Journal of Applied Sciences

ISSN 1812-5654

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (4): 513-524, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.513.524

On the Best Predictive General Linear Model for Data Analysis: A Tolerance Region Algorithm for Prediction

[1]C.P. Kitsos and [2]Vasilios Zarikas
[1]Department of Mathematics, Academic Technology Institute of Athens (ATEI of Athens),
12210 Aigaleo, Athens, Greece
[2]Department of Electrical Engineering, Theory Division, Academic Technology,
Institute of Lamia (ATEI of Lamia), 35100 Lamia, Greece

Abstract: There is a constant need for correct and meaningful statistical prediction. The General Linear Model (GLM) is a commonly used method to fit the data although most of the times the target is to construct a linear model in order to "predict" the value of the dependent variable; a goal for which GLM has not been designed for. The aim of the present study is to work on best model for a future observation, adopting the tolerance regions concept. A new method is explained and demonstrated, which is an alternative approach for choosing the optimal order of a response polynomial. The present study proposes a novel algorithm, which selects the best response polynomial, as far as prediction is concerned. The beta expected tolerance region is applied. The proposed computational approach has been applied for several data sets. This analysis, confirms the utility and the advantage of the method which provides non trivial results.

Key words: General linear regression, predictive models, tolerance intervals, algorithms, case studies

INTRODUCTION

In the well-known linear regression analysis (Seber, 1977; Stewart and Gill, 1991; Maddala, 1992) among others, the main target is to obtain a fitted response function for the collected data. The objective is to choose that model among the candidate models which fits the data and work so that this model "best" predicts a future response. The selected model should provide the "best", under some criterion, future response, for a given future input variable. In the present study, the best predictive linear model (Geisser, 1993; Yu and Ally, 2009) will be analyzed in order to be algorithmically determined.

Various potential uses of the regression equation are considered: pure description, prediction, control, model building etc. The current study focuses on prediction considering that the underlying model is linear. The procedure which is going to be developed adopts the expected β-content tolerance region for the future response (Guttman, 1970b) and the minimax criterion, in the sense that among all the "worst cases", those with the maximum tolerance region choose the minimum, the best one.

One important remark is that in the vast majority of the applied cases, a "make sense" model to fit the data of the problem is needed. However, this demands a general linear model to be constructed selecting the best fitting polynomial according to a criterion that ensures best future predictions. This model differs significantly from the best fitting polynomial with respect to a criterion of curve minimizing "distance" from data. The potential use of the proposed algorithm is prediction.

MATERIALS AND METHODS

Background and theory: The classical general linear model (GLM), is defined as:

$$Y = X\theta + \sigma e \tag{1}$$

where, $Y\in\Re^{n\times1}$, $X\in^{\Re n\times(p+1)}$, $\beta\in\Re^{(p+1)\times1}$, $e\in\Re^{k\times1}$ with $\Re^{k\times l}$ being the set of k×l matrices. From a statistical point of view Y is the observed random vector of responses, X is a matrix of known constants based on the p input variables, θ is a vector of unknown parameters and e is an unobserved random vector of errors with:

$$E(e) = 0,\ E(ee') = I_n \tag{2}$$

with $0\in\Re^{nx1}$ is a vector of zeros, $I_{n=}$ diag (1,1,...) is the unit matrix and $\sigma^2>0$ unknown. Usually, when statistical inference is performed, the following assumption is made:

Corresponding Author: Vasilios Zarikas, Department of Electrical Engineering Theory Division,
Academic Technology Institute of Lamia (ATEI of Lamia), 35100 Lamia,
Greece Tel: 00302231060265 Fax: 00302231033945

$$e \sim N\ (0,\ I_n) \tag{3}$$

with mean vector 0 and covariance matrix I_n . Given a realization of Y, a joint (1-α) 100% confidence region of the parameters θ can be constructed:

$$C(\theta)=\left\{\theta:(\theta-\hat{\theta})(X'X)(\theta-\hat{\theta})\le ps^2F(p,v;1-\alpha)\right\} \tag{4}$$

with v = n-p, α is the significant level and F, as usually, the F distribution, with p and v degrees of freedom and s^2 the (unbiased) estimate of σ^2. The ellipsoid (4) plays an important role either to impose experimental design criteria, or to decide which input variable X_1, X_2,…, $X_p \in \Re^{nx1}$ will be assumed participating to the model, (Hocking, 1976).

The confidence intervals are widely used in statistical estimation. The idea of a confidence interval is based on defining a region which contains the parameters under investigation with a certain probability level, usually (1-α)100% with α = 0.05. However in many applications like industrial applications it is desirable to have a region that contains a certain portion of the production with a predefined probability. That is the idea of the tolerance region, (Wilks, 1962; Guttman, 1970a; Boente and Farall, 2008) which seems the appropriate one.

In principle a statistical tolerance region is a statistic Q(X, y) from $\Re^n$ to the Borel σ-algebra B in $\Re^n$ and therefore exist functions L (X_1,...,X_n), U (X_1,..., X_n) such that:

$$Q(X,y)=Q(X_1,\ldots,X_n;y)=\left(L(X_1,\ldots,X_n),U(X_1,\ldots,X_n)\right) \tag{5}$$

$$\left(L(X_1,\ldots,X_n),U(X_1,\ldots,X_n)\right)=(L,\ U) \tag{6}$$

Sampling the random variables from a continuous cumulative distribution a tolerance region is of the form, (Wilks, 1962):

$$Q(X_1,\ X_2,\ldots,X_n;y)=\left[\left(X_{(k_1)},\ X_{(k_1+k_2)}\right)\right] \tag{7}$$

$$F\ \left(X_{(k_1+k_2)}\right)-F\left(X_{(k_1)}\right)\ \sim\ \text{Beta}\ (k_2,\ n-k_2+1) \tag{8}$$

with Beta (k, m) being the Beta distribution and $X_{(j)}$ is the jth order statistics. If a sample is taken from a continuous distribution function, with

$$k_1=r<\frac{n+1}{2}$$

then the confidence level ξ is:

$$\gamma=1-I_\beta(n-2r+1,2r)$$

where, I_β (p, q) is the incomplete Beta distribution. It seems more appropriate in applications to be referring to β-content tolerance region at confidence level γ if and only if:

$$P[P_X\{\ L(X_1,\ldots,X_n),\ U(X_1,\ldots,X_n)\}\ge\beta]=\gamma \tag{10}$$

For the β-expected tolerance region a dominant role plays the pioneering Lemma of Paulson (1943) for the defined lower and upper limit L = L (X) and U = U (X), respectively and for a given function T with distribution function G, the expected value of is β, E (A) = β:

$$A=\int_L^U dG(t) \tag{11}$$

Ellerton *et al.* (1986) introduced the idea to apply tolerance region for choosing the best linear model with one variable, while Kitsos (1994) adopted the invariance principle to extent this idea. Muller and Kitsos (2004) adopted the β-content tolerance regions to construct optimum experimental designs. Moreover, they proved that invariant tolerant regions are equal in both frameworks: Bayesian and Classical. In this study the β-content tolerance regions is adopted to construct the best predictive linear model.

β-expectation tolerance regions: Consider the General Linear Model (1). Given a realization y of Y it is desirable to construct a region Q(X, y) such that the vector Y* of the future observations Y_1^*, Y_2^*,…, Y_m^* will lie in Q(X, y)⊂ $\Re^m$ with a high probability. It is assumed that the vector Y* of future responses will follow model (1). Therefore for a given matrix X* of the input observations will hold:

$$Y^*=X^*\theta+\sigma e^* \tag{12}$$

with e*~ N(0, I_m), $0\in\Re^{mx1}$.

Thus, the distributions of Y*, is defined by the same parameters θ, σ from the parameter space $\Theta = \Re^P x \Re^+$ with elementsϑ = (θ, σ). Moreover for given parameter vector θ, Y and Y^* are assuming independent. Since the aim is to construct a region Q(X, y) it should be emphasize that is impossible to construct the region Q(X, y) so that:

- $Y^* \in$Q(X, y) with high probability
- The above is true for every parameter vectorϑ = (θ, σ) and every realization y of Y

The probability that Y* given lies in Q(X, y) is $P_{\grave{e}}[Y^{*} \in Q(X, y)]$. As Q(X, y), the tolerance region, cannot satisfy simultaneously 1 and 2 as above, the average tolerance region known as β-expectation tolerance region is used. Thus, by definition, Q(X,y) obeys to:

$$\int P_{\theta}[Y^{*} \in Q(X,y)] f_{Y|\theta}(y)dy = \beta \tag{13}$$

for every $\theta \in \Re^{p}$ (Guttman, 1970b).

The so defined β-expectation tolerance region can be proved that is also a prediction region (Muller and Kitsos, 2004).

Theorem: For the linear model (1) the β-expectation tolerance region Q(X, y), Classical or Bayesian, is evaluated:

$$Q(X, y) = \{w \in \Re^{m} : (w - X^{*}\hat{\theta})' S(X)(w - X^{*}\hat{\theta}) \leq \frac{m}{n-P} F_{m,n-p_1,\beta}\} \tag{14}$$

where, $\theta = (X'X)^{-1}X'y$ is the Least Square Estimate (LES) of θ and s^2, S(X) equals:

$$s^2 = \hat{\sigma}^2(X, y) = (y - X\hat{\theta})'(y - X\hat{\theta}) \tag{15}$$

$$S(X) = I_m + X^{*}(X'X)^{-1}X^{*\prime} \tag{16}$$

with $F_{m,\, n\text{-}p;\, \beta}$ the β quantile of the F distribution with m and n-p degrees of freedom. It can be also proved, that:

$$S(X) = I_m - X^{*}(X'X + X^{*\prime}X^{*})^{-1}X^{*} = I_m - M(x) \tag{17}$$

with the definition of M (x) obviously obtained. This notation is followed at the presented algorithm. It can further be proved that the β-expectation tolerance region defined as in (14), is also a Bayesian one with respect to θ. Following (14) and working for the one variable degree polynomial the length L_p (x) of the tolerance region can be derived equal to:

$$\begin{aligned} L_p(x_o) &= 2t_{n\text{-}p}(ß/2)(n\text{-}p)^{-1/2}\, s_p^{\,1/2}\left(\frac{RSS_p}{S^{-1}(x)}\right)^{1/2} \\ &= \text{const}\left(\frac{1}{S^{-1}(x)}\right)^{1/2} \approx \left(\frac{1}{1-M(x)}\right)^{1/2} \end{aligned} \tag{18}$$

but it is easy to prove that (Ellerton *et al.*, 1986):

$$M(x) = X'_{op}CX_{op} - \frac{(X'_{op}CX_{op})(X'_{op}CX_{op})}{1 + X'_{op}CX_{op}} \tag{19}$$

with $C = (X'X)^{-1}$. If $m(x) = X'_{op}CXt$ then:

$$M(x) = \frac{m(x)}{1 + m(x)} \tag{20}$$

To get the maximum of M(x) it is required that $M(x) = 0 \Leftrightarrow X'_{op}Cx_{op} = 0$ with the dot sign denoting the corresponding derivative. Thus:

$$\ddot{M}(x) < 0 \Leftrightarrow \ddot{X}'_{op}\, C\, X'_{op} + \dot{X}'_{op}\, C\, \dot{X}_{op} < 0 \tag{21}$$

For a complete presentation see also the constructed algorithm Appendices A and B.

Algorithm for the best predictive model: For a given future response, under the linear model (1), the β-expectation tolerance region can be constructed. To identify the best predictive model the following algorithm is proposed based on the volume of the "future" ellipsoid. The largest volume of the β-expectation tolerance region corresponds to the worst case of the input variables set, as far as prediction concerns. The minimum β-expectation tolerance among the worst models is those with max β-expectation tolerance region is the best one.

An algorithm based on the mini-max criterion can be constructed now, in order to find the "best" linear model, based on the following basic steps:

- **Step 1:** Fit all possible linear models for the subsets with k variables from p, k = 1, 2,..., p, normalizing x:X∈[-1,1]
- **Step 2:** For the corresponding k variables calculate the β-expectation tolerance region Q_k and select that k which corresponds to the largest β-expectation tolerance region
- **Step 3:** Among the max tolerance regions for the different k = 1, 2, .., .p choose the minimum one. So choose the best subset of variables which corresponds to $k_0 = \min_p \max_j.\ \{Q_{kj}\ (X,y),\ k = 1,2,..., p; j = 1,2\}$

The above Algorithm is applied to a number of applications and different datasets (see section 3), investigating its behavior on various circumstances. Based on the previous discussion, the following algorithm has been introduced, for fitting the "best" predictive model for n measurements $\{x_i, y_i\}$.

```
1. Read data X and data Y
2. Normalize data X in the interval [-1,1]
3. For p = 0 to k
4. Evaluate matrix $X=X_p$ for the pth order model
5. Define matrix $X_{op}$
6. Evaluate vector $\theta$ with the estimators of GLM model of pth order
   $\hat{\theta}=(X'X)^{-1} X'Y$
7. Solve the equation:
   $\dot{X}'_{op} (X'X)^{-1} X_{op} = 0$
8. Check if its roots satisfy the relation:
   $\ddot{M}(x) = \ddot{X}'_{op} (X'X)^{-1} X'_{op} + \dot{X}'_{op} (X'X)^{-1} \dot{X}_{op} < 0$
9. Evaluate the length $L_p$ at the point x which satisfy steps 7,8
   $L_p(x_o) = 2t_{n-p;1-\beta/2} (n-p)^{-1/2} s_p^{1/2}\{(I-X'_{op}(X'X+X'_{op}X_{op})^{-1} X_{op})^{-1}\}^{1/2}$
   Where, $s_p = RSS = (Y-X\hat{\theta})'(Y-X\hat{\theta})$
10. E aluate function $L_p$ at the end points-1 and 1.
11. Store the maximum $L_p(.)$ value of those obtained from (8) (9) and (10).
12. Equivalently steps (7), (8), (9) and (10) can be replaced by calling a subroutine that evaluates x that gives
    Max $[X'_{op} (X'X)^{-1} X_{op}]$ with $-1 \leq x \leq 1$
    $RMS = \frac{RSS}{\sqrt{n-p-1}}$
13. Evaluate RMS for the best fitted polynomial of p-th order
14. Repeat parts (4)-(13) for p = 0,1,2,…,k.
15. Choose the minimum of the stored maximum "lengths".
```

The corresponding p value is the degree of the response function for the best predictive model. In addition the best fit model according to the conventional method is the one with the minimum RMS.

The flow chart of a critical part of the algorithm is given in the Appendix A, while a particular implementation in Mathematica is shown in Appendix B.

RESULTS APPLYING THE METHOD

In this section various results are presented that show the added value of the proposed methodology compared with the commonly used practice to choose the fitting model. The databases in use, have been selected to carry different characteristics, in order to test the algorithm and the method.

Dataset I : Strong linear correlation: First, the dataset I is studied. It comprises data that exhibit a strong linear correlation. Data X: x = 1, 1, 2, 3, 4, 5, 7. Data Y: y = 7.1, 7, 10.1, 12.1, 15.1, 18.1, 23.1. Table 1 shows the estimations based on the algorithm and the corresponding program in the Appendix B. Based on the procedure discussed in Section 2 the value $[L_p(x)] = 1.29$ is estimated (fourth column of Table 1) and hence the corresponding model is the second order polynomial:

$$\hat{Y} = 15.1447 + 8.047x$$

This is the model which according to the proposed prediction criterion (based on tolerance regions framework) "best" predicts the future observation, (Fig. 1). However, this is not the model according to the traditional distance RMS criterion. The latter criterion selects the fifth order polynomial (Figure 2). Most of the methods which use a distance criterion for choosing the most appropriate polynomial have a tendency to peak a large order polynomial, as it is shown in Fig. 2. It is obvious that even for this very simple dataset the proposed method suggests a non trivial difference. If the designer of an experiment wills to peak a polynomial that best fits data, as far prediction is concerned, then the proposed method gives distinctive results.

Dataset II: Not strong correlation: Here the dataset II is considered:

- **Data x:** 1,1,1,2,3,4,5,6,6,6,7,7,7,7,7,7,8,9,8,8,9,9,9, 9,10,10,4,4,5,6,11,11,12,12,13,13,14,14,15,15,16,16,17, 17,18,18,19,19,20,21,22,21,22,24,20,21,22,23,24,25,26, 27,28,29,30
- **Datay:** 14,13,12,10,11,10,9,9,8,8,7,7,6,7,7,6,12,11,10, 9,8,7,6,8,9,6,10,7,7,6,6,5,5,7,7,6,6,7,7,6,6, 8,8,6,6,7,5,5,4,3.5,4.5,5,5,5,6,6,6,7,7,8,6,7,8,7,8

This dataset comprises no strong correlation as it can be observed in Figures 3, 4. The relevant mathematical quantities have been evaluated and presented in Table 2, where the critical quantity min {max [max $[L_p(x)]$} = 6.536 is depicted. Hence, the corresponding model is the second order polynomial:

$$\hat{Y} = 5.84795 - 1.70747\,x + 3.8513\,x^2$$

This is the model which according to the suggested criterion "best" predicts the future observation, (Fig. 3). However, the model according to the traditional RMS criterion is the fifth order (Fig. 4).

Table 1: General Linear models up to fifth order and their evaluation according to both criteria for dataset I. RMS_p denotes the factor RMS of the p-th order polynomial while $maxL_p$ denotes the maximum value of the length L_p of the p-th order polynomial

P	Fitted Model	RMS_p	$maxL_p(\bullet)$
0	$Y = 13.23$	35.349	25.407
1	$Y = 15.1447 + 8.047x$	0.066	1.290
2	$Y = 15.166+8.042x-0.043x^2$	0.081	1.561
3	$Y = 15.183+8.299x-0._{061x2}-0.291x^3$	0.102	1.735
4	$Y = 14.936+ 8.915x+2.895x^2-0.896x^3-2.753x^4$	0.040	1.549
5	$Y =15.1+9.806x+0.003x^2-7.940x^3-0.028x^4+6.159x^5$	0.005	1.320

Table 2: General Linear models up to fifth order and their evaluation according to both criteria for dataset II. RMS_p denotes the factor RMS of the p-th order polynomial while $maxL_p$ denotes the maximum value of the length L_p of the p-th order polynomial

P	Fitted Model	RMS_p	$maxL_p(\bullet)$
0	Y = 7.32308	4.573	8.779
1	Y = 6.98406-2.16625 x	3.246	7.651
2	Y = 5.84795-1.70747 x+3.8513 x^2	2.105	6.536
3	Y = 5.88263-1.22885 x+3.75222 x^2-0.853556 x^3	2.119	6.961
4	Y = 6.356-1.729 x-0.538 x^2-0.099 x^3+4.777 x^4	1.992	7.109
5	Y =6.278-3.754 x+0.517 x^2+9.571 x^3+3.527 x^4-8.407 x^5	1.900	7.263

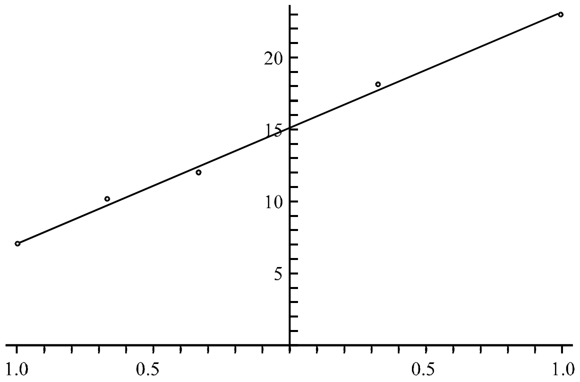

Fig. 1: Best fitting polynomial for prediction Y = θ X for dataset I.×represents normalized X to the interval [-1,1]

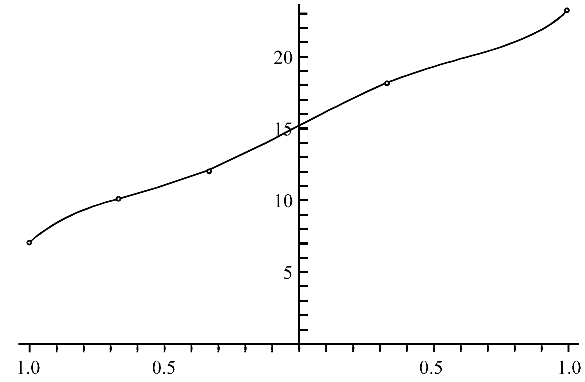

Fig. 2: Best fitting polynomial Y = θ X according to the RMS (Root Mean Square) criterion for dataset I.× represents normalized X to the interval [-1,1]

As it can be seen in the scatter diagrams, Fig. 3 and 4, there is not a strong correlation between Y and X data in the measurement space. In such cases, the selected RMS model differs almost always from the "best" model for prediction in the initial region of interest.

Dataset III: Strong nonlinear correlation:

- **Data x:** 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 , 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30
- **Data y:** 6, 7, 8, 9, 10, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 3.5,4.5, 6, 6, 6, 7, 7, 8, 6, 7, 8, 7, 8

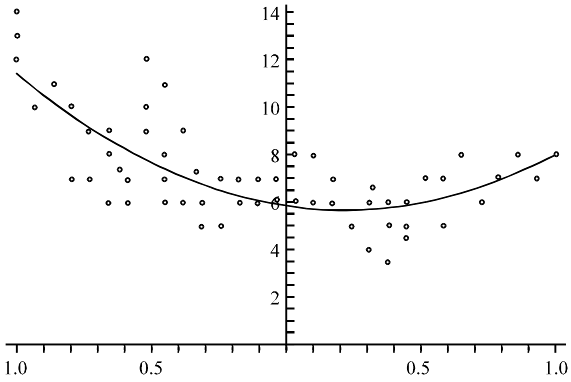

Fig. 3: Best fitting polynomial for prediction Y = θ X for dataset II×represents normalized X to the interval [-1,1]

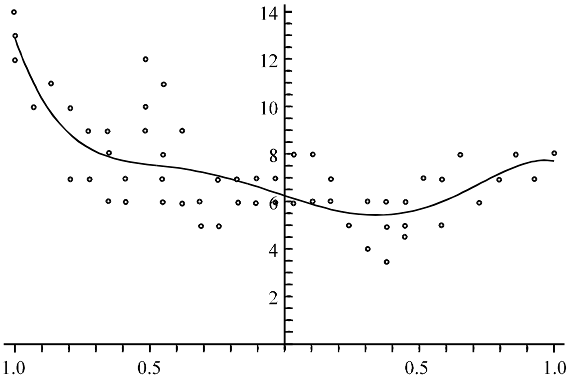

Fig. 4: Best fitting polynomial Y = θ X according to the RMS (Root Mean Square) criterion for dataset II× represents normalized X to the interval [-1,1]

It can be understood from the scatter diagram, (Fig. 5), that there is a quite strong nonlinear correlation between Y and X data for a considerable portion of the data. Based on the analyzed procedure the value min {max [max [L_p (x)]} = 5.334 is associated to the fourth order polynomial:

$$\hat{Y}= 4.519 -5.720 x+17.61 x^2+6.832 x^3-16.336 x^4$$

This is the model, shown in Fig. 5, which according to the proposed criterion "best" predicts the

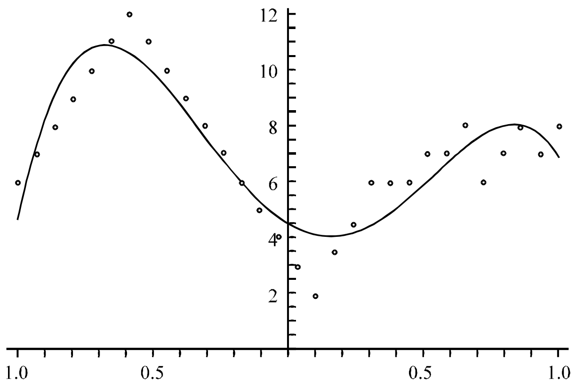

Fig. 5: Best fitting polynomial for prediction Y = θ X for dataset III.×represents normalized X to the interval [-1,1]

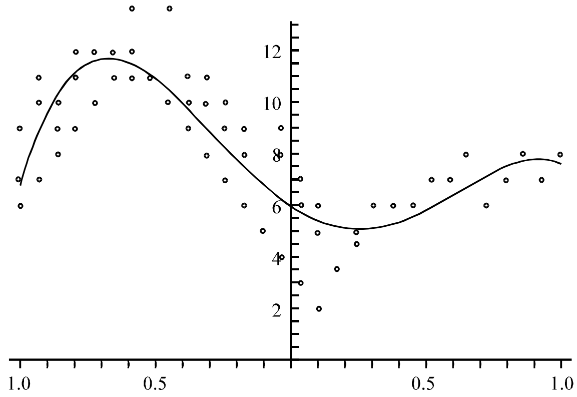

Fig. 6: Best fitting polynomial for prediction Y = θ X for dataset IV.×represents normalized X to the interval [-1,1]

future observation. In this case, this is the model according to the traditional RMS criterion too.

Dataset IV: Not strong nonlinear correlation:

- **Data x:** 4, 4, 5, 5, 6, 1, 2, 2, 3, 3, 1, 1, 1, 2, 3, 4, 5, 6, 7, 7,7, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 19, 19, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 15, 15, 16, 16, 17, 17,10, 10, 11, 11, 12, 12, 13, 13
- **Data y:** 11, 12, 12, 13, 12, 9, 10, 11, 9, 10, 6, 7, 7, 7, 8, 9, 10, 11, 12, 13, 11, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 3.5, 4.5, 5, 5,5, 6, 6, 6, 7, 7, 8, 6, 7, 8, 7, 8, 8, 9, 6, 7, 5, 6, 11, 10, 11, 10, 10, 9, 9, 8

In this case, the algorithm evaluates the value min {max [max [L_p (x)]} = 6.849 and hence the corresponding model is the fourth order polynomial:

$$\hat{Y}= 5.981 -6.955 x+12.309 x^2+7.359 x^3-11.065 x^4$$

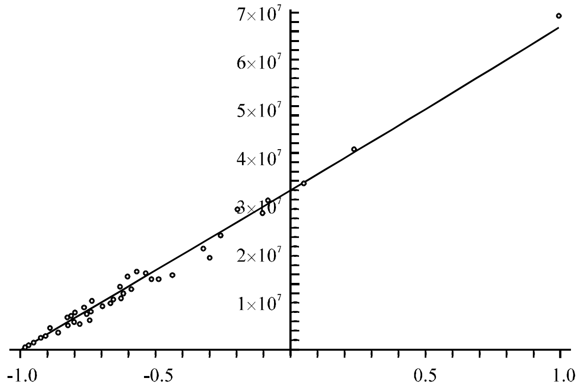

Fig. 7: Best fitting polynomial for prediction Y = θ X.× represents the normalized to the interval [-1, 1] X which is the number of industry employees in USA (2007) and Y their annual payroll

This is the model, illustrated in Fig. 6, which "best" predicts the future observation in conformity with the discussed criterion. However, the model according to the traditional RMS criterion is the fifth order. As can be seen in Fig. 6, there is no strong nonlinear correlation between Y and X data. It is now apparent from the comparison of both Fig. 5 and 6 that the proposed methodology suggests different polynomial as the best for prediction in cases of data that are more dispersed and with nonlinear behavior.

A business example: In this example data refer to Manufacturing-Selected Industry Statistics by USA State (http://www.census.gov/compendia/statab/2012/tables/12s1008.pdf). A sociologist in order to write a study regarding employment needs to have a general linear model that relates the number of employees in USA (X) as a function of their annual payroll (Y). It is easy to understand that a scientist often does not seek to confirm an underlying fundamental law (as in physics) but an expression that relates two quantities which correlate. It is obvious that a correct model is provided by a polynomial ensuring best prediction (new values for X within experimental region). This is the model a scientist should search for.

The program of Appendix B provides results, (Table 3), showing that the best model for prediction is the first order polynomial, shown in Fig. 7, while the RMS criterion suggests the second order model. Thus the proposed method leads to a different simpler model but most importantly best as far as new "observations" are concerned. Table 3 encompasses results where the

Table 3: General Linear models up to fifth order and their evaluation according to both criteria. RMS_p denotes the factor RMS of the p-th order polynomial while max L_p denotes the maximum value of the length L_p of the p-th order polynomial (number of employees in USA, x, and their annual payroll, Y, at 2007)

P	Fitted model	RMS_p	$maxL_p(\bullet)$
0	$Y = 1.2\ 10^7$	$1.72\ 10^{14}$	$5.40\ 10^7$
1	$Y = 3.41\ 10^7+3.45\ 10^7x$	$1.57\ 10^{12}$	$6.04\ 10^6$
2	$Y = 3.37\ 10^7+3.56673*10^7x+2.07\ 10^6x^2$	$1.31\ 10^{12}$	$6.44\ 10^6$
3	$Y = 3.38\ 10^7+3.60\ 10^7x+1.93\ 10^6x^2-478079x^3$	$1.33\ 10^{12}$	$6.65\ 10^6$
4	$Y=3.39\ 10^7+3.86\ 10^7x+6.26\ 10^6x^2-2.96\ 10^6x^3-4.58\ 10^6x^4$	$1.32\ 10^{12}$	$8.99\ 10^6$
5	$Y =3.40\ 10^7+3.81\ 10^7x+419270x^2-9.31\ 10^6x^3+ 1.01\ 10^6x^4+6.87*10^6x^5$	$1.34\ 10^{12}$	$2.1\ 10^7$

Table 4: General Linear models up to fifth order and their evaluation according to both criteria. RMS_p denotes the factor RMS of the p-th order polynomial while $maxL_p$ denotes the maximum value of the length L_p of the p-th order polynomial (x is the number of available engineers and Y denotes Industry R and D expenditures for 2006)

P	Fitted Model	RMS_p	$maxL_p(\bullet)$
0	$Y = 4781.43$	$7.87\ 10^7$	36502
1	$Y = 22606+23672.4x$	$1.35\ 10^7$	18629
2	$Y = 19710.6+27071.2\ x+8003x^2$	$9.88\ 10^6$	17783
3	$Y = 13115.1+12198.5\ x+15457.4x^2+17235.2x^3$	$6.78\ 10^6$	14987
4	$Y = 13638.5-195.1\ x-6922.29x^2+29383.4x^3+22510.5x^4$	$6.14\ 10^6$	21210
5	$Y =13857-549\ x-10598\ x^2+26242x^3+25949x^4+3515x^5$	$6.28\ 10^6$	74300

minimum value of the fourth column belongs to the first order polynomial. Figure 7 shows the best predictive fitting polynomial.

An investment decision example: In this example, data concern USA Science and Engineering Indicators of year 2006, (http://nsf.gov/statistics/seind08/c0/c0a.htm). A bank loan department in order to drive strategy regarding industrial research funding, wants to have a crude estimation regarding the relation of the number of available engineers as a function of industry R and D expenditures. It is obvious in this case too, that the target is to find a best fitting polynomial allowing for trusty predictions and not a polynomial "passing" closer to existing observations.

The algorithm of Appendix B suggests that the best model for prediction is the third order polynomial while the RMS criterion suggests choosing the fourth order model, Fig. 8. Thus the suggested method leads to a simpler model and most importantly best for predicting values inside the experimental space. Table 4 encompasses results, indicating a minimum value in the fourth column equal to 14987 and a minimum value in the third column equal to $6.14\ 10^6$. Figure 8 shows the best fitting first order polynomial.

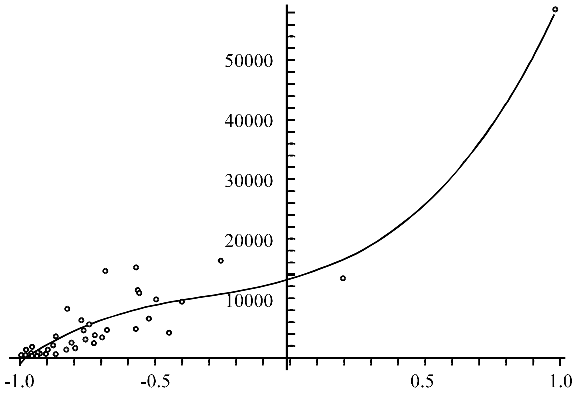

Fig. 8: Best fitting polynomial for prediction $Y = \theta X$.× represents the normalized to the interval [-1, 1] X which is the number of available engineers in USA States and Y the amount of industry R and D expenditures (for year 2006)

A policy making example: In this example, data concern Top U.S. Foreign Trade Freight Gateways by Value of Shipments (Current $ billions) referring to 2008, (http://www.bts.gov/publications/national_transportation_statistics/).

A government policy maker is interested to find a vague model with the help of a General Linear Model that elates exports (x) with imports (y), based on data from various major water and air USA gateways. It is obvious that a best fitting polynomial according to the proposed prediction criterion is the one that is needed to be found. The implemented algorithm of Appendix B reveals that the best model for prediction is the first order polynomial since the minimum value minimum v alue min {max [L_p (x)]} = 111.3 belongs to the p = 1 polynomial in Table 5. Furthermore, the RMS best model suggests choosing the fourth order model, since the minimum RMS value is equal to 498, as shown in Table 5. Therefore the presented method leads not only to a simple model but also to the best for predicting values inside the experimental space. Figure 9 shows the best fitting, first order polynomial, as well as the data exhibiting large dispersion.

Table 5: General Linear models up to fifth order and their evaluation according to both criteria. RMS_p denotes the factor RMS of the p-th order polynomial while $maxL_p$ denotes the maximum value of the length L_p of the p-th order polynomial (x represents the exports and Y denotes imports (2008)

P	Fitted model	RMS_p	$maxL_p(\bullet)$
0	$Y = 32.91$	1012.83	130.93
1	$Y = 57.87+47.19x$	588.38	111.28
2	$Y = 63.06+40.93x-18.01\ x^2$	579.79	125.56
3	$Y = 69.43+69.53x-29.27x^2-44.49x^3$	555.02	132.05
4	$Y = 78.93+45.77x-131.65x^2-20.91x^3+109.51x^4$	498.987	128.13
5	$Y =80.68\ 54.94x-144.79x^2-64.00x^3+120.13x^4+35.57x^5$	509.30	129.91

Table 6: General Linear models up to fifth order and their evaluation according to both criteria. RMS_p denotes the factor RMS of the p-th order polynomial while $maxL_p$ denotes the maximum value of the length L_p of the p-th order polynomial (Y abdominal circumference and x gestational age)

P	Fitted model	RMS_p	$maxL_p(\bullet)$
0	$Y = 363.3$	347.51	76.69
1	$Y = 364.08-5.58x$	345.04	79.60
2	$Y = 367.6-4.97x-11.28x^2$	342.90	84.81
3	$Y = 367.46-16.31x-9.78x^2+20.42x^3$	341.39	90.59
4	$Y = 364.57-19.71x+17.40x^2+23.80x^3-31.91x^4$	342.86	97.17
5	$Y =364.42-24.0x+16.94x^2+43.17x^3-31.30x^4-17.07x^5$	350.28	102.92

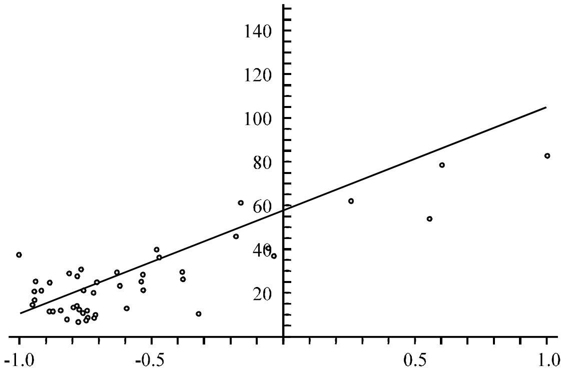

Fig. 9: Best fitting polynomial for prediction Y = θ X× represents the normalized to the interval [-1, 1] X which is the exports with Y denoting imports. Data from various major water and air USA gateways Top U.S. Foreign Trade Freight Gateways by Value of Shipments (Current $ billions, year 2008)

A medical example: In this example, data concern modeling abdominal circumference as a function of gestational age in weeks (Hosmer and Royston, 2003). A gynecologist is interested to find a polynomial relation between abdominal circumference and gestational age for the last time length of the gestational period (abd.circ.>40). Since this empirical law will be used for future observations and predictions the doctor need to find the best fitting polynomial according to the proposed method. It is easy to find using the algorithm of Appendix B that the best model for prediction is the first order polynomial $y = 364-5.6\ x$ while the RMS method suggests to choose the model $y = 367.46-16.31\ x-9.78x^2+20.42\ x^3$. Thus the current study leads to a simpler and most appropriate for prediction model. Table 6 encompasses results where it can be seen that the minimum value of the fourth column happens for p = 1 polynomial.

DISCUSSION

It is important in this section to discuss findings providing extensive interpretation and description of differences with respect to other methods of polynomial data fitting.

At this point it is considered useful to give a brief remind of the three distinct intervals that appear in statistical analysis of data. In case of fitting a parameter to a model, the accuracy or precision can be expressed as a confidence interval, a prediction interval or a tolerance interval which are quite distinct. Confidence intervals provide information about how well the best-fit parameter determined by regression has been estimated. Taking many samples from a Gaussian distribution it is expected about 95% of those intervals to include the true value of the population best fit parameter. The crucial remark is that the confidence interval informs about the likely location of the true population parameter. Prediction intervals on the other hand inform where you can expect to find the next data point sampled. Collecting many samples (Gaussian distribution), it is expected next value to lie within that prediction interval in 95% of the samples. The key remark is that the prediction interval informs about the distribution of values, not the uncertainty in determining the population parameter. The richest interval is tolerance interval. It is determined by two different percentages. The first determines “how sure” it is desired the value to be and the second expresses what fraction of the values the interval will contain. In case the first value (how sure) is set to 50%, then a tolerance interval is the same as a prediction interval. If it is set to a higher value (say 95%) then the tolerance interval is wider.

Most of the statistical criteria in model selection for applications (Maddala, 1992) among others, are working towards the target: find the model, which "best", under some criteria, fits the data. These criteria are, in principle, functions of the Residual Sum of Squares (Stigler, 1981).

When the investigated mechanism between the controlled variables and the (single) response is an engineering or economic functional model (Draper and Smith, 1998; Urbain, 1989), this statistical approach seems suitable. For dynamic models and non-independent errors the Econometric Models for example are suitable. Usually these best fitting models are applied for prediction too, although the "distance" criteria are not designed for this purpose. In the present study the problem is tackled from a different perspective. The chosen model is the one that best predicts on the average the future value, which lie on a certain interval with some probability. This is achieved using beta expected tolerance regions. So, while the regression oriented prediction is based on the extrapolation or interpolation of the best model fitting the data, the proposed method is based on a probabilistic reasoning and provides that model which best predicts next value within experimental region.

They key concept that allows for the developed alternative approach is the tolerance regions Chew (1966) provided a comparison between confidence regions and tolerance regions. However, the entire background is completely different (Guttman, 1970a). Various methods have been established to obtain the minimum ellipsoid (Pronzato and Walter, 1994; Bland *et al.,* 1981; Cheung *et al*., 1993) from an optimum design approach. As far as the tolerance regions concern it has been proved (Muller and Kitsos, 2004) that the classical tolerance regions coincide with the Bayesian one. But the essential difference of the current study is that the target is completely different: the beta-expected tolerance region for the future observation it is not used as a design criterion, but as a model fitting criterion and provides, as it is discussed, the best predictive general linear model. This is the essential difference of the work attempted with the existed theory: there is no model fitting attempted so far under the tolerance region framework.

As a last remark it should be mentioned that the cross validation method (Shao, 1993) should not be confused with the present study. In this method the background is completely different since the idea is to try to test how well a polynomial performs for prediction using a portion of data. On the other hand, the present study utilizes all data and selects the best polynomial for prediction.

Thus the presented method differs nontrivially as far as the interpretation is concerned from all other methods that best fit the data with a distance criterion. Furthermore, the current method suggests in many cases different models. This was shown for all investigated data cases which were belonged to different fields of application. Furthermore, evaluation with the help of the developed algorithm reveals that there is enough evidence that the proposed method works well. In summary, the findings of the numerical study of the methodology reveal that the selected polynomial model according to the best prediction criterion does not have the inherit property to peak the largest order polynomial as the best model. This is a not a desired property usually associated with methods using distance criteria (RMS). In addition, the analysis of all datasets shows that for data with large dispersion, the proposed method peaks always a polynomial of different order from this suggested by an RMS criterion.

CONCLUSION

The study of the results revealed an affirmative conclusion for using the proposed method in scientific and technological applications. There is a strong theoretical background that ensures the success of the method to any applied field. It was shown that for several datasets the selected polynomial differs from the commonly selected one, if the choice respects the criterion of "the best predictive model". This is not to mean that there are no cases where the two methods suggest the same polynomial.

Therefore, it would be safe to conclude that for most models in sciences, such as Economics, Medicine, Psychology, Sociology as well as in Industry and Quality control, the presented proposal provides a powerful insight into scientific/research practice, adding a real, indispensable value to it. As a result, in cases such as the ones outlined in this study, it is not considered useful and correct modeling to find a curve as closely as possible to the data. Instead, it is desirable to guarantee that a curve is going to fit to the given set of data based on a best fit polynomial that is most adequate for the prediction of the Y value for a certain X (within the experimental region), establishing a specific degree of high probability.

As a future research it is worth to generalize the proposed method for problems with multiple independent variables or for cases like (Gikas and Stratakos, 2012; (Zarikas *et al.,* 2010). It would also be interesting to develop integration with the experiment design approach (Mead, 1991; Mejza and Mejza, 2012). Although, in most of the cases in classical experiment design theory the models are linear, still these approaches are based on the typical regression analysis (Oliveira and Oliveira, 2012; Valente and Oliveira, 2011; Pereira *et al.,* 2012) and not to the tolerance regions adopted in this study, beyond the regression analysis to fit the model. Another interesting investigation is to develop a strategy handling extrapolation. Extrapolation needs special care for the proposed method since an extension of the experiment/measurement space is needed.

ACKNOWLEDGMENT

V. Zarikas acknowledges that this work has been supported by the research program “Optimum distributed genration design (DGRES)” MIS 380360 , within the Greek Research Activity “Archimedes III”, funded by the NSRF 2007-2013.

APPENDIX A. FLOW CHART

A critical part of the algorithm regarding the appropriate selection of the order of the polynomial is shown in the form of a flow chart in Fig. 10, 11.

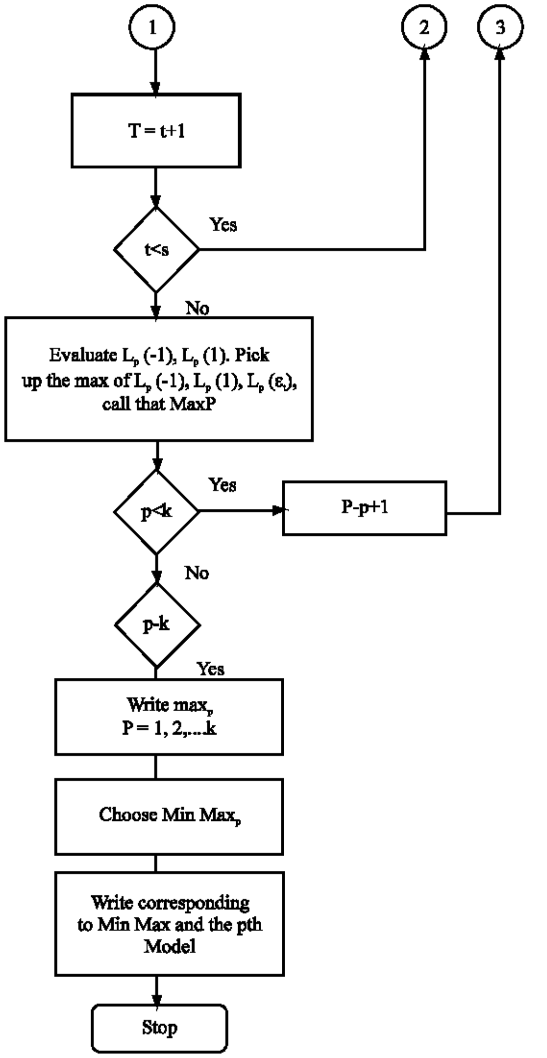
1
2
3
T = t+1
t<s
Yes
No
Evaluate L_p (-1), L_p (1). Pick up the max of L_p (-1), L_p (1), L_p (ε_t), call that MaxP
p<k
Yes
P-p+1
No
p-k
Yes
Write $\max_p$ P = 1, 2,...k
Choose Min Max_p
Write corresponding to Min Max and the pth Model
Stop

Fig. 10: A critical part of the algorithm regarding the appropriate selection of the order of the polynomial. Flow chart, part 1 of 2

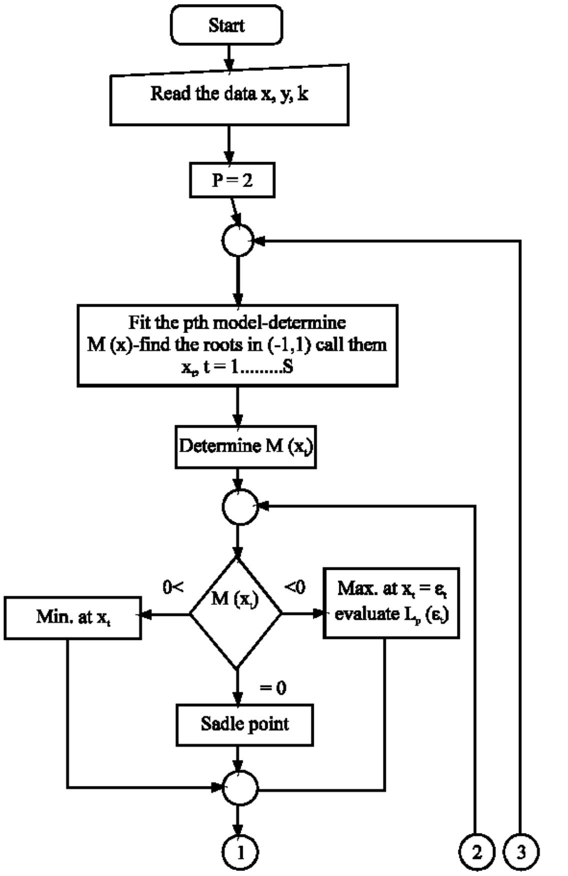
Start
Read the data x, y, k
P = 2
Fit the pth model-determine M (x)-find the roots in (-1,1) call them x_t, t = 1.........S
Determine M (x_t)
0<
M (x_t)
<0
Min. at x_t
Max. at $x_t = \varepsilon_t$ evaluate L_p (ε_t)
= 0
Sadle point
1
2
3

Fig. 11: A critical part of the algorithm regarding the appropri te selection of the order of the polynomial. Flow chart, part 2 of 2

APPENDIX B. CODE FOR “MATHEMATICA”

```
datax = ……

datay = …..
data = Table[{xTRN[[m]], datay[[m]]}, {m, Dimensions[datax][[1]]}];

Y = Table[{datay[[n]]}, {n, Dimensions[datax][[1]]}]
tstud[n_] :=
Sqrt[-n+ (1/n*(0.05*(Sqrt[n]*Beta[n/2, 1/2]))^(2/(1+ n)))^(-1)];

n := Dimensions[datax][[1]];

DDOWN = Min[datax];
UUP = Max[datax];

A := (UUP+ DDOWN)/2 ;
B := UUP-A;
xTRN := (datax-A)/B

X[0] := Table[{1}, {i, Dimensions[datax][[1]]}];

X[1] := Table[{1, xTRN[[i]]}, {i, Dimensions[datax][[1]]}];

X[2] := Table[{1, xTRN[[i]], xTRN[[i]]^2}, {i,
 Dimensions[datax][[1]]}];
X[3] := Table[{1, xTRN[[i]], xTRN[[i]]^2, xTRN[[i]]^3}, {i,
Dimensions[datax][[1]]}];
X[4] := Table[{1, xTRN[[i]], xTRN[[i]]^2, xTRN[[i]]^3,
 xTRN[[i]]^4}, {i, Dimensions[datax][[1]]}];

X[5] := Table[{1, xTRN[[i]], xTRN[[i]]^2, xTRN[[i]]^3, xTRN[[i]]^4,
 xTRN[[i]]^5}, {i, Dimensions[datax][[1]]}];

Xop[0] := {{1}}; Xop[1] := {{1}, {t}}; Xop[2] := {{1}, {t}, {t^2}};

Xop[3] := {{1}, {t}, {t^2}, {t^3}};

Xop[4] := {{1}, {t}, {t^2}, {t^3}, {t^4}};

Xop[5] := {{1}, {t}, {t^2}, {t^3}, {t^4}, {t^5}};

EXPR := (Xop[i]\[Transpose].Inverse[(X[i]\[Transpose].X[i])]. Xop[i])[[1]];

MAX[nn_] := (i = nn; NMaximize[{EXPR[[1]],-1 <= t <= 1}, t] );

LP := 2*tstud[Dimensions[datax][[1]]-i]/Sqrt[
 Dimensions[datax][[1]]-i] ((INVS1p)^(1/2))*(RSSp)^(1/2);

β := Inverse[X[i]\[Transpose]. X[i]].(X[i]\[Transpose]. Y);

RSSp := Transpose[Y-X[i]. â] . (Y-X[i]. â);

RMS := RSSp/(Dimensions[datax][[1]]-i-1);

INVS1p := Inverse[
 1-Xop[i]\[Transpose].Inverse[(X[i]\[Transpose].X[i]+
 Xop[i].Xop[i]\[Transpose])]. Xop[i]];

 Do[ i = k;
 Print["i=", i, " max", NMaximize[{EXPR[[1]],-1 <= t <= 1}, t] ];
 gg = Plot[EXPR[[1]], {t,-1, 1}]; g1 = ListPlot[data];
 g2 = Plot[Xop[i]\[Transpose].β, {t,-1, 1}];
 Print[Xop[i]\[Transpose].β];
 asd := NMaximize[{EXPR[[1]],-1 <= t <= 1}, t] [[2]][[1]];
 Print["RMS=", RMS]; t = t /. asd; Print["LP=", LP]; Print[Show[gg]];
 Print[Show[g1, g2, PlotRange-> All]]; t =., {k, 0, 5}]
```

REFERENCES

Bland, G.R., D. Goldfarb and M.J. Todd, 1981. The ellipsoid method: A survey. Oper. Res., 29: 1039-1091.

Boente, G. and A. Farall, 2008. Robust multivariate tolerance regions: Influence function and Monte Carlo study. Technometrics, 50: 487-500.

Cheung, F.M., S. Yurkovitch and M.K. Passino, 1993. An optimal volume ellipsoid algorithm for parameter set estimation. IEEE Trans. Autom. Control, 38: 1292-1296.

Chew, V., 1966. Confidence, prediction and tolerance regions for the multivariate normal distribution. J. Am. Stat. Assoc., 61: 605-617.

Draper, N.R. and H.S. Smith, 1998. Applied Regression Analysis. 3rd Edn., Wiley and Sons, New York, USA.

Ellerton, R.R.W., C.P. Kitsos and S. Rinco, 1986. Choosing the optimal order of a response polynomical-structural approach with minimax criterion. Commun. Stat. Theory Meth., 15: 129-136.

Geisser, S., 1993. Predictive Inference: An Introduction. Chapman and Hall, London, ISBN: 0412034719, Pages: 264.

Gikas, V. and J. Stratakos, 2012. A novel geodetic engineering method for accurate and automated road/railway centerline geometry extraction based on the bearing diagram and fractal behavior. IEEE Trans. Intell. Transp. Syst., 13: 115-126.

Guttman, I., 1970a. Construction of β-content tolerance regions at confidence level β for large samples for k-Variate normal distribution. Ann. Math. Stat., 41: 376-400.

Guttmann, I., 1970b. Statistical Tolerance Region: Classical and Bayesian. Grifffin Ltd., London, ISBN: 0852641729, Pages: 150.

Hocking, R.R., 1976. The analysis of selection of variables in linear regression. Biometrics, 32: 1-49.

Hosmer, D.W. and P.R. Royston, 2003. Using fractional polynomials to model continuous covariates in regression analysis. http://www.umass. edu/statdata/statdata/data/ac.txt.

Kitsos, C.P., 1994. An Algorithm for Construct the Best Predictive Model. In: Softstat'93: Advances in Statistical Software, Faulbaum, F. (Eds.). Stuttgart, New York, pp: 535-539.

Maddala, G., 1992. Introduction to Econometrics. 2nd Edn., Macmillan, New York, USA., Pages: 663.

Mead, R., 1991. The Design of Experiments. Cambridge University Press, USA.

Mejza, S. and I. Mejza, 2012. Individual control treatments in designed agricultural experiments, COMPSTAT 2012. Proceedings of the 20th International Conference on Computational Statistics, August 27-31, 2012, Limassol, Cyprus.

Muller, C.H. and C.P. Kitsos, 2004. Optimal Design Criteria Based on Tolerance Regions. In: mODa 7-Advances in Model-Oriented Design and Analysis, Bucchianico, A., H. Lauter and H.P. Wynn (Eds.). Physica-Verlag, USA., pp: 107-115.

Oliveira, T. and A. Oliveira, 2012. Ineffectiveness at the FIM in selecting optimal BIB designs for testing block effects COMPSTAT 2012. Proceedings of the 20th International Conference on Computational Statistics, August 27-31, 2012, Limassol Cyprus.

Paulson, E., 1943. A note on tolerance limits. Ann. Math. Stat., 14: 90-93.

Pereira, D.G., P.C. Rodrigues, S. Mejza and J.T. Mexia, 2012. A comparison between joint regression analysis and the AMMI model: A case study with barley. J. Stat. Comput. Simul., 82: 193-207.

Pronzato, L. and E. Walter, 1994. Minimal-volume ellipsoids. Int. J. Adaptive Control Signal Process., 8: 15-30.

Seber, G.A.F., 1977. Linear Regression Analysis. John Wiley and Sons, New York.

Shao, J., 1993. Linear model selection by cross-validation. J. Amer. Statist. Assoc., 88: 486-494.

Stewart, J. and L. Gill, 1991. Econometrics. Prentice Hall Europe, USA.

Stigler, S.M., 1981. Gauss and the invention of least squares. Ann. Stat., 9: 465-474.

Urbain, J.D., 1989. Model selection criteria and granger causality tests: An empirical note. Econ. Lett., 29: 317-320.

Valente, V. and T.A. Oliveira, 2011. Hierarchical linear models: Review and applications. Proceedings of the 9th International Conference of Numerical Analysis and Applied Mathematics, September 19-25, 2011, Halkidiki, Greece, pp: 1549-1552.

Wilks, S.S., 1962. Mathematical Statistics. John Wiley and Sons, New York.

Yu, K. and A. Ally, 2009. Improving prediction intervals: Some elementary methods. Am. Stat., 63: 17-19.

Zarikas, V., V. Gikas and C.P. Kitsos, 2010. Evaluation of the optimal design `cosinor model` for enhancing the potential of robotic theodolite kinematic observations. Measurement, 43: 1416-1424.