Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 48 (2012) 500 – 506

Procedia Engineering

www.elsevier.com/locate/procedia

MMaMS 2012

Modelling using polynomial regression

Eva Ostertagová[a]*

[a]Technical University of Košice, Faculty of Electrical Engineering and Informatics, Department of Mathematics and Theoretical Informatics, Němcovej 32, 042 00 Košice, Slovak Republic

Abstract

This paper is concentrated on the polynomial regression model, which is useful when there is reason to believe that relationship between two variables is curvilinear. The polynomial regression model has been applied using the characterisation of the relationship between strains and drilling depth. Parameters of the model were estimated using a least square method. After fitting, the model was evaluated using some of the common indicators used to evaluate accuracy of regression model. The data were analyzed using computer program MATLAB that performs these calculations.

Keywords: multiple regression model, mean absolute percentage error, root mean squared error, *R*-squared, adjusted *R*-squared.

Nomenclature

e	error vector
h	hole depth (mm)
MAPE	mean absolute percentage error
RMSE	root mean squared error
R^2	*R*-squared
R^{*2}	adjusted *R*-squared
X	design matrix
Y	response vector
Greek symbols	
β	vector of regression parameters
ε	strain (μm/m)

1. Introduction

Regression analysis involves identifying the relationship between a dependent variable and one or more independent variables. It is one of the most important statistical tools which is extensively used in almost all sciences. It is specially used in business and economics to study the relationship between two or more variables that are related causally. A model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation.

* Corresponding author. Tel.: +421-55-602-2447
E-mail address: eva.ostertagova@tuke.sk.

doi:10.1016/j.proeng.2012.09.545

Various tests are then employed to determine if the model is satisfactory. Model validation is an important step in the modelling process and helps in assessing the reliability of models before they can be used in decision making.

2. The multiple regression

Multiple regression refers to regression applications in which there are more than one independent variables. Multiple regression includes a technique called polynomial regression. In polynomial regression we regress a dependent variable on powers of the independent variables.

2.1. The multiple regression model

The basic multiple regression model of a dependent (response) variable Y on a set of k independent (predictor) variables $X_1, X_2, \dots, X_k$ can be expressed as

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + e_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + e_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + e_n \end{cases} \tag{1}$$

i.e.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \text{, for } i = 1, 2, \dots, n \tag{2}$$

where y_i is the value of the dependent variable Y for the ith case, x_{ij} is the value of the jth independent variable X_j for the ith case, β_0 is the Y-intercept of the regression surface (think multidimensionality), each β_j, $j = 1, 2, \dots, k$, is the slope of the regression surface with respect to variable X_j and e_i is the random error component for the ith case. In basic equations (1) we have n observations and k predictors ($n > k + 1$).

The assumptions of the multiple regression model are similar to those for the simple linear regression model. Model assumptions [1]:

- For each observation the errors e_i are normally distributed with mean zero and standard deviation σ and are independent of the error terms associated with all other observations. The errors are uncorrelated with each other. That is $e_i \sim N(0, \sigma^2)$ for all $i = 1, 2, \dots, n$, independent of other errors.
- In the context of regression analysis, the variables X_j are considered fixed quantities, although in the context of correlation analysis, they are random variables. In any case, X_j are independent of the error term. When we assume that X_j are fixed quantities, we are assuming that we have realizations of k variables X_j and that the only randomness in Y comes from the error term.

In matrix notation, we can rewrite model (1) as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{3}$$

where response vector $\mathbf{Y}$ and error vector $\mathbf{e}$ are column vectors of length n, vector of parameters $\boldsymbol{\beta}$ is column vector of length $k+1$ and design matrix $\mathbf{X}$ is n by $k+1$ matrix (with its first column having all elements equal to 1, the second column being filled by the observed values of X_1, etc.). We want to estimate unknown values of $\boldsymbol{\beta}$ and $\mathbf{e}$.

2.2. Least squared error approach in matrix form

We estimate the regression parameters by the method of least squares. This is an extension of the procedure used in simple linear regression. First, we calculate the sum of the squared errors and, second, find a set of estimators that minimize the sum.

Using equation (3) we obtain for the errors

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\,. \tag{4}$$

Find estimator $\hat{\boldsymbol{\beta}}$ we want to minimize the sum of squares of the errors

$$\mathbf{e}^T\mathbf{e} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{5}$$

where the symbol $()^T$ denotes the transpose of the matrix.

Here $\mathbf{e}^T\mathbf{e}$ is scalar. We can take the first derivate of this object function with respect to the vector $\boldsymbol{\beta}$. Making these equal to $\mathbf{0}$ (a vector of zeros) we obtain normal equations

$$\mathbf{X}^T\mathbf{X}\,\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}\,. \tag{6}$$

Multiply the inverse matrix of $(\mathbf{X}^T\mathbf{X})^{-1}$ on the both left sides in equation (6), and we have the least squared estimator for the multiple regression model in matrix form [8]

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\,. \tag{7}$$

Vector $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. The fitted (predicted) values for the mean of $\mathbf{Y}$ (let us call them $\hat{\mathbf{Y}}$), are computed by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\,\mathbf{Y} \tag{8}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. We call this the hat matrix because is turns $\mathbf{Y}$ into $\hat{\mathbf{Y}}$. Matrix $\mathbf{H}$ is symmetric, i.e. $\mathbf{H} = \mathbf{H}^T$ and idempotent, i.e. $\mathbf{H}^2 = \mathbf{H}$.

The fitted values for error terms e_i are residuals $\hat{e}_i$, $i = 1, 2, \ldots, n$, that are computed by

$$\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \tag{9}$$

where $\mathbf{I}$ is an identity matrix.

The sum of squares of the residuals $SSE = \hat{\mathbf{e}}^T\hat{\mathbf{e}}$ has the χ^2 distribution with $df_E = n - (k+1)$ degrees of freedom, and is independent of $\hat{\boldsymbol{\beta}}$.

2.3. Polynomial regression model and evaluating of its accuracy

Polynomial regression is a special case of multiple regression, with only one independent variable *X*. One-variable polynomial regression model can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_k x_i^k + e_i\,, \text{ for } i = 1, 2, \ldots, n \tag{10}$$

where *k* is the degree of the polynomial. The degree of the polynomial is the order of the model.

Effectively, this is the same as having a multiple model with $X_1 = X$, $X_2 = X^2$, $X_3 = X^3$, etc.

The mean squared error *MSE* is an unbiased estimator of the variance σ^2 of the random error term and is defined in equation

$$MSE = \frac{SSE}{df_E} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - (k+1)} \tag{11}$$

where y_i are observed values and $\hat{y}_i$ are the fitted values of the dependent variable *Y* for the *i*th case. Since the mean squared error is the average squared error, where averaging is done by dividing by the degrees of freedom, *MSE* is a measure of how well the regression fits the data. The square root of *MSE* is an estimator of the standard deviation σ of the random error term. The root mean squared error $RMSE = \sqrt{MSE}$ is not an unbiased estimator of σ, but it is still a good estimator. *MSE* and *RMSE* are measures of the size of the errors in regression and do not give a indication about the explained component of the regression fit [1].

Mean absolute percentage error *MAPE* is the most useful measure to compare the accuracy of forecasts between different items or products since it measures relative performance [5]. It is one measure of accuracy commonly used in quantitative methods of forecasting. This measure is defined in equation

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{12}$$

If *MAPE* calculated value is less than 10 %, it is interpreted as excellent accurate forecasting, between 10 – 20 % good forecasting, between 20 – 50 % acceptable forecasting and over 50 % inaccurate forecasting [4].

The *R*-squared R^2 (coefficient of determination) of the multiple regression is similar to the simple regression where the coefficient of determination R^2 is defined as

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{13}$$

where *SST* is the total sum of squares and $\overline{y}$ is the arithmetic mean of the *Y* variable. R^2 measures the percentage of variation in the response variable *Y* explained by the explanatory variable *X*. Thus, it is an important measure of how well the regression model fits the data. The value of R^2 is always between zero and one, $0 \le R^2 \le 1$. An R^2 value of 0.9 or above is very good, a value above 0.8 is good, and a value of 0.6 or above may be satisfactory in some applications, although we must be aware of the fact that, in such cases, errors in prediction may be relatively high. When the R^2 value is 0.5 or below, the regression explains only 50 % or less of the variation in the data; therefore, prediction may be poor [1, 9].

Adjusted *R*-squared R^{*2} is computed by

$$R^{*2} = R^2 - \frac{(1-R^2)\,k}{n-(k+1)}. \tag{14}$$

Formula (14) shows explicitly the „adjustment" process, and also demonstrates that the adjusted *R*-squared is always smaller as *R*-squared. R^{*2} is adjusted for the number of variables included in the regression equation. If the value of R^{*2} is much lower than R^2 value, it is an indication our regression equation may be over-fitted to the sample, and of limited generalization. R^{*2} is always preferred to R^2 when data are being examined because of the need to protect against spurious relationships [1, 9].

3. Application of polynomial regression model

The principle of the hole-drilling method lies in determination of stress state alteration which occurs when drilling a hole into the structural element in which residual stresses are found. Detailed description of the procedure can be found in various writings devoted to this method [2, 7, 11, 12]. The hole-drilling method was applied for determination of residual stresses in case of the transverse beam of the casting ladle supporting structure directly in a metallurgical plant. Strain gauge rosette applied to the structural element revealed strain values ε_a, ε_b, ε_c in particular directions marked as *a*, *b*, *c*. Stress state alteration was identified after the hole of 0.5 mm was drilled into the surface of the structural element and was registered even in the depth (drilling stage) of 5 mm. The strain values measured in particular drilling stages (hole depths) are listed in Table 1 [3].

Table 1. Measured strain values in particular drilling stages

drilling stage *h* (mm)	strain values in particular directions (μm/m)		
	ε_a	ε_b	ε_c
0.50	−16.00	−9.00	−6.00
1.00	−38.00	−22.00	−8.00
1.50	−50.00	−32.00	−5.00
2.00	−65.00	−48.00	−2.00
2.50	−72.00	−55.00	2.00

3.00	−80.00	−63.00	4.00
3.50	−85.00	−67.00	5.00
4.00	−89.00	−81.00	6.00
4.50	−93.00	−83.00	8.00
5.00	−100.00	−85.00	9.00

The purpose of this study was to determine the relationship between strains ε_a, ε_b, ε_c in particular directions marked as a, b, c and hole depth h. All analyses were done using MATLAB and with its Curve Fitting Toolbox too [10].

It is recommended that data analysts should endeavour to always plot a simple scatter diagram before using any regression model in order to know the type of relationship that exists between the variable of interest. Figures 1(a), 2(a), 3(a) show the comparison of polynomial regression models with measured data in particular directions marked as a, b, c. Looking at this data we may suspect a simple linear model may not be the best choice here. So, instead of simple linear regression here it makes sense to consider polynomial regression with degree of the polynomial $k > 1$.

Thus, when applied polynomial regression in this example, we fit a linear, quadratic, cubic, maybe a quartic polynomial, and then see if can reduce the model by a few terms. In this case, the polynomial may provide a good approximation of the relationship.

The basic statistical outputs for particular directions a, b, c are, respectively, shown in Tables 2 – 4.

Table 2. Polynomial regression results for direction a

	polynomial model		
	linear	quadratic	cubic
RMSE	7.876	3.011	1.295
MAPE	14.9473	4.8526	1.5763
R^2	0.9233	0.9902	0.9984
R^{*2}	0.9137	0.9874	0.9977

Table 3. Polynomial regression results for direction b

	polynomial model		
	linear	quadratic	cubic
RMSE	5.357	2.542	2.732
MAPE	13.5912	3.0394	2.6997
R^2	0.9638	0.9929	0.9929
R^{*2}	0.9593	0.9908	0.9894

Table 4. Polynomial regression results for direction c

	polynomial model			
	linear	quadratic	cubic	quartic
RMSE	1.501	1.516	1.319	0.656
MAPE	26.2045	24.0227	19.7495	8.1552
R^2	0.9467	0.9524	0.9691	0.9936
R^{*2}	0.94	0.9388	0.9537	0.9885

- Direction a:
  The cubic polynomial regression model outperforms the other two models with lowest error statistics and highest deterministic coefficient.
  Least squares parameter estimates for this model are $\hat{\boldsymbol{\beta}} = (9.2000, -56.9503, 12.3007, -1.0521)^T$.

- Direction b:
  We find that the quadratic polynomial regression model appears to fit the data best.
  Least squares parameter estimates for this model are $\hat{\boldsymbol{\beta}} = (5.8667, -30.2242, 2.3636)^T$.

- Direction c:
  The quartic polynomial regression model is here the best.
  Least squares parameter estimates for this model are $\hat{\boldsymbol{\beta}} = (0.5000, -20.9751, 17.0268, -4.2906, 0.3590)^T$.

There are several possible uses of a regression model. One is understand the relationship between the two or more variables. A more common use of a regression analysis is prediction, providing estimates of values of the dependent

variable (variables) by using the prediction equation. Point predictions are not perfect and are subject to error. The error is due to the uncertainty in estimation as well as the natural variation of points about the regression line.

We can compute e.g. 95 % prediction interval for strains ε_a, ε_b, ε_c in particular directions marked as a, b, c (see the formula in 1, 10). Figures 1(b), 2(b), 3(b) show the 95 % prediction interval for strains in particular directions by using the best polynomial regression model.

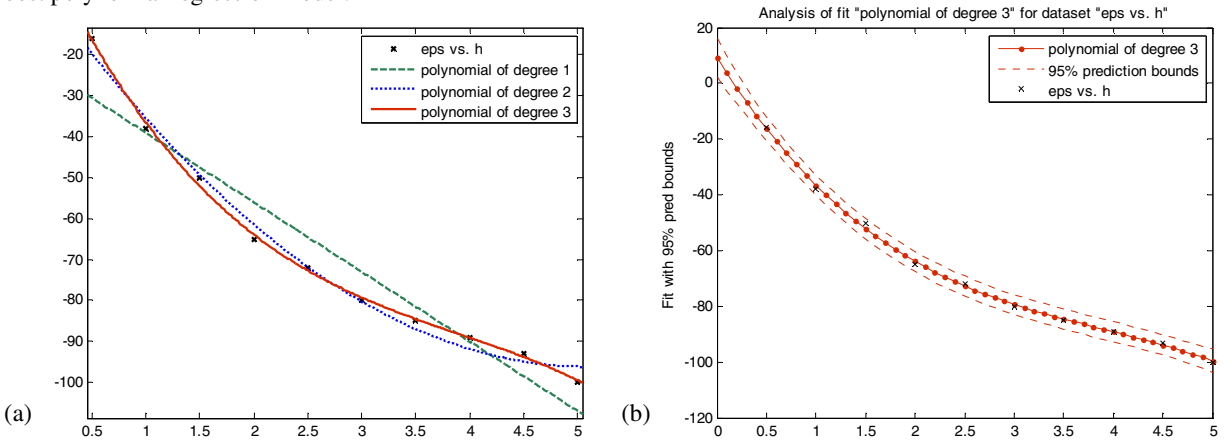

Fig. 1. (a) comparison of polynomial models with measured data – direction a, (b) 95 % prediction interval using cubic polynomial – direction a

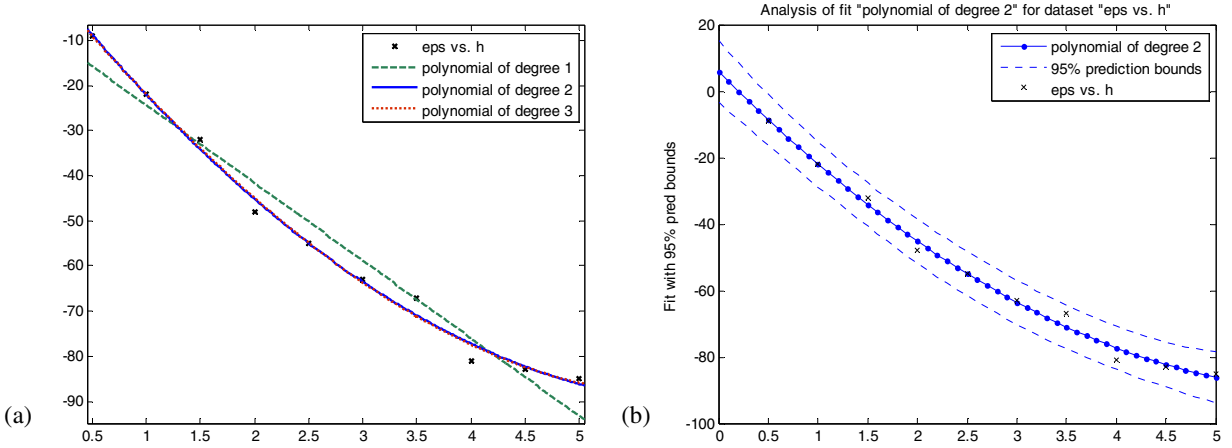

Fig. 2. (a) comparison of polynomial models with measured data – direction b, (b) 95 % prediction interval using quadratic polynomial – direction b

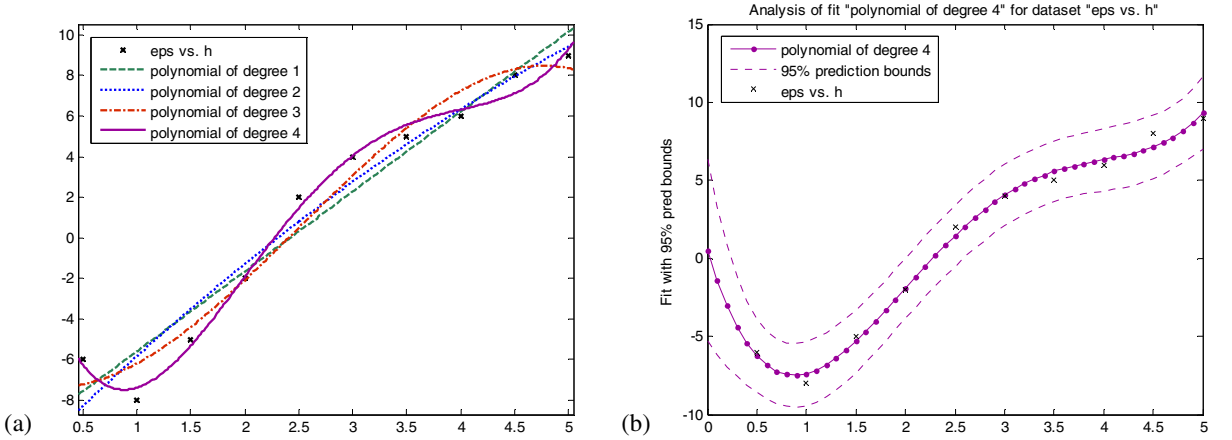

Fig. 3. (a) comparison of polynomial models with measured data – direction c, (b) 95 % prediction interval using quadratic polynomial – direction c

4. Conclusion

Regression analysis is a statistical tool for the investigation of relationships between variables. The multiple regression analysis is a useful method for generating mathematical models where there are several (more than two) variables involved. Polynomial regression model is consisting of successive power terms. Each model will include the highest order term plus all lower order terms (significant or not). We can view polynomial regression as a particular case of multiple linear regression. Polynomial models are an effective and flexible curve fitting technique.

The most widely used method of regression analysis is ordinary least squares analysis. This method works by creating a "best fit" line through all of the available data points and parameter estimates are chosen to minimize error sum of squares.

Fitting a regression model requires several assumptions. Estimation of the model parameters requires the assumption that the errors are uncorrelated random variables with mean zero and constant variance. Tests of hypotheses and interval estimation require that the errors are normally distributed. There are a number of advanced statistical tests that can be used to examine whether or not these assumptions are true for any given regression equation.

Acknowledgements

This article was created by implementation of the grant project VEGA no. 1/1205/12 Numerical modelling of machatronic systems.

References

[1] Aczel, A. D., 1989. Complete Business Statistics. Irwin, p. 1056. ISBN 0-256-05710-8.
[2] Frankovský, P., Kostelníková, A., Šarga, P., 2010. The use of strain-gage method and PhotoStress method in determining residual stresses of steel console. Metalurgija. Vol. 49, no. 2, p. 208-212.
[3] Frankovský, P., 2010. The use of experimental methods of mechanics for determination of residual stresses. (in Slovak), Dissertation, Košice, Slovak Republic.
[4] Lewis, C. D., 1982. Industrial and Business Forecasting Methods, London, Butterworths.
[5] Makridakis, S., Wheelwright, S. C., Hyndman, R. J., 1998. Forecasting methods and applications, New York, Wiley.
[6] Novotný, L., 2008. Utilization of finite element method in deformation calculation of assembly platform (in Slovak), Acta Mechanica Slovaca, pp. 607-612.
[7] Ostertag, O., Sivák, P., 2010. Degradation processes and fatigue life prediction (in Slovak), Typopress Košice, Slovakia, ISBN 978-80-553-0486-1.
[8] Ostertagová, E., 2005. Probability and Mathematical Statistics with examples (in Slovak), Elfa Košice, Slovakia, p. 123, ISBN 80-8086-005-X.
[9] Ostertagová, E., 2011. Applied Statistic (in Slovak). Elfa Košice, Slovakia, p. 161, ISBN 978-80-8086-171-1.
[10] Ostertagová, E., 2012. Applied Statistic with MATLAB (in Slovak). Equilibria Košice, Slovakia, p. 193, ISBN 978-80-8143-006-0.
[11] Trebuňa, F., et al., 2009. The solution for life extension of casting ladle supporting structure on the continuous casting machine 2 while taking advantage of until now available conclusions and proposals for life extension of the casting ladle supporting structure, final report, Technical University of Košice, Košice, p. 130.
[12] Trebuňa, F., Masláková, K., Frankovský, P., 2011. "Residual stress measurements," Modelling of Mechanical and Mechatronical Systems, Proceedings of the 4th international conference, Herľany, Slovakia, pp. 487-491.