Journal of Mathematical Problems, Equations and Statistics

E-ISSN: 2709-9407
P-ISSN: 2709-9393
JMPES 2025; 6(2): 35-39

www.mathematicaljournal.com
Received: 18-05-2025
Accepted: 21-06-2025

Mehmet Pakdemirli
Department of Mechanical Engineering, Manisa Celal Bayar University, Muradiye, Manisa, Turkey

Corresponding Author:
Mehmet Pakdemirli
Department of Mechanical Engineering, Manisa Celal Bayar University, Muradiye, Manisa, Turkey

Polynomial regression analysis based on constrained data

Mehmet Pakdemirli

DOI: https://www.doi.org/10.22271/math.2025.v6.i2a.219

Abstract
An arbitrary order polynomial regression analysis is presented. Under the assumption of satisfaction of some reference points exactly by the polynomial, the general theory is given for the analysis. The polynomials are expressed in a suitable form so that the lower order coefficients can be calculated easily from the restrictions. The remaining coefficients are calculated from the error minimization. The theory is then applied to linear, quadratic and cubic polynomial regression sample problems having constrained data. Compared to the unrestrained regression analysis, the method reduces the computational cost of calculating coefficients.

Keywords: Polynomials, linear and nonlinear regression, constrained regression, data approximation

1. Introduction

In statistics, bivariate analysis is used to represent the empirical relationship between two variables [1]. One of the common methods of bivariate analysis is the regression analysis [1, 2]. The experimental data is obtained in a discrete form. The aim of the regression analysis is to obtain a continuous relationship between the variables so that the high cost of experimentation can be avoided for intermediate points and possible extrapolations with the aim of reducing the experimentation deviations to a reasonable approximate form. The simplest of all is the linear regression in which a line is searched which minimizes the total square of errors of each point. The nature of the problem may require a nonlinear relationship in which case appropriate nonlinear functions with sufficient number of parameters are employed to express approximate analytical representation of the data. Rather than dealing with formulas, a conceptual qualitative approach for nonlinear regression analysis was discussed [3]. When the linear regression is inadequate, polynomials of higher orders are used in expressing the relationship more precisely. Researchers worked on the problem of determining the minimum highest order of polynomials to reduce the complexity and computational cost of determining high number of parameters. Anderson [4] proposed a procedure in which a predetermined order is selected for the polynomial. If, in the regression analysis, the coefficient of the highest order is equal to zero, then the degree of the polynomial is lowered by one. The procedure is repeated until the coefficient of highest order is nonzero.

This rather restrictive assumption of vanishing highest order coefficient is interchanged with the negligibly small highest order coefficient recently [5]. Using the normalized data, perturbation theory [6, 7] is used to develop the necessary criteria for elimination of the highest orders. The analysis of [5] was then generalized to arbitrary multiple base functions instead of polynomials with the aid of perturbation theory [8]. In fact, perturbation theory was proven to be extremely useful in estimating magnitudes of roots of polynomials [9]. Other works based on Anderson's approach were due to Dette [10] and Dette and Studden [11]. Non-parametric methods were also used for polynomial regressions [12]. Another approach is to employ Chebyshev polynomials for reduction of the degree of the polynomials [13]. A new method based on neural networks was also proposed for polynomial regressions [14]. In an application data of batteries, it is shown that higher order polynomials better represent the relationship between current and voltage [15].

Constraints can also be imposed on the regression analysis. Shape constraints such as non-negativity/non-positivity, monotonically increasing/decreasing, convexity/concavity of the regression function were discussed [16]. In this study, data point constraints for polynomial regressions are treated for the first time. It may happen that, some of the data may be a reference data where the polynomial function should pass exactly through these reference points without an error. In those cases, the number of parameters to be calculated in the regression analysis is reduced by the number of reference data points. Suitable alternative expressions for such trial polynomials are given which enables easy calculation of the non-regression coefficients. Regression analysis is then applied to a reduced order system in terms of the parameters. The general theory is discussed first. Applications of the theory to linear, quadratic and cubic polynomials are depicted via worked examples. The method enables usage of higher order polynomials with reduced computational algebra.

2. Theory

Consider an n'th order polynomial regression of data (x_i, y_i), i=0,1,2,…m, from which the first k+1 points (x_j, y_j), j=0,1,2,…k are satisfied exactly by the polynomial approximation. For consistency, note that $k<n<m$. The polynomial is cast into a suitable form

$$p_n(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \cdots + a_k(x-x_0)(x-x_1)\ldots(x-x_{k-1}) + a_{k+1}(x-x_0)^2(x-x_1)\ldots(x-x_{k-1}) + \cdots + a_{2k}(x-x_0)^2(x-x_1)^2\ldots(x-x_{k-1})^2 + \cdots + a_n(x-x_0)^{n_0}(x-x_1)^{n_1}\ldots(x-x_{k-1})^{n_k} \quad (1)$$

Where

$$n = \sum_{i=0}^{k} n_i \,, n_i - n_{i+1} = 1 \vee 0 \,, i = 0,1,2,\ldots k-1 \,, \quad (2)$$

The powers n_i should be evenly distributed in the representation. In this suitable form, the first $a_0 - a_k$ coefficients can easily be calculated via the below recursive relations

$$a_0 = y_0, a_1 = \frac{y_1 - a_0}{x_1 - x_0}, a_j = \frac{y_j - a_0 - \sum_{i=1}^{j-1} a_i \prod_{\ell=0}^{i-1}(x_j - x_\ell)}{\prod_{\ell=0}^{j-1}(x_j - x_\ell)} \,, j = 2,3,\ldots k \quad (3)$$

which are derived by exact satisfaction of the polynomial function at the reference data points.
As an example, for seventh order polynomial with two reference points, the polynomial suggested will have the below form

$$p_7(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + a_3(x-x_0)^2(x-x_1)$$
$$+a_4(x-x_0)^2(x-x_1)^2 + a_5(x-x_0)^3(x-x_1)^2$$
$$+a_6(x-x_0)^3(x-x_1)^3 + a_7(x-x_0)^4(x-x_1)^3 \quad (4)$$

The remaining coefficients are calculated from minimizing the sum of square errors

$$\frac{\partial(e^2)}{\partial a_i} = 0, i = k+1, \ldots, n \,, \quad (5)$$

where

$$e^2 = \sum_{i=k+1}^{m} (y_i - p_n(x_i))^2 \quad . \quad (6)$$

The standard error in the regression analysis is [17]

$$S_{y/x} = \left[\frac{1}{m-(n+1)} \sum_{i=k+1}^{m} (y_i - p_n(x_i))^2\right]^{1/2} \,, \quad (7)$$

3. Linear Regression

Assume that one of the data points (x_0, y_0) is a reference point and one requires the line to pass through it. Then, the specific form of the equation is from (1)

$$p_1(x) = a_0 + a_1(x-x_0) \quad (8)$$

with $a_0 = y_0$ from (3). The remaining coefficient which is the slope of the line has to be determined from the regression analysis. Substituting (8) into (6) and then into (5), the coefficient is

$$a_1 = \frac{\sum_{i=1}^{m}(y_i - y_0)(x_i - x_0)}{\sum_{i=1}^{m}(x_i - x_0)^2} \quad (9)$$

The following example is given.

Example 1. Consider the data x=[3 0 1 2 4 5 6 7 8 9 10]; y=[2 1 2 1 3 3 4 3 4 5 5] with a reference point of $(x_0, y_0) = (3,2)$. The linear equation passing through this reference point is

$$p_1(x) = 2 + a_1(x-3) \quad (10)$$

From (9), $a_1 = 0.4286$. The plot of the data and the approximate line is given in Figure 1.

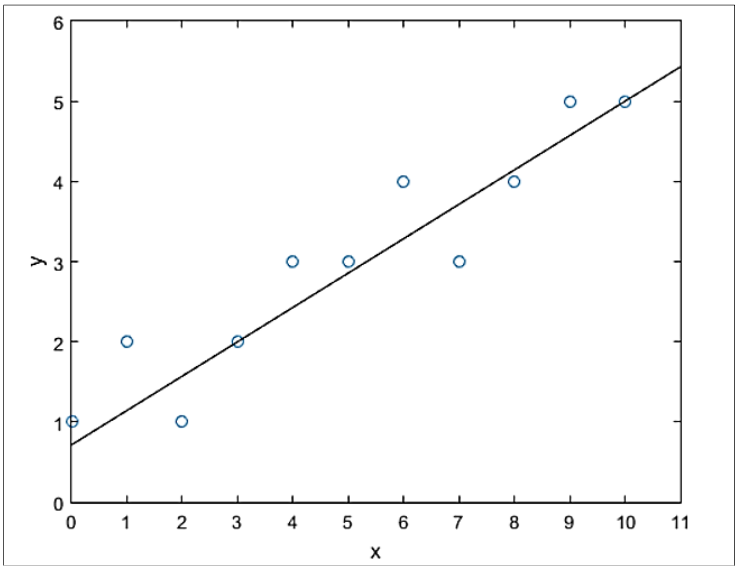

Fig 1: Comparison of the original data (o) with the regression line subject to the constraint $(x_0, y_0) = (3,2)$

From (7), the standard error is $S_{y/x} = 0.5492$.

3. Quadratic Regression

The analysis is done for only one reference data point (x_0, y_0). The quadratic polynomial is

$$p_1(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 \tag{11}$$

with $a_0 = y_0$ from (3). The remaining coefficients are determined from the regression analysis. Substituting (11) into (6) and then into (5), solving the linear algebraic equations, the coefficients are

$$a_1 = \frac{\sum_{i=1}^{m}(y_i - y_0)(x_i - x_0)\sum_{i=1}^{m}(x_i - x_0)^4 - \sum_{i=1}^{m}(y_i - y_0)(x_i - x_0)^2\sum_{i=1}^{m}(x_i - x_0)^3}{\sum_{i=1}^{m}(x_i - x_0)^2\sum_{i=1}^{m}(x_i - x_0)^4 - \left(\sum_{i=1}^{m}(x_i - x_0)^3\right)^2} \tag{12}$$

$$a_2 = \frac{\sum_{i=1}^{m}(y_i - y_0)(x_i - x_0)^2\sum_{i=1}^{m}(x_i - x_0)^2 - \sum_{i=1}^{m}(y_i - y_0)(x_i - x_0)\sum_{i=1}^{m}(x_i - x_0)^3}{\sum_{i=1}^{m}(x_i - x_0)^2\sum_{i=1}^{m}(x_i - x_0)^4 - \left(\sum_{i=1}^{m}(x_i - x_0)^3\right)^2} \tag{13}$$

The following data is given.

Example 2. Consider the data *x*=[4 0 1 2 3 5 6 7 8 9 10]; *y*=[9 4 4 2 7 13 26 30 49 60 70] with a reference point of $(x_0, y_0) = (4,9)$. The quadratic polynomial passing through this reference point is

$$p_1(x) = 9 + a_1(x - 4) + a_2(x - 4)^2 \tag{14}$$

From (12) and (13), $a_1 = 5.0283$ and $a_2 = 0.9518$. The data and its approximation function are contrasted in Figure 2.

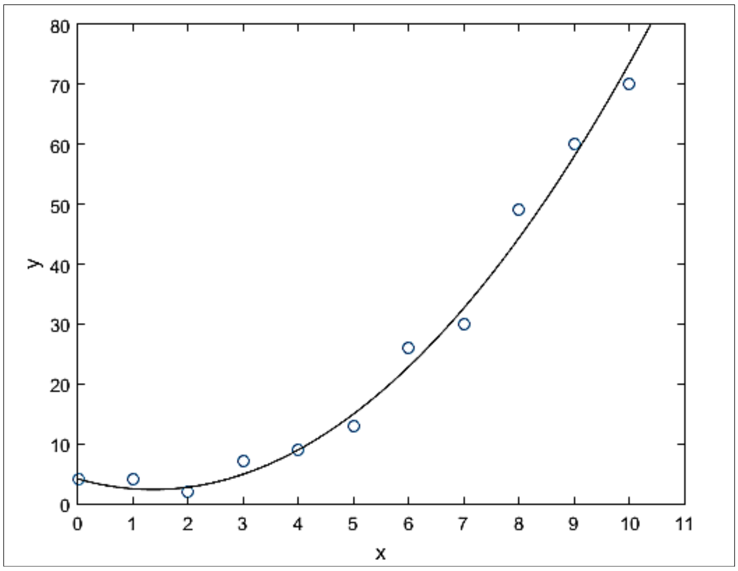

Fig 2: Comparison of the original data (o) with the quadratic regression subject to the constraint $(x_0, y_0) = (4,9)$

From (7), the standard error is $S_{y/x} = 2.8664$.

Three points determine a parabola in the two dimensional space. Hence, the maximum number of constraints can be at most 2 for a quadratic regression analysis. Depending on the specific data, increasing the number of constraints may increase the standard error also.

4. Cubic Regression

The analysis is done for two reference data point (x_0, y_0) and (x_1, y_1). The suitable form of the cubic polynomial is

$$p_1(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)^2(x - x_1) \qquad (15)$$

with $a_0 = y_0$ and $a_1 = \frac{y_1 - a_0}{x_1 - x_0}$ from (3). The remaining coefficients are determined from the regression analysis. Substituting (15) into (6) and then into (5), solving the linear algebraic equations, the coefficients are

$$a_2 = \frac{\begin{bmatrix} \sum_{i=2}^{m}(y_i - a_0 - a_1(x_i - x_0))(x_i - x_0)(x_i - x_1) & \sum_{i=2}^{m}(x_i - x_0)^3(x_i - x_1)^2 \\ \sum_{i=2}^{m}(y_i - a_0 - a_1(x_i - x_0))(x_i - x_0)^2(x_i - x_1) & \sum_{i=2}^{m}(x_i - x_0)^4(x_i - x_1)^2 \end{bmatrix}}{\begin{bmatrix} \sum_{i=2}^{m}(x_i - x_0)^2(x_i - x_1)^2 & \sum_{i=2}^{m}(x_i - x_0)^3(x_i - x_1)^2 \\ \sum_{i=2}^{m}(x_i - x_0)^3(x_i - x_1)^2 & \sum_{i=2}^{m}(x_i - x_0)^4(x_i - x_1)^2 \end{bmatrix}} \qquad (16)$$

$$a_3 = \frac{\begin{bmatrix} \sum_{i=2}^{m}(x_i - x_0)^2(x_i - x_1)^2 & \sum_{i=2}^{m}(y_i - a_0 - a_1(x_i - x_0))(x_i - x_0)(x_i - x_1) \\ \sum_{i=2}^{m}(x_i - x_0)^3(x_i - x_1)^2 & \sum_{i=2}^{m}(y_i - a_0 - a_1(x_i - x_0))(x_i - x_0)^2(x_i - x_1) \end{bmatrix}}{\begin{bmatrix} \sum_{i=2}^{m}(x_i - x_0)^2(x_i - x_1)^2 & \sum_{i=2}^{m}(x_i - x_0)^3(x_i - x_1)^2 \\ \sum_{i=2}^{m}(x_i - x_0)^3(x_i - x_1)^2 & \sum_{i=2}^{m}(x_i - x_0)^4(x_i - x_1)^2 \end{bmatrix}} \qquad (17)$$

The following data is given.

Example 3. Consider the data *x*=[0 9 1 2 3 4 5 6 7 8 10]; *y*=[13 13 -5 -26 -58 -90 -100 -120 -94 -61 129] with reference points of $(x_0, y_0) = (0,13)$ and $(x_1, y_1) = (9,13)$. For the cubic polynomial passing through this reference point, $a_0 = 13$, $a_1 = 0$, and hence

$$p_3(x) = 13 + a_2 x(x - 9) + a_3 x^2(x - 9) \qquad (18)$$

From (16) and (17), $a_2 = 0.7956$ and $a_3 = 1.0437$. The data and its approximation function are contrasted in Figure 3.

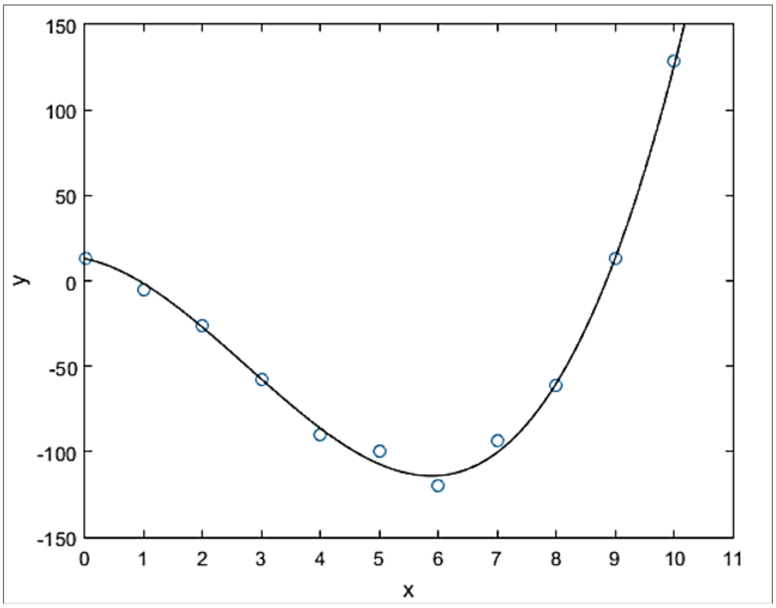

Fig 3: Comparison of the original data (o) with the cubic regression subject to the constraints $(x_0, y_0) = (0,13)$ and $(x_1, y_1) = (9,13)$

From (7), the standard error is $S_{y/x} = 4.9227$.

Instead of the four coefficients in the calculations of the regular regression, calculating only two coefficients reduced the algebra much. This approach may be used if one is confident about the preciseness of the reference points.

5. Concluding Remarks

A restricted form of regression analysis is presented in this work. Reference data points are employed to reduce the number of coefficients to be calculated in the analysis. The theory is given for an arbitrary order of polynomial functions with several constrained data points. The general analysis is then applied to linear, quadratic and cubic polynomials. The algorithm can be summarized below:

1. Select the degree of the polynomial for a given data set.
2. Decide for the reference data points to be used as constraints.
3. Write the polynomial in the special form given in section 2.
4. Calculate the coefficients associated with the reference data points first.
5. Apply the regression analysis to calculate the remaining coefficients.
6. Plot the data and the polynomial in the same figure for visual comparison.

This approach reduces the algebra involved in the regression analysis. One should be cautious in determining the zero error data points (reference points) and should not use too many data points not to end up with an imprecise approximation. Determining the precise degree of a polynomial to be used was already addressed in a previous work [5].

Declarations

- **Availability of Supporting Data-** There is no associated additional data with the study.
- **Competing Interests-** Author declares no competing interests.
- **Funding-** No funding has been received for the study.
- **ORCID ID-** Mehmet Pakdemirli: 0000-0003-1221-9387

6. References

1. Ben A. Bivariate analysis techniques: correlation, regression and their applications. Math Eterna. 2024;14(2).
2. Holman JP. Experimental methods for engineers. Tokyo: McGraw-Hill; 1978.
3. Motulsky HJ, Ransnas LA. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. FASEB J. 1987;1:365-374.
4. Anderson TW. The choice of the degree of a polynomial regression as a multiple decision problem. Ann Math Stat. 1962;33:255-265.
5. Pakdemirli M. A new perturbation approach to optimal polynomial regression. Math Comput Appl. 2016;21(1):1-8.
6. Nayfeh AH. Introduction to perturbation techniques. New York: John Wiley and Sons; 1981.
7. Pakdemirli M. Strategies for treating equations with multiple perturbation parameters. Math Eng Sci Aerosp (MESA). 2023;14(4):1-18.
8. Pakdemirli M. Optimal multiple functional regression analysis using perturbation theory. J Taibah Univ Sci. 2024;18(1):2366522.
9. Pakdemirli M, Sarı G. A comprehensive perturbation theorem for estimating magnitudes of roots of polynomials. LMS J Comput Math. 2013;16:1-8.
10. Dette H. Optimal designs for identifying the degree of a polynomial regression. Ann Stat. 1995;23:1248-1266.
11. Dette H, Studden WJ. Optimal designs for polynomial regression when the degree is not known. Stat Sin. 1995;5:459-473.
12. Jayasuriya BR. Testing for polynomial regression using nonparametric regression techniques. J Am Stat Assoc. 1996;91:1626-1631.
13. Tomašević N, Tomašević M, Stanivuk T. Regression analysis and approximation by means of Chebyshev polynomial. Informatologia. 2009;42:166-172.
14. Ivanov A, Ailuro S. TMPNN: High-order polynomial regression based on Taylor map factorization. In: Proc. 38th AAAI Conf Artif Intell; 2024. p. 12726-12734.
15. Karnehm D, Neve A. A resource-constrained polynomial regression approach for voltage measurement compression in electric vehicle battery packs. Batteries. 2024;10(9):305.
16. Haider C, de França FO, Burlacu B, Kronberger G. Using shape constraints for improving symbolic regression models. arXiv Prepr arXiv:2107.09458. 2021.
17. Chapra SC, Canale RP. Numerical methods for engineers. New York: McGraw Hill; 2014.