^3 & 1.1025 \cdot 10^4 \\ 105 & 1.015 \cdot 10^3 & 1.1025 \cdot 10^4 & 1.27687 \cdot 10^5 \\ 1.015 \cdot 10^3 & 1.1025 \cdot 10^4 & 1.27687 \cdot 10^5 & 1.53982 \cdot 10^6 \\ 1.1025 \cdot 10^4 & 1.27687 \cdot 10^5 & 1.53982 \cdot 10^6 & 1.90923 \cdot 10^7 \end{bmatrix}.$$

From a comparison of these matrices, we can see that using the proposed modification we obtain a simple matrix with limited values of its components. Besides it, the condition number, (which can be as multiplied coefficients of random influences) is equal to about 53 and $5.1 \cdot 10^7$, i.e. about $\cdot 10^6$ bigger. I.e., even using the usual method of solving a system of regression equations is simpler and more accurate.

The values of estimates of coefficients $b_{\chi,j}$ (29) are:

$b_{\chi,0}$= 3.8212, $b_{\chi,1}$= 2.7210, $b_{\chi,2}$= -0.88293, $b_{\chi,3}$= -0.32766.

The values of estimates recalculation by (47) are:

b_0= 0.009478, b_1= 0.53074, b_2= 0.005947, b_3= -0.001193,

which are consistent with [3].

From (41) the biased estimate of variation: $s_y^2 = 0.009756$. Therefore from (49) the standard deviations of these estimates are:

$s(b_0)$ = 0.1676, $s(b_1)$ = 0.09343, $s(b_2)$ = 0.01422, $s(b_3)$ = 0.000625.

which are consistent with values determined by the classical method in [3].

From (53) using the values of parameters $\bar{x} = \bar{t} = 7.5$, V=6.5, $\overline{\chi^2} = 0.38462$, $\overline{\chi^4} = 0.26445$; $\overline{\chi^6} = 0.21498$ and $\overline{\chi^2} V^2 = 16.25$, $\overline{\chi^4} V^2 = 11.1731$ $\left(\overline{\chi^4} - \left(\overline{\chi^2}\right)^2\right) V^4 = 208$, $\left(\overline{\chi^6} \cdot \overline{\chi^2} - \left(\overline{\chi^4}\right)^2\right) V^4 = 22.7625$ the standard deviations of function related to first replicate is:

$$s(y1(t)) = \sqrt{\frac{0.009756}{10}\left[1 + \frac{\left((t-7.5)^2 - 16.25\right)^2}{208} + \frac{(t-7.5)^2}{16.25} \cdot \left(1 + \frac{\left(0.38462 \cdot (t-7.5)^2 - 11.1731\right)^2}{22.7625}\right)\right]}.$$

Expanded uncertainty of function related to the first replicate is:

$$U_p(y1(t)) = t_p(10)\sqrt{\frac{0.009756}{10}\left[1 + \frac{\left((t-7.5)^2 - 16.25\right)^2}{208} + \frac{(t-7.5)^2}{16.25} \cdot \left(1 + \frac{\left(0.38462 \cdot (t-7.5)^2 - 11.1731\right)^2}{22.7625}\right)\right]}.$$

For the second replicate the values of estimates are:

b_0= -0.55173, b_1= 0.69885, b_2= -0.01263, b_3= -0.0006796.

For the second replicate, the biased estimate of variation is: $s_y^2 = 0.007202$. Therefore from (49) the standard deviations of these estimates are:

$s(b_0)$= 0.144, $s(b_1)$= 0.0803, $s(b_2)$= 0.0122, $s(b_3)$= 0.000537.

From (53) and (54) we can see that the same values of input quantity standard deviation and expanded uncertainty of function differ only by values of estimated standard deviation $s_y = \sqrt{s_y^2}$. Therefore the standard deviation and expanded uncertainty of function related to the second replicate are:

$$s(y2(t))=\sqrt{\frac{0.007202}{10}\left[1+\frac{\left((t-7.5)^2-16.25\right)^2}{208}+\frac{(t-7.5)^2}{16.25}\cdot\left(1+\frac{\left(0.38462\cdot(t-7.5)^2-11.1731\right)^2}{22.7625}\right)\right]},$$

$$U_p(y2(t))=t_p(10)\sqrt{\frac{0.007202}{10}\left[1+\frac{\left((t-7.5)^2-16.25\right)^2}{208}+\frac{(t-7.5)^2}{16.25}\cdot\left(1+\frac{\left(0.38462\cdot(t-7.5)^2-11.1731\right)^2}{22.7625}\right)\right]}.$$

The expanded uncertainties $U_p(y(t))$ of both regression functions for a confidence level of p=0.95, which $k_{0.95}=t_{0.95}(10)=2.228$, are sown in Fig. 1, a. Expanded uncertainty with the part for the forward prediction is shown in Fig. 1,b.

From the results obtained and from Fig.1, we see that despite the different values of the input quantity in the two data series, in both cases the standard deviations and the expanded uncertainties are described by the same relationships only with different values of the estimated variances as scale factors. In addition, Fig. 1, b shows that if based on the estimated regression function we want to forecast the values of the function beyond the given range of variability of the input quantity, the uncertainty of such a forecast increases rapidly. For example, if we set a forecast for the 16th day (2 days ahead), the forecast uncertainty increases almost 3 times! This is a known effect [7] but is sometimes forgotten.

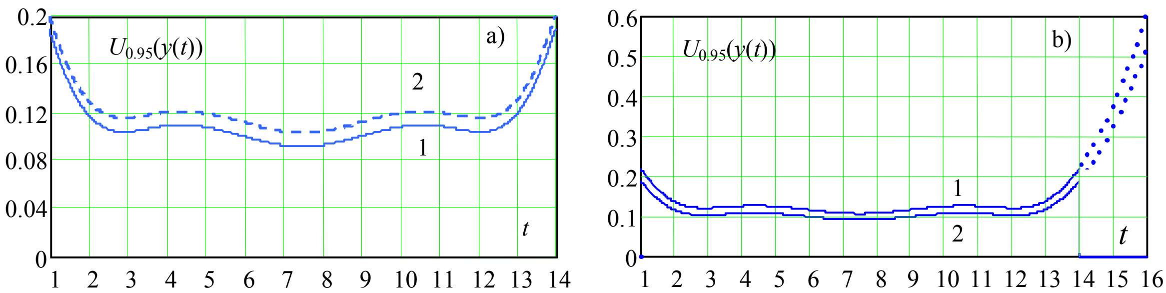

Fig. 1. Estimated expanded uncertainties of functions (a), expanded uncertainty for the prediction (b)
1 – first replicate, 2 – second replicate

5. Conclusions

The obtained results can be applied to polynomial regression of order up to 3 with a constant interval of the independent (input) variable.

The use of a modification of the values of the input quantity, which consists in centering these values by the mean and subsequent normalization to the value of half of the range, provides the possibility of deriving analytical relationships both for the determination of coefficients and regression functions as well as for determining all values of estimated variances, standard deviations, and uncertainties.

Relationships have been derived for the direct calculation of parameters, related to the regression analysis, the main ones are:

– coefficient estimators and regression functions of orders 1, 2, and 3:

– standard deviations of coefficient estimators and regression functions:

– uncertainties of the expanded regression function.

The derived formulas depend only on the means of the 2nd, 4th and 6th powers of the normalized values of the input quantity, and the means of the products of the output quantity and zero, 1st, 2nd, and 3rd powers of the normalized values of input quantity, as well as the mean and half of range of the input quantity.

I.e., a complete regression analysis can be performed easier without the necessity to directly solve the corresponding system of equations and calculate the corresponding inverse matrices.

The obtained results show that for the same number of measurement points (with a constant distance), the standardized values of the input quantity and the associated coefficient used to calculate the regression coefficients and functions and also corresponding estimated standard deviations are the same for regressions of a given regression. This simplifies performing regression analysis for different sets of the output quantity.

A numerical example, taken from a literature source, shows all the steps of the analysis, concerning polynomial regression of order 3, performed only by analytical dependences which gives the same results without performing operations with systems of equations and calculating the inverse matrix.

6. Gratitude

The author expresses gratitude to the Staff of the Department of Information-Measuring Technologies of Lviv Polytechnic National University.

References

[1] M.G. Kendall and A. Stuart. (1966) *The advanced theory of statistics*. Vol. 2. *Inference and relationship*. Second Edition. Chares Griffin and Company Limited, London. https://archive.org/details/in.ernet.dli.2015.212877.

[2] DRAPER, N. and SMITH, H. (1981) *Applied Regression Analysis*. 2nd ed., Wiley, New York. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118625590

[3] J.O. Rawlings, S. G. Pantula, D. A. Dickey. (1998) Applied Regression Analysis: A Research Tool, Second Edition, Springer-Verlag. New York, Berlin, Heidelberg. https://doi.org/10.1007/b98890.

[4] Handbook of Applicable Mathematics. Volume VI: Statistics, Part A, (1984). Edited by Emlyn Lloyd, John Wiley and Sons. https://www.amazon.com/Handbook-Applicable-Mathematics-Statistics-6/dp/0471900249.

[5] J.H. Pollard. (1977) Handbook of Numerical and Statistical Techniques. Cambridge University Press. https://cc.bingj.com/cache.aspx?q=J.H.+Pollard.+(1977)+Handbook+of+Numerical+and+Statistical+Techniques.&d=4579616625657925&mkt=en-WW&setlang=en-US&w=lYKqOOVXlYsikZD-qV7p9HJrRgt2oDA3.

[6] Applied Linear Regression, 3rd ed. Willey, Hoboken (2005). https://openlibrary.org/books/OL3306403M/Applied_linear_regression.

[7] J. R. Taylor (1982). *An introduction to error analysis*. University Science Books Mill Valley, California 1982. https://openlibrary.org/books/OL3786923M/An_introduction_to_error_analysis

[8] B. Forbes, (2009) Parameter estimation based on least-squares methods. In F. Pavese and A. B. Forbes, editors, Data modeling for metrology and testing in measurement science, New York. Birkauser-Boston. https://link.springer.com/chapter/10.1007/978-0-8176-4804-6_5.

[9] G. Mejer. (2008) *Smart Sensor Systems.* John Wiley, 2008. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470866931/

[10] Evaluation of measurement data—Guide to the expression of uncertainty in measurement Joint Committee for Guides in Metrology, JCGM 100: 2008. https://www.iso.org/sites/JCGM/GUM/JCGM100/.

[11] JCGM 101:2008. Evaluation of measurement data—Supplement 1 to the 'Guide to the Expression of Uncertainty in Measurement'— propagation of distributions using a Monte Carlo method'. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML. https://www.bipm.org/documents/20126/2071204/JCGM.

[12] Dorozhovets M. Study of the effect of correlation of observation results on the uncertainty of linear regression. Pomiary, Automatyka, Kontrola. N12, 2008, s.31-34 (in Polish). https://bibliotekanauki.pl/issues/8204.

[13] Dorozhovets M. Include measurement uncertainty of both quantities in linear regression. Pomiary, Automatyka, Kontrola. N9, 2008, pp.-612-615 (in Polish). https://bibliotekanauki.pl/issues/8089.