Graphically Oriented System for Textile Processes Models Building.

Jiří Militký,
Dept. of Textile Materials, Textile Faculty
Technical University of Liberec
46117 Liberec, Czech Republic
E-mail: jiri.militky@vslib.cz

KEY WORDS
Linear regression, Models building, Prediction, Partial regression graphs

ABSTRACT

Multiple linear and nonlinear models building in textile branch belongs to the most complex problems solved in practice. Interactive approach to model building can be divided into the following steps (Meloun, Militky and Forina 1998):

i) Selection of provisional models,

ii) Analysis of assumptions about model, data and used regression methods (regression diagnostic),

iii) Extension and modification of model, data and regression method,

iv) Testing of model validity, their prediction capability, etc.

Some interactive strategy of multiple regression model building based on the above steps is described in book (Meloun, Militky and Forina 1998). Many problems in realization of step i) are caused by strong multicollinearity. Multicollinearity in multiple linear regression analyses is defined as approximate linear dependencies among the explanatory variables (columns of design matrix **X**). It is well known that under strong multicollinearity the individual scatter plots between response y and explanatory variables x_j cannot be used for model building. Models of textile processes are usually created by the classical methods of experimental design. This approach enabling the optimization of experimental conditions is formally very general but in practice often leads to the incorrect models containing often too parameters. In this contribution, the graphically oriented method of textile type models building will be presented. This method is based on the special projection enabling the investigation of partial dependence of response on the selected exploratory variable. The aim of graphical analysis is to evaluate the type of nonlinearities due to function of predictors describing well the experimental data. For selection of suitable model the characteristics based on the cross validation principle will be proposed. The program MULTIREG in MATLAB is mentioned. This methodology is demonstrated on the example of PET/cotton type yarns tenacity prediction.

INTRODUCTION

A regression type model building is a relatively specific discipline capable to solve a lot of practical problems in textile research. Classical tasks solved by the regression model building are:

•Description of dependence between fibers properties and properties of fibrous structures.

•Quantification of influence of process parameters on the structural parameters and properties of fabrics

•Prediction of directly non-measurable properties of textiles form some directly measurable ones (eg. hand or comfort prediction) called multiple calibration

•Optimization of technological processes based on the models of Taylor expansion type (experimental design approach)

In all above-mentioned cases the interdependencies are very complex. And therefore data based models with good predictive capability are useful. Data based multiple linear and nonlinear model building belongs generally to the most complex problems solved in practice. In many cases is not possible to construct the mathematical form of model based on the information about system under investigation. In these cases the interactive approach to regression type models building could be attractive. In the proposed strategy of regression models building, the graphically oriented methods for estimation of model correctness and identification of spurious data are selected. These methods are based on the special projections enabling the investigation of partial dependencies of response on the selected exploratory variable. Classical ones are partial regression graphs or partial residual graphs. Nonlinear or special patterns in these graphs can be used for extension of regression model and including nonlinear terms or interactions. For identification of spurious data the so-called LR graphs can be used as well. For evaluation of model quality the characteristics derived from predictive capability are used. Some statistical tools for realization of above-mentioned techniques are described in the book [1].

Proceedings 14th European Simulation Symposium
A. Verbraeck, W. Krug, eds. (c) SCS Europe BVBA, 2002

For realization of regression models building in practice is necessary to have software for simple and interactive data analysis by linear and nonlinear regression with extensions for above mentioned graphically oriented strategy of model building and evaluation of their quality. Example of this software type is ADSTAT, which was built on the ground of the author long time experience with regression modeling and teaching of this topic at technical university. Using of the same strategy the program MULTIREG in Matlab was created. The application of this program for graphically oriented strategy of model building is demonstrated on the example of prediction of PET/cotton type yarns tenacity from selected process and raw materials characteristics.

SUMMARY OF LINEAR REGRESSION

A *linear* regression model is a model which is formed by a linear combination of explanatory variables **x** or their functions, $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_m x_{m,i} + \varepsilon_i$, $i = 1, ..., n$, written in matrix notation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Vector $\boldsymbol{y}$ has dimensions (n H 1) and matrix $\boldsymbol{X}$ (n H m). Linear means linear according to model parameters. Therefore for linear models the following conditions are valid

$$g_j = \frac{\delta f(x,\beta)}{\delta \beta_j} = \text{constant},\quad \text{j} = 1, ..., \text{m}.$$

If for any parameter, β_j, the partial derivative is not a constant, we say that the regression model is *nonlinear*. For additive model of measurements errors the linear regression model has the form

$$y = \mathbf{X}\beta + \varepsilon \qquad (1)$$

In eqn. (1) the n x m matrix **X** contains the values of m explanatory (predictor) variables at each of n observations, **β** is the m x 1 vector of regression parameters and ε_i is n x 1 vector of experimental errors. The **y** is n x 1 vector of observed values of the dependent variable (response). Columns $\boldsymbol{x}_j$ i. e. individual explanatory variables define geometrically the m-dimensional co-ordinate system or the hyperplane L in n-dimensional Euclidean space E^n. The vector $\boldsymbol{y}$ usually does not have to lie in this hyperplane L. The least squares are the most frequently used method in regression analysis. For a linear regression, the parameter estimates $\boldsymbol{b}$ may be found by minimization of measure between the vector $\boldsymbol{y}$ and the hyperplane L. This is equivalent to finding the minimal length of the residual vector $e = y - y_P$, where $y_p = X\,b$ is the *predictor vector*. This is equivalent to requirement of minimal length of residual vector

$$\mathbf{e} = \mathbf{y} - \mathbf{y}_\mathrm{P}$$

In Euclidean space is the length of residual vector expressed as

$$d = \sqrt{\sum_{i=1}^{n} e_i^2}$$

Geometry of linear least squares is shown on fig. 1.

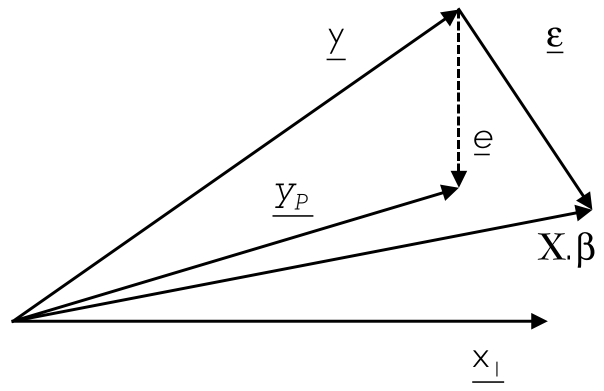

Figure 1: Geometry of linear least squares

The classical least squares method is based on the following assumptions:

1. regression parameters **β** are not restricted,

2. regression model is linear in parameters and additive model of measurements is valid(see eqn (1)),

3. design matrix **X** has a rank equal to n,

4. errors ε_i are i.i.d. random variables with zero mean $E(\varepsilon_i)=0$ and diagonal covariance matrix $D(\boldsymbol{\varepsilon})=\sigma^2$ **E,** where $\sigma^2 < \infty$

For testing purposes it is assumed that errors ε_i have normal distribution $N(0, \sigma^2)$. When these four assumptions are valid the parameter estimates **b** found by minimization of least squares criterion

$$S(\mathbf{b}) = \sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{m} x_{ij} b_j\right]^2 \qquad (2)$$

are called as best linear unbiased estimators (BLUE). The conventional least squares estimator **b** has the form

$$\mathbf{b} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{y} \qquad (3)$$

where symbol $\mathbf{A}^{-1}$ denotes inversion of matrix **A**. The term *best* estimates $\boldsymbol{b}$ means that any linear combination of these estimates has the smallest variance of all linear unbiased estimates. That is, the variance of the individual estimates $D(b_j)$ are the smallest from all possible linear unbiased estimates (Gauss-Markov theorem). The term *linear* estimates means that they can be written as a linear combination of measurements $\boldsymbol{y}$ with weights Q_{ij} which depend only on the location of variables $\boldsymbol{x}_j$, $j = 1, ..., m$, and $\boldsymbol{Q} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$ for the weight matrix, we can then say $b_j = \sum_{i=1}^{n} Q_{ij}\, y_i$. Each estimate b_j is the weighted sum of all measurements. Also, the estimates $\boldsymbol{b}$ have an asymptotic multivariate normal distribution with covariance matrix $D(\mathbf{b}) = \sigma^2 (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}$. The term *unbiased* estimates means that $E(\boldsymbol{\beta} - \boldsymbol{b}) = 0$ and the mean value of an estimate vector $E(\boldsymbol{b})$ is equal to a vector of regression parameters **β**. It should be noted that there exist *biased estimates*, the variance of which can be smaller than the variance of estimates $D(b_j)$.

The perpendicular projection of *y* into hyperplane L can be made using projection matrix ***H*** and may be expressed as

$$\mathbf{y}_P = \mathrm{X}\,\mathbf{b} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathbf{y} \quad (4)$$

where **H** is projection matrix. Residual vector **e** = **y** - $\mathbf{y}_P$ is orthogonal to subspace L and has the minimal length. Variance matrix corresponding to prediction vector $\mathbf{y}_P$ has the form D($\mathbf{y}_P$) = σ^2 **H** and variance matrix for residuals is D(**e**) =σ^2 (**E** - **H**). Residual sum of squares has the form $\mathbf{RSC} = S(\mathbf{b}) = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T(\mathbf{E}-\mathbf{H})\mathbf{y} = \mathbf{y}^T\mathbf{P}\mathbf{y}$, and its mean value is $\mathrm{E}(\mathbf{RSC}) = \sigma^2(\mathrm{n} - \mathrm{m})$. Unbiased estimator of measurements variance σ^2 is equal to

$$s^2 = \frac{S(\mathbf{b})}{n-m} = \frac{\mathbf{e}^T\mathbf{e}}{n-m}$$

Statistical analysis related to least squares is based on normality of estimates **b**. Quality of regression is often (not quite correctly) described by the multiple correlation coefficient R defined by relation

$$R^2 = 1 - \frac{RSC}{\sum(y_i - \sum y_i / n)^2} \qquad (5)$$

For model building the multiple correlation coefficient is not suitable. It is non-decreasing function of number of predictors and therefore the over-parameterized model results. Prediction ability of regression model can be characterized by quadratic error of prediction (MEP) defined for linear models by relation

$$MEP = \sum_{i=1}^{n}(y_i - x_i^T b_{(i)})^2 / n \qquad (6)$$

Here $\mathbf{b}_{(i)}$ is the estimate of regression model parameters when all points except the i-th are used. The statistics MEP for linear models uses the prediction $y_{Pi} = \mathbf{x}_i^T\, \mathbf{b}_{(i)}$ which was constructed without the information about the i-th point. The estimate $\mathbf{b}_{(i)}$ can be computed from least squares estimate **b**

$$\mathbf{b}_{(i)} = \mathbf{b} - [(\mathbf{X}^T\mathbf{X})^{-1}\, \mathbf{x}_i\, e_i] / [1 - \mathrm{H}_{ii}] \qquad (7)$$

Here H_{ii} is a diagonal element of projection matrix **H**. Optimal model has minimal value of MEP. The MEP can be used for definition of the predicted multiple correlation coefficient PR (Meloun, Militky and Forina 1998). The PR is attractive especially for empirical model building. Analysis of various types of the regression residuals, or some transformation of the residuals, is very useful for detecting inadequacies in the model or problems in data. The true errors in the regression model are assumed to be normally and independently distributed random variables with zero mean and common (i. e. constant) variance ε . N(**0**, **I**σ^2).

(a) Classical residuals e_i are defined by the expression $e_i = y_i - x_i b$, where x_i is the *i*th row of matrix ***X***. Classical analysis is based on the wrong assumption that residuals are good estimates of errors ε_i. Reality is more complex, the residuals e are a projection of vector ***y*** into a subspace of dimension (*n* - *m*),

$$e = P\,y = P\,(X\,\beta + \varepsilon) = P\,\varepsilon = (\,E - H\,)\,\varepsilon$$

and therefore, for the *i*th residual it is valid

$$e_i = (\,1 - H_{ii}\,)\,y_i - \sum_{j \neq i}^{n} H_{ij}\,y_j = (\,1 - H_{ii}\,)\,\varepsilon_i - \sum_{j \neq i}^{n} H_{ij}\,\varepsilon_j$$

Each residual $\hat{e}_i$ is a linear combination of all errors ε_i. The distribution of residuals depends on

(a) the error distribution,
(b) the elements of the projection matrix ***H***,
(c) the sample size *n*.

Because the residual e_i represents a sum of random quantities with bounded variance, the *supernormality effect* appears for small sample size: Even when the errors **ε** do not have a normal distribution, the distribution of residuals is close to normal. In small samples, the elements of the projection matrix ***H*** are larger and the main role of an actual point is to influence the sum of terms $H_{ii}\ \varepsilon_i$. The distribution of this sum is closer to a normal one than the distribution of errors ε. For large sample size, where 1/*n* is approaching to 0, we find that $e_i \rightarrow \varepsilon_i$ and analysis of the residual distribution gives direct information about the distribution of errors. Classical residuals are always associated with non-constant variance; they sum to be more normal and may not indicate strongly deviant points. The common practice is to use residuals for investigation of model quality and for identification of nonlinearities As it has been shown above for small and moderate sample sizes the classical residuals are not good for diagnostic or identification of models quality.

(b) Normalized residuals $e_{N,i} = e_i / s$ are often recommended in practice. It is falsely assumed that these residuals are normally distributed quantities with zero mean and variance equal to one. In reality these residuals have non-constant variance. When normalized residuals are used, the rule of 3σ is classically recommended: "quantities with $e_{N,i}$ of magnitude greater than 3σ are classified as the outliers". This approach is quite misleading and may cause wrong decision about data.

(c) Standardized residuals $e_{S,i} = e_i / (s\sqrt{1 - H_{ii}})$ exhibit constant unit variance and their statistical properties are the same as those of classical residuals. Here H_{ii} is the *i*th diagonal element of ***H*** matrix. The standardized residuals behave much like a Student random variable except for the fact that the numerator and denominator of $e_{S,i}$ are not independent.

(d) Jackknife residuals $e_{J,i} = e_{S,i}\sqrt{\dfrac{n - m - 1}{n - m - e_{S,i}^2}}$, are residuals which with an assumption of normality of errors, have the Student distribution with (*n* - *m* - *1*) degrees of freedom. The principle is standardization each residual with an estimate of its standard deviation that is independent of the residual. This is accomplished by using, as the estimate of σ^2 for the *i*th residual, the residual mean square from an analysis where that observation has been omitted. This variance is labeled $s^2_{(i)}$, where the subscript in parentheses indicates that the *i*th observation has been omitted for the estimate of σ^2. The result is *jackknife residual* or also called the *fully Studentized residual*. It is distributed as Student *t* with (*n* - *m* - *1*) degrees of freedom when normality of

errors ε holds. As with e_i and $e_{S,i}$, residuals $e_{J,i}$ are not independent of each other. The $e_{J,i}$ are called the cross-validatory or jaccknife residuals (Atkinson 1985) and are often used for identification of outliers

GRAPHICAL AIDS FOR MODEL CREATION

In multiple regression one usually starts with assumption that response y is linearly related to each of predictors. The aim of graphical analysis is to evaluate the type of non-linearity due to function of predictors describing well the experimental data. The power type function of predictors is suitable when relation is monotone. Several diagnostic plots have been proposed for detection of curve between y and x_j (Berk and Booth 1995, Atkinson 1985). Very useful for designed experiments without marked collinearities is partial regression plot (PRL). This plot uses the residuals from the regression of y on the predictor x_j, graphed against the residuals from the regression of x_j on the other predictors. This graph is now the standard part of modern statistical packages and can be constructed without recalculating of least squares. To discuss the properties of this plot type let assume the regression model in the matrix notation

$$\mathbf{y} = \mathbf{X}_{(j)}\,\boldsymbol{\beta}^* + \mathbf{x}_j\, c + \varepsilon_i \qquad (8)$$

Here $\mathbf{X}_{(j)}$ is matrix formed by leaving out the j-th column $\mathbf{x}_j$ from matrix $\mathbf{X}$, $\boldsymbol{\beta}^*$is (n-1) x 1 parameter vector and c is regression parameter corresponding to the j-th variable $\mathbf{x}_j$. For the investigation of partial linearity between y and j-th variable $\mathbf{x}_j$ the projection into subspace L orthogonal to space defined by columns of matrix $\mathbf{X}_{(j)}$ is used. Corresponding projection matrix into space L has the form $\mathbf{P}_{(j)} = \mathbf{E} - \mathbf{X}_{(j)}\,(\mathbf{X}_{(j)}^T\,\mathbf{X}_{(j)})^{-1}\,\mathbf{X}_{(j)}^T$.
Using this projection to the both sides of eqn.(8) the following relation results

$$\mathbf{P}_{(j)}\,\mathbf{y} = \mathbf{P}_{(j)}\,\mathbf{x}_j\, c + \mathbf{P}_{(j)}\,\boldsymbol{\varepsilon} \qquad (9)$$

The product $\mathbf{P}_{(j)}\,\mathbf{X}_{(j)}\,\boldsymbol{\beta}^*$ is equal to zero because the space spanned by $\mathbf{X}_{(j)}$ is orthogonal to the residuals space. It is clear that the term $\mathbf{v}_j = \mathbf{P}_{(j)}\,\mathbf{x}_j$ is the residual vector of regression of variable $\mathbf{x}_j$ on the other variables which form columns of the matrix $\mathbf{X}_{(j)}$ and the term $\mathbf{u}_j = \mathbf{P}_{(j)}\,\mathbf{y}$ is the residual vector of regression of variable $\mathbf{y}$ on the other variables which form columns of the matrix $\mathbf{X}_{(j).}$ The partial regression graph is then dependence of vector $\mathbf{u}_j$ on vector $\mathbf{v}_j$. If the term $\mathbf{x}_j$ is correctly specified the partial regression graph forms straight line. Systematic nonlinearity is indication of incorrect specification of $\mathbf{x}_j$. Random pattern shows unimportance of $\mathbf{x}_j$ for explaining the variability of $\mathbf{y}$. The partial regression graph (PRL) has the following properties:

1. The slope c in PRL is identical with estimate b_j in a full model.

2. The correlation coefficient in PRL is equal to the partial correlation coefficient R_{yxj}.

3. Residuals in PRL are identical with residuals for full model.

4. The influential points, nonlinearities and violations of least squares assumptions are markedly visualized.

Therefore the PRL are useful for inspection of data quality and model quality as well. Form nonlinearities in PRL graphs the proper transformation or inclusion of nonlinear functions of explanatory variables cane be deduced.

MATLAB PROGRAM

The program MULTIREG serves for creation of linear and linearized regression models, estimation of their parameters and corresponding statistical analysis. For linear regression, special algorithms have been implemented. The least squares method is the only special case among a series of biased parameter estimation, controlled by a single parameter. Before computations, the data can be transformed to a polynomial form, the Taylor expansion (up to quadratic terms) or generally (any variable is transformed by a user function). A matrix left division is used for solving of over determined system of equations

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdot & x_{1m} \\ x_{21} & x_{22} & \cdot & x_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & x_{n2} & \cdot & x_{nm} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_m \end{bmatrix}$$

in the least squares sense. Variety of regression characteristics including partial regression graphs is computed. The PRL is created very simply from matrix $\mathbf{P}_{(j)}$ created by using of matrix oriented expressions in MAATLAB For diagnostic purposes, the plenty of graphs for proving the assumptions about data, models and least squares criterion have been included (Meloun, Militky and Forina 1998).

EXAMPLE

The main aim is the description of PET/cotton yarn tenacity (response) in dependence on the following parameters (rotor diameter (x_1), rotor speed (x_3), yarn fineness (x_2) ,PET fibers length (x_5) and fineness (x_4)). The n = 189 experimental points were specified Details of yarn creation and tenacity measurements are given in thesis (El Shahat 1994). For the purpose of right setting of rotor machine and selection of cotton fibers is necessary to known the dependence of tenacity on the above-mentioned explanatory variables especially from point of view of trends. In the first run by using of MULTIREG program the linear regression model was created. The partial regression graph for rotor speed is in Fig 2.

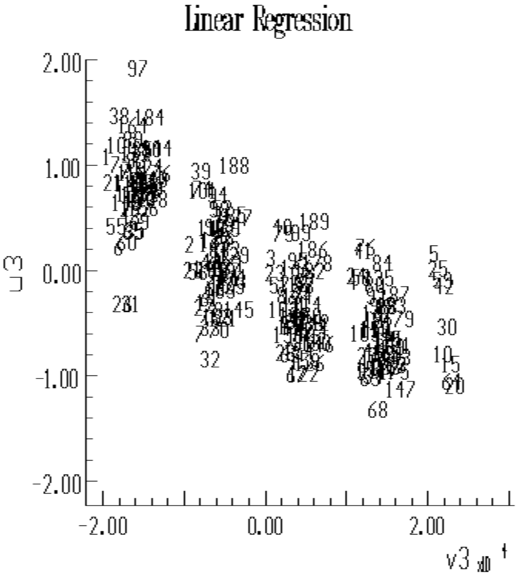

Figure 2: Partial regression graph for rotor speed

Formally, the linear regression model is fully acceptable. All regression parameters are significant on the significance level 0.05 and multiple correlation coefficient is equal to R= 0.9425. Other statistical characteristics are:

Predicted correlation coefficient equal to PR= 0.93862
Mean quadratic error of prediction MEP =0.19738.

It is known that the increase of rotor speed leads to increase of tenacity. The negative sign of coefficient for rotor speed variable is therefore not acceptable from the point of view of the textile interpretation.. From partial regression graphs is clear that there are some nonlinearities in variables x_2 and x_3 mainly. In the second run the full quadratic model equivalent to the Taylor expansion of the unknown function to the quadratic terms was used. Partial regression graph for rotor speed is shown on the fig. 3.

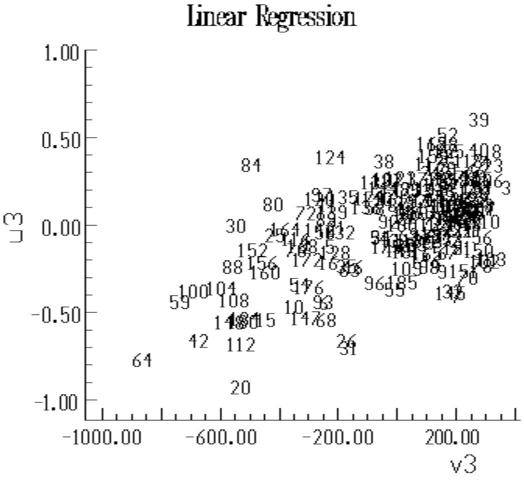

Fig 3 Partial regression graph for rotor speed

It is clear that x_3 (rotor speed, see Fig. 3) is now significantly linear and has the right positive sign. The significance of the other variables is hidden in the interactions or quadratic terms (see table 1). In the table 1 there are only significant terms of the full quadratic model computed by least squares.

Table 1: Significant Parameters for Full Quadratic Model

Parameter	Variable	Estimate	σ Estimated
B[2]	x_2	1.1017E+00	1.2989E-01
B[3]	x_3	5.8685E-04	6.6282E-05
B[6]	$x_1\ x_2$	-7.8860E-03	2.8638E-03
B[7]	$x_1\ x_3$	-1.1992E-05	1.2090E-06
B[9]	$x_1\ x_5$	-1.3611E-02	5.4958E-03
B[10]	$x_2\ x_3$	-9.8261E-07	3.4311E-07
B[17]	x_2^2	-1.3584E-02	1.0781E-03
B[18]	x_3^2	-1.1339E-09	1.3846E-10
B[19]	x_4^2	-5.1025E+00	8.4249E-01
B[20]	x_5^2	-1.6550E-02	.6391E-03

The multiple correlation coefficient is R= 0.98455. Other statistical characteristics are:

Predicted correlation coefficient PR= 0.98451
Mean quadratic error of prediction MEP =0.05097.

Quadratic model has here therefore better predictive ability than linear one. Contributions of individual terms exhibit no nonlinear trends (see PRL graphs). The right interpretation of this model is now without problems.

CONCLUSION

The utilization of partial regression graphs and suitable criterion expressing the predictive ability is very useful for building of statistical models especially based on the experimental design arrangements. The MATLAB program MULTIREG is useful for interactive building of empirical regression models. These nonlinear models are often very simple and are attractive especially for selection of optimal technological processes conditions.

ACKNOWLEDGEMENT

This work was supported by the research project “Textile center” of Czech Ministry of Education LN00B090

REFERENCES

Atkinson A.1985. *Plots, Transformations and Regression*, Claredon Press, Oxford,

Berk K., Booth D.E. 1995 “*Seeing a curve in Multiple Regression*” Technometrics 37, 385 -396

El Shahat I. 1994 “*OE yarns tenacity prediction*”, Thesis, Mansoura University

Meloun M., Militký J.and Forina M. 1994. *Chemometrics in Analytical Chemistry vol. II, Interactive Model Building and Testing on IBM PC*, Ellis Horwood, Chichester