Article

A New Perturbation Approach to Optimal Polynomial Regression

Mehmet Pakdemirli

Applied Mathematics and Computation Center, Celal Bayar University, Muradiye, Manisa, 45140, Turkey; mpak@cbu.edu.tr

Abstract: A new approach to polynomial regression is presented using the concepts of orders of magnitudes of perturbations. The data set is normalized with the maximum values of the data first. The polynomial regression of arbitrary order is then applied to the normalized data. Theorems for special properties of the regression coefficients as well as some criteria for determining the optimum degrees of the regression polynomials are posed and proven. The new approach is numerically tested, and the criteria for determining the best degree of the polynomial for regression are discussed.

Keywords: polynomial regression; perturbation analysis; degree of a polynomial

1. Introduction

In experimentation, one usually obtains a data set in the form of (x_i, y_i) $i = 1, 2, \dots N$. The continuous functional relation of the dependent variable y_i on the independent variable x_i is of practical importance, since data are needed for the missing intervals of the variables. Regression analysis is one of the most widely used methods to determine such a relationship, the most common of which being the linear regression. Depending on the nature of the data, nonlinear regression might be inevitable since the data may not always be perfectly approximated by a straight line. Nonlinear regression may be performed with usage of any simplified functional relationship as well as a polynomial with degree n, $n \geqslant 2$.

In this study, the polynomial regression problem is handled with a slightly different approach than the existing analysis in the literature. First, the given data are normalized by the division of the maximum values of each variable. The polynomial regression is then applied to this normalized data. Note that this normalization process does not alter the form of the polynomial representation; rather, it has direct effects on the magnitudes of the coefficients only. By this process, the regression coefficients acquire some important properties that are outlined in three theorems. The theorems employ the concept of the order of magnitudes of the perturbation theory and are about the properties of the regression coefficients. The additional two theorems are about the error introduced in the regression analysis. The theory developed is applied next to several data sets to outline the algorithm and determine the optimal value of the degree n in a polynomial regression.

A good review of the nonlinear regression with concepts rather than the mathematical formulations is presented by Motulsky and Ransnas [1]. The Anderson's procedure [2] for determining the optimum degree in a polynomial regression is to start from a prescribed certain n value and the procedure is to test in sequence whether the coefficients are 0. If the coefficient of the highest degree is zero, one lower degree polynomial is taken until a non-zero coefficient for the highest degree polynomial term is obtained. In the present analysis, the requirement of a zero highest degree coefficient is somewhat relaxed for the normalized data; if the coefficient of the highest degree is small, *i.e.*, of O(ε), $\varepsilon \ll 1$, then a lower degree polynomial can be selected. For some theoretical results on the topic and optimal designs for the Anderson's procedure, see Dette [3] and Dette and Studden [4]. Nonparametric

regression techniques were utilized to test the validity of a *k*th-order polynomial regression model by Jayasuri [5]. Chebyshev polynomials have been used to approximate the polynomial obtained by the regression analysis to a lower degree polynomial with required accuracy, keeping the polynomial coefficient within the estimated intervals at the same time by Tomašević *et al.* [6].

2. Regression Analysis

Assume that a set of data is given or obtained from experimentation (x_i, y_i) $i = 1, 2, \ldots N$ and that the data can be approximated by an *n*th degree polynomial:

$$y^{(n)} = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 \tag{1}$$

The total difference with squaring all the differences between the data and the approximation is

$$e = \sum_{i=1}^{N} \left(y_i - y^{(n)}(x_i) \right)^2 \tag{2}$$

The aim is to find the best coefficients that make the above total differences minimum; hence,

$$\frac{\partial e}{\partial a_m} = 0, \quad m = 0, 1, 2, \ldots n \tag{3}$$

which is called the method of least squares [7]. The above set of equations lead to a matrix equation

$$\mathbf{X}(n)\mathbf{A} = \mathbf{Y} \tag{4}$$

where

$$\mathbf{X}(n) = \begin{bmatrix} N & \sum x_i & \ldots & \sum x_i^n \\ \sum x_i & \sum x_i^2 & \ldots & \sum x_i^{n+1} \\ \vdots & \vdots & & \vdots \\ \sum x_i^n & \sum x_i^{n+1} & \ldots & \sum x_i^{2n} \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \vdots \\ \sum y_i x_i^n \end{bmatrix} \tag{5}$$

The optimum coefficients are then

$$\mathbf{A} = \mathbf{X}^{-1}(n)\mathbf{Y} \tag{6}$$

The standard regression error is defined to be [7]

$$S_{y/x} = \left[\frac{1}{N - (n+1)} \sum_{i=1}^{N} \left(y_i - y^{(n)}(x_i) \right)^2 \right]^{1/2} \tag{7}$$

where $N - (n+1)$ is called the degree of freedom.

3. New Theorems

In perturbation theory, one deals with dimensionless quantities rather than the dimensional ones to enable healthy comparisons of magnitudes of terms. The data set obtained from experimentation may have any unit, and the simplest way to obtain dimensionless quantities is to divide the data by the maximum value in the set

$$\overline{x}_i = \frac{x_i}{x_{\max}}, \quad \overline{y}_i = \frac{y_i}{y_{\max}} \tag{8}$$

where $\overline{x}_i$, $\overline{y}_i$ are called the normalized data. If one assumes that the measured quantities are all positive, the normalized data set is confined in a square region in the first quadrant $(0 \leqslant \overline{x}_i \leqslant 1, \quad 0 \leqslant \overline{y}_i \leqslant 1)$. The regression analysis will be carried out for the normalized data. The normalization is crucial and many interesting features about the coefficients can be derived with this normalization.

Theorem 1

For the polynomial regression of degree n for the normalized data,

$$\overline{y}^{(n)} = a_n \overline{x}^n + a_{n-1} \overline{x}^{n-1} + ... + a_1 \overline{x} + a_0 \tag{9}$$

if all $a_i \geqslant 0$, then $a_i \sim O(\varepsilon^{k_i}), k_i \geqslant 0 \, (\varepsilon << 1)$.

Proof

The theorem states that there cannot be a large coefficient with an order of magnitude greater than 1 if all coefficients turn out to be positive. Since the data are confined in a square, for $\overline{x} = 1, \quad \overline{y}^{(n)} \leqslant O(1)$. Hence, from (9)

$$a_n + a_{n-1} + ... + a_1 + a_0 \leqslant O(1) \tag{10}$$

or replacing the magnitudes of the terms

$$O(\varepsilon^{k_n}) + O(\varepsilon^{k_{n-1}}) + ... + O(\varepsilon^{k_1}) + O(\varepsilon^{k_0}) \leqslant O(1) \tag{11}$$

If at least one of the k_i is less than zero, there is an unbalanced large term of $O(\frac{1}{\varepsilon^{-k_i}}), -k_i > 0$, which spoils the inequality. Hence, if all coefficients are positive, the coefficients can be at most $O(1)$ □

Theorem 2

For the polynomial regression of degree n for the normalized data

$$\overline{y}^{(n)} = a_n \overline{x}^n + a_{n-1} \overline{x}^{n-1} + ... + a_1 \overline{x} + a_0 \tag{12}$$

if there are large coefficients, *i.e.*, $a_i \sim O(\varepsilon^{k_i}), k_i < 0 \, (\varepsilon << 1)$, then the coefficients cannot have the same signs, and other large term(s) $a_m \sim O(\varepsilon^{k_m}), k_m < 0$ should appear with opposite signs.

Proof

For $\overline{x} = 1, \quad \overline{y}^{(n)} \leqslant O(1)$. Hence, from (12)

$$O(\varepsilon^{k_n}) + O(\varepsilon^{k_{n-1}}) + ... + O(\varepsilon^{k_1}) + O(\varepsilon^{k_0}) \leqslant O(1) \tag{13}$$

If there is a large term with k_i less than zero, this term must be balanced by other large term(s) $a_m \sim O(\varepsilon^{k_m}), k_m < 0$ with opposite signs so that the sum adds to at most an $O(1)$ term □

Theorem 3

For the polynomial regression of degree n for the normalized data,

$$\overline{y}^{(n)} = a_n \overline{x}^n + a_{n-1} \overline{x}^{n-1} + ... + a_1 \overline{x} + a_0 \tag{14}$$

the sum of the regression coefficients are bounded such that

$$\sum_{i=1}^{n} a_i \leqslant O(1)$$

with the equality sign holding for the specific set of data where $x_N = x_{\max}, y_N = y_{\max}$.

Proof

For $\bar{x} = 1$, due to normalization, $\bar{y} \leqslant O(1)$. Using the normalized regression equation and substituting $\bar{x} = 1$,

$$\bar{y}(1) = \sum_{i=1}^{n} a_i \leqslant O(1) \tag{15}$$

If y_{max} corresponds to x_{max}, then $\bar{y}(1) \cong 1$ due to the approximation of the regression; hence, the equality sign holds for (15) □

The previous three theorems are all about features of the regression coefficients. They can be used to check the preciseness of the determination of the coefficients. The following theorem is about the standard regression error for the normalized data.

Theorem 4

For a good polynomial regression of the normalized data, the standard regression error is bounded by

$$S_{\bar{y}/\bar{x}} \leqslant O(\varepsilon) \quad (\varepsilon << 1) \tag{16}$$

Proof

If the representation is good, the curve passes close enough to each data point; since the data are within a square box of length 1, the distance between each datum and the curve can only be a small fraction of 1. Hence,

$$\left| \bar{y}_i - \bar{y}^{(n)}(\bar{x}_i) \right| \leqslant O(\varepsilon) \quad (\varepsilon << 1) \tag{17}$$

Substituting the above into (7) yields,

$$S_{\bar{y}/\bar{x}} \leqslant \left[\frac{N}{N - (n+1)} O(\varepsilon^2) \right]^{1/2} \tag{18}$$

For a good representation, the number of data points N should be much larger than the degrees of freedom $n + 1$; hence, $\dfrac{N}{N - (n+1)} \sim O(1)$. Substituting the order of magnitude into (18) yields, $S_{\bar{y}/\bar{x}} \leqslant O(\varepsilon)$ □

Theorem 4 can be used roughly to check the appropriateness of the polynomial regression. As $\varepsilon \to 0$, the fit becomes better.

The last theorem gives a criterion for using a polynomial with degree n instead of $n + 1$.

Theorem 5

Assume that the normalized data are expressed with a polynomial regression of degree n:

$$\bar{y}^{(n)}(\bar{x}) = b_n \bar{x}^n + b_{n-1} \bar{x}^{n-1} + \ldots + b_1 \bar{x} + b_0 \tag{19}$$

with $b_n \sim O(1)$. For the polynomial regression of degree $n + 1$ of the same data,

$$\bar{y}^{(n)}(\bar{x}) = a_{n+1} \bar{x}^{n+1} + a_n \bar{x}^n + \ldots + a_1 \bar{x} + a_0 \tag{20}$$

if $a_{n+1} \sim O(\varepsilon)$, $(\varepsilon << 1)$, then the average approximation error for using the degree n polynomial instead of the degree $n+1$ polynomial is bounded by

$$e(n, n+1) = \frac{1}{N} \sum_{i=1}^{N} \left| \bar{y}^{(n+1)}(\bar{x}_i) - \bar{y}^{(n)}(\bar{x}_i) \right| \leqslant O(\varepsilon) \tag{21}$$

Proof

Since $\bar{x}_i \sim O(1)$, due to normalization, $\bar{y}^{(n)}(\bar{x}_i) \cong \sum_{i=1}^{n} b_i \sim O(1)$ via Theorem 3. $\bar{y}^{(n+1)}(\bar{x}_i) \cong a_{n+1} + \sum_{i=1}^{n} a_i \sim O(1)$ also via Theorem 3. Since $a_{n+1} \sim O(\varepsilon)$, $\sum_{i=1}^{n} a_i \sim O(1)$. Replacing the order of magnitudes to (21), $e(n, n+1) \leqslant \frac{1}{N} N\, O(\varepsilon) \leqslant O(\varepsilon)$ □

Theorem 5 leads to a useful criterion for determining the degree of the regression polynomial. If the leading coefficient in a regression polynomial of degree n comes out to be of order 1 and by increasing the degree to $n + 1$, if the leading coefficient becomes much smaller than 1, there is no point in using a more complicated regression formula. Hence, one may stop at degree n.

4. Numerical Applications

Applications of the theory are depicted by numerical simulations. In this work, the study is confined to a polynomial type regression analysis. Usually, the data obtained from an experiment do not give precise information about the degree of the polynomial or whether a polynomial representation is suitable at all or not. To outline our algorithm, carefully selected data of polynomial and functional types are examined. The data are somewhat distorted so that an exact match between the data points and regression curves are not possible in accordance with the real cases. Applications of the theorems are indicated on each data set.

In Table 1, an artificial data set distorted from a linear relationship is chosen. n represents the degree of the polynomial. For each data set, the determinant of $\mathbf{X}(n)$ (see Equation (5)), the regression coefficients, the sum of the coefficients, the standard regression errors, and the difference of the $n-1$th degree regression error and nth degree regression error are given. For $n = 1$, since the data are suitable for a linear relationship, all indicators are good, *i.e.* the determinant is not too singular, the coefficients are all positive and not larger than $O(1)$ (Theorem 1), the highest order coefficient is of order 1 (Theorem 5), the sum of the coefficients is close to 1 (Theorem 3), and the standard error is small (Theorem 4). When one tries to fit a parabolic curve to the same data set, the coefficient of the highest order term becomes a small term ($a_2 = -0.0628$), an obvious indication that the data set does not need a polynomial of degree 2 (Theorem 5). The standard error increased slightly, which is another indicator that the parabolic curve is unnecessary. The difference of the standard errors between the two steps is negative, which indicates that one should stop at this stage and use the regression analysis of the previous step ($n = 1$).

Table 1. Polynomial regression for normalized data close to a linear function.

Data 1	x = [0 1 2 3 4 5 6 7] y = [1 2.2 2.9 4 5.2 6.1 6.8 7.9]		Data close to $y = x + 1$		
n	$\det(\mathbf{X}(n))$	$a_o \ldots \ldots \ldots .a_n$	$\Sigma\, a_i$	$(S_{y/x})_n$	$\Delta(S_{y/x})_n = (S_{y/x})_{n-1} - (S_{y/x})_n$
1	6.8571	0.1382 0.8660	1.0042	0.0184	
2	0.4798	0.1292 0.9288 −0.0628	0.9953	0.0188	−0.0004

Table 2 is produced from a distorted set of parabolic data. When $n = 2$, the coefficient of highest order term is not small, the sum of the coefficients are approximately 1, and the standard error is reduced to a much smaller number compared to the previous case of $n=1$ (Theorem 4). However, when a cubic polynomial is fit to the data, one sees that the coefficient of the highest order term is small compared to 1 ($a_3 = 0.1414$) (Theorem 5), the sum of the coefficients start deviating from 1 (Theorem 3), the standard error increased compared to the previous step, the difference of the standard errors becomes negative, and the determinant is much more singular, all indicating that $n = 3$ is not a suitable

choice and one should stop at $n = 2$. Note that, if the data were exactly parabolic, $a_3 = 0$. Hence, the criterion should not be to require the highest degree coefficient to be zero; instead, if it appears to be small, there is no point in using a more complex representation. Higher representations may lower the standard error but may lead to a phenomenon called polynomial wiggling, which casts doubts about the preciseness of the missing intermediate data.

Table 2. Polynomial regression for normalized data close to a parabolic function.

Data 2	$x = [0\ 1\ 2\ 3\ 4\ 5\ 6\ 7]$ $y = [0\ 1\ 5\ 9\ 14\ 25\ 37\ 49]$		Data close to $y = x^2$		
n	$\det(\mathbf{X}(n))$	$a_o \ldots \ldots .a_n$	$\Sigma\ a_i$	$(S_{y/x})_n$	$\Delta(S_{y/x})_n = (S_{y/x})_{n-1} - (S_{y/x})_n$
1	6.8571	−0.1429 1.0000	0.8571	0.1148	
2	0.4798	0.0068 −0.0476 1.0476	1.0068	0.0216	0.0932
3	0.0024	0.0025 0.0317 0.8355 0.1414	1.0111	0.0237	−0.0020

Table 3 is produced from a distorted data of a cubic polynomial. At $n = 4$, the determinant becomes very singular, the highest degree coefficient appears to be small ($a_4 = 0.1626$) (Theorem 5), the sum of the coefficients start deviating from 1 (Theorem 3), the standard regression error is higher than the previous case, and the difference turns out to be negative, all indicating that $n = 3$ is the best choice.

Table 3. Polynomial regression for normalized data close to a cubic function.

Data 3	$x = [0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8]$ $y = [1\ 1.1\ 4.9\ 20\ 51\ 98\ 175\ 300\ 455]$		Data close to $y = x^3 - x^2 + 1$		
n	$\det(\mathbf{X}(n))$	$a_o \ldots \ldots .a_n$	$\Sigma\ a_i$	$(S_{y/x})_n$	$\Delta(S_{y/x})_n = (S_{y/x})_{n-1} - (S_{y/x})_n$
1	8.4375	−0.1887 0.9175	0.7288	0.1682	
2	0.6345	0.0425 −0.6673 1.5848	0.9599	0.0390	0.1292
3	0.0035	0.0005 0.0562 −0.3339 1.2791	1.0019	0.0070	0.0310
4	1.2100e−06	0.0014 0.0166 −0.1317 0.9539 0.1626	1.0029	0.0077	−0.0007

In Table 4, functional-type data are considered. The data are a distorted form of a logarithmic relationship. Given the previous criterion, the ideal representation of the data is a cubic polynomial because the determinant is too much singular for $n = 4$, the sum of the coefficients start deviating from 1 and the difference of the standard errors become negative. Note that, since the original data are not of a polynomial form, the highest order coefficient at $n = 4$ is not small, but, at this stage, one has larger opposite sign coefficients (Theorem 2), which is an indicator that one should stop and take the n value of the previous stage.

Table 4. Polynomial regression for normalized data close to a logarithmic function.

Data 4	x = [0 1 2 3 4 5 6 7] y = [0 0.7 1.2 1.3 1.6 1.8 1.9 2.2]			Data close to $y = ln(1 + x)$	
n	det($\mathbf{X}$(n))	$a_o \ldots \ldots .a_n$	$\Sigma\, a_i$	$(S_{y/x})n$	$\Delta(S_{y/x})n = (S_{y/x})_{n-1} - (S_{y/x})n$
1	6.8571	0.1629 0.8902	1.0530	0.0967	
2	0.4798	0.0587 1.6193 −0.7292	0.9489	0.0615	0.0352
3	0.0024	0.0069 2.5694 −3.2686 1.6930	1.0007	0.0333	0.0282
4	7.6071e-07	0.0002 2.8906 −4.9333 4.3800 −1.3435	0.9940	0.0359	−0.0026

Finally, the distorted data of a function with fractional power are considered in Table 5. Starting at $n = 4$, the determinant is very singular, there are large opposite sign coefficients (Theorem 2), the sum starts deviating from 1, and, although the difference of the errors does not turn to be negative, the gain is very marginal. Therefore, $n = 2$ or $n = 3$ may be used to express the given data.

Table 5. Polynomial regression for normalized data close to a fractional power function.

Data 5	x = [0 1 2 3 4 5 6 7] y = [0 1 1.5 1.6 2.1 2.4 2.5 2.8]			Data close to $y = x^{1/2}$	
n	det($\mathbf{X}$(n))	$a_o \ldots \ldots .a_n$	$\Sigma\, a_i$	$(S_{y/x})n$	$\Delta(S_{y/x})n = (S_{y/x})_{n-1} - (S_{y/x})n$
1	6.8571	0.1696 0.9018	1.0714	0.0982	
2	0.4798	0.0640 1.6414 −0.7396	0.9658	0.0627	0.0355
3	0.0024	0.0233 2.3879 −2.7348 1.3302	1.0065	0.0517	0.0110
4	7.6071e-07	0.0058 3.2227 −7.0613 8.3136 −3.4917	0.9890	0.0478	0.0039
5	1.2003e-11	−0.0032 4.5848 −18.6930 41.3990 - 41.4880 15.1990	0.9981	0.0401	0.0077

5. Concluding Remarks

For the case of a normalized data, new theorems are posed and proven for the polynomial regression coefficients and errors. The concepts of orders of magnitudes in the perturbation theory is utilized. These theorems turn out to be useful in determining the optimum value of the highest degree in a polynomial representation. The steps of the algorithm is given below:

(1)　Start from a linear relationship with $n = 1$ and increase n by one at each step;

(2)   Calculate the determinant, coefficients, sum of coefficients, standard regression error, and difference of errors (for $n \geqslant 2$);

(3)   Check the singularity of the determinant, the highest degree coefficient, the remaining coefficients, the sum of the coefficients, the standard regression error, and the difference of errors;

(4)   Stop at n if the highest order coefficient turns out to be small, and/or the difference of errors is negative, and use $n - 1$ as an ideal representation; and

(5)   Stop at n if the determinant is too singular, there are opposite sign large coefficients, the sum of coefficients start deviating from the ideal value, or the standard error is small. One may use $n - 1$ as an ideal representation.

If the data are suitable for a polynomial representation, then indicators at stage 4 are observed and the decision of the optimal n value is clearer. When the functional dependence is not suitable for a polynomial representation, one criterion, or both criteria, of stage 4 need not be observed, and the fainter criterion of stage 5 may be employed.

For the normalized data, this work suggests a simple and alternative way of determining optimal n values in contrast to the standard statistical approaches.

Conflicts of Interest: The author declares no conflict of interest.

References

1.   Motulsky, H.J.; Ransnas, L.A. Fitting curves to data using nonlinear regression: A practical and nonmathematical review. *FASEB J.* **1987**, *1*, 365–374. [PubMed]

2.   Anderson, T.W. The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Stat.* **1962**, *33*, 255–265. [CrossRef]

3.   Dette, H. Optimal designs for identifying the degree of a polynomial regression. *Ann. Statist.* **1995**, *23*, 1248–1266. [CrossRef]

4.   Dette, H.; Studden, W.J. Optimal designs for polynomial regression when the degree is not known. *Stat. Sin.* **1995**, *5*, 459–473.

5.   Jayasuriya, B.R. Testing for polynomial regression using nonparametric regression techniques. *J. Am. Stat. Assoc.* **1996**, *91*, 1626–1631. [CrossRef]

6.   Tomašević, N.; Tomašević, M.; Stanivuk, T. Regression analysis and approximation by means of chebyshev polynomial. *Informatologia* **2009**, *42*, 166–172.

7.   Chapra, S.C.; Canale, R.P. *Numerical Methods for Engineers*; Mc Graw Hill: New York, NY, USA, 2014.