New Criteria for Polynomial Regression

MEHMET PAKDEMİRLİ
Applied Mathematics and Computation Center
Celal Bayar University
Muradiye, Yunusemre, Manisa
TURKEY

Abstract: - The order of a polynomial for approximating a given data is important in a polynomial regression analysis. By normalizing the data and employing the order of magnitudes from the perturbation theory, new theorems are posed and proven. The theorems outline the basic features of the regression coefficients for the normalized data. Using the theorems and the described algorithm, the optimal degree of a polynomial can be determined. This task is a multiple criteria decision task and numerical examples are given to outline the basics of the algorithm.

Key-Words: - Polynomial Regression, Orders of Magnitudes, Optimal Criteria

1 Introduction

Regression analysis is one of the most fundamental methods used to approximate a given data set. The most common one is the linear regression which is well established. If the data is not suitable for a linear relationship, the nonlinear regression analysis is inevitable. Usually a degree n ($n\geq 2$) polynomial or a basic simple function with a few parameters is used to approximate the data. The regression analysis, if done properly, enables to make interpolations and extrapolations of the given data.

As a nonlinear regression function, only polynomial type regressions are considered here. One of the important issues in polynomial regression is to determine the degree of the polynomial. According to the Anderson's procedure [1], one starts with a polynomial of certain n value and the coefficients are calculated for that degree. If the highest degree coefficient is zero, one resorts to a polynomial of degree n-1. The algorithm is terminated for the specific value of degree m for which the highest degree coefficient is nonzero. Here, in the present analysis, the zero highest degree coefficient requirement is somewhat relaxed and if the highest degree coefficient is very small compared to 1 for the normalized data, one may try a lower degree polynomial.

Motulsky and Ransnas [2] presented a good review of the nonlinear regression with mathematical formulations kept to a minimum. Optimal designs for the Anderson's procedure are given by Dette [3] and Dette and Studden [4]. The validity of a k'th order polynomial regression model was tested by utilizing nonparametric regression techniques [5].

In this study, new theorems are given which can be used as simple tests for the appropriateness of the polynomial regressions. A crucial point is to normalize the data with the maximum values in each set. After normalization, the regression coefficients possess some important properties which are exploited through theorems. By employing the theorems, the optimal degree of a polynomial as well as the appropriateness of the data for a polynomial regression can be determined. The usage of the algorithm is outlined through specific examples.

2 Regression Analysis

For a given set of data (x_i, y_i) i=1,2,…N, if the data is approximated by an n'th degree polynomial

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

the optimum coefficients are calculated as [6]

$$\mathbf{A} = \mathbf{X}^{-1}(n)\mathbf{Y} \tag{2}$$

where

$$\mathbf{X}(n) = \begin{bmatrix} N & \sum x_i & .... & \sum x_i^n \\ \sum x_i & \sum x_i^2 & .... & \sum x_i^{n+1} \\ \vdots & \vdots & & \vdots \\ \sum x_i^n & \sum x_i^{n+1} & .... & \sum x_i^{2n} \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \vdots \\ \sum y_i x_i^n \end{bmatrix} \tag{3}$$

The standard regression error is

$$S_{y/x} = \left[\frac{1}{N-(n+1)} \sum_{i=1}^{N} \left(y_i - y^{(n)}(x_i) \right)^2 \right]^{1/2} \tag{4}$$

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

3 Theorems

For the theorems to be applicable, the data is normalized first by dividing by the maximum values in each set

$$\bar{x}_i = \frac{x_i}{x_{\max}}, \quad \bar{y}_i = \frac{y_i}{y_{\max}} \ . \tag{5}$$

For positive quantities, the data is confined in a square region in the first quadrant.

Theorem 1

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

$$\bar{y}^{(n)} = a_n \bar{x}^n + a_{n-1}\bar{x}^{n-1} + ... + a_1 \bar{x} + a_0 \tag{6}$$

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

Proof

If all coefficients are positive, the theorem states that there cannot be a large coefficient. Since the data is confined in a square, for $\bar{x} = 1, \quad \bar{y}^{(n)} \leq O(1)$. Hence from (6)

$$a_n + a_{n-1} + ... + a_1 + a_0 \leq O(1) \tag{7}$$

or replacing the magnitudes

$$O(\varepsilon^{k_n}) + O(\varepsilon^{k_{n-1}}) + ... + O(\varepsilon^{k_1}) + O(\varepsilon^{k_0}) \leq O(1) \tag{8}$$

If at least one of the k_i is less than zero, there is an unbalanced large term 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 given in Eq. (6), if there exists 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

As stated in the previous theorem, coefficients satisfy Eqs. (7) and (8) for the normalized data. If there exists 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 2 is a complimentary theorem of Theorem 1. The next theorem states the additive property of the coefficients.

Theorem 3

For the polynomial regression of degree *n* for the normalized data given in Eq. (6), the sum of the regression coefficients are bounded such that

$$\sum_{i=1}^{n} a_i \leq O(1) \tag{9}$$

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, $\bar{y} \leq O(1)$. Using the normalized regression equation and substituting $\bar{x} = 1$

$$\bar{y}(1) = \sum_{i=1}^{n} a_i \leq O(1) \tag{10}$$

If y_{max} corresponds to x_{max}, then $\bar{y}(1) \cong 1$ due to the approximation of the regression and hence the equality sign holds for (10) □

In the previous three theorems, the properties of the regression coefficients for the normalized data are outlined. They can be used to check the calculations. The last theorem sets up a criterion for the standard regression error.

Theorem 4

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

$$S_{\bar{y}/\bar{x}} \leq O(\varepsilon) \quad (\varepsilon << 1) \tag{11}$$

Proof

If there is a perfect representation, the curve passes close enough to each data point and since the data is within a square box of length 1, the distance between each real and approximated data can only be a small fraction of 1, hence

$$\left| \bar{y}_i - \bar{y}^{(n)}(\bar{x}_i) \right| \leq O(\varepsilon) \quad (\varepsilon << 1) \tag{12}$$

Substituting the above into (4) yields

$$S_{\bar{y}/\bar{x}} \leq \left[\frac{N}{N-(n+1)} O(\varepsilon^2) \right]^{1/2} \tag{13}$$

For a good representation, the number of data points *N* should be much larger than the degrees of freedom *n*+1 and hence $\frac{N}{N-(n+1)} \sim O(1)$. Substituting the order of magnitude into (13) yields $S_{\bar{y}/\bar{x}} \leq O(\varepsilon)$ □

Theorem 4 can be used as a rough criterion to determine the suitability of the polynomial regression.

4 Applications

For a given experimental data, usually the distribution of data does not precisely give a clue about the degree of the polynomial to be used. Sometimes, a polynomial regression is not suitable at all. With worked examples and the theorems given, guidelines for selection of the degree of a polynomial regression are established in this section.

Table 1 is produced from an approximately cubic polynomial data. The major indicators are the determinant, the coefficients, the sum of the coefficients, the regression error and the difference of the regression errors between the n-1'th degree and n'th degree. At n=4, the determinant becomes very singular, the highest degree coefficient appears to be small (a_4=0.1626), 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.

In Table 2, a functional type of data is considered. The data is an approximation 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 is 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.

5 Concluding Remarks

The basics of the algorithm can be summarized as follows.

1) Try first a linear relationship and increase the degree by one at each stage.
2) Form a similar table as given in Tables 1 and 2.
3) Check the singularity of the determinant, the highest degree coefficient, the magnitudes of the coefficients, the sum of the coefficients, the standard regression error and the differences of errors at each step.
4) Stop at degree n when the highest degree coefficient is small, and/or the difference of the errors is negative and use n-1 as the ideal degree.
5) Although not compulsory, may stop at n when the determinant is too singular, there are opposite sign large coefficients, the sum of the coefficients start deviating from the ideal value, the standard error is small and may use n-1 as an ideal representation.

Acknowledgment: - The support of the Turkish Academy of Sciences (TÜBA) for the expenses of the conference is highly appreciated.

References:

[1] T. W. Anderson, The choice of the degree of a polynomial regression as a multiple decision problem, *The Annals of Mathematical Statistics*, Vol.33, 1962, pp. 255-265.
[2] H. J. Motulsky and L. A. Ransnas, Fitting Curves to Data Using Nonlinear regression: A Practical and Nonmathematical Review, *FASEB Journal*, Vol.1, 1987, pp. 365-374.
[3] H. Dette, Optimal designs for identifying the degree of a polynomial regression, *The Annals of Statistics*, Vol.23, 1995, 1248-1266.
[4] H. Dette and W. J. Studden, Optimal designs for polynomial regression when the degree is not known, *Statistica Sinica*, Vol.5, 1995, 459-473.
[5] B. R. Jayasuriya, Testing for polynomial regression using nonparametric regression techniques, *Journal of the American Statistical Association*, Vol.91, 1996, 1626-1631.
[6] S. C. Chapra and R. P. Canale, *Numerical Methods for Engineers*, Mc Graw Hill, 2014.

Table 1- Polynomial Regression for the Normalized Data Close to a Cubic Function

Data 1	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(**X**(n))	a_o…….a_n	Σa_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

Table 2- Polynomial Regression for the Normalized Data Close to a Logarithmic Function

Data 2	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(**X**(n))	a_o…….a_n	Σa_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