Data based model creation in cotton research

J. Militký
Department of Textile Materials, Textile Faculty, Technical University of Liberec, Liberec
CZECH REPUBLIC
Correspondence author jiri.militky@vslib.cz

ABSTRACT

The problems connected with application of sophisticated statistical analysis of HVI data are presented. The graphically oriented methods of prediction type models building are described. First method is based on the special projection enabling the investigation of partial dependence of response on the selected exploratory variable (partial regression graphs). For selection of predictive model the mean error of prediction MEP and predictive correlation coefficient PR are proposed Second method use PCA for definition of new orthogonal set of variables. Regression models constructed from these new variables are structurally simpler and the non-linearities are more visible. The re analyzing of published data about HVI shows that proposed approach leads to the model with better predictability and simpler structure in comparison with classical regression approach.

Introduction

Data based models building is a relatively specific discipline capable to solve a lot of practical problems in cotton research. Classical tasks solved these techniques are:

- Description of dependence between cotton fiber properties and properties of fibrous structures.
- Creation of calibration models for HVI instruments.
- Quantification of influence of cultivation parameters on the structural parameters of cotton fibers.
- Prediction of directly non-measurable properties of cotton fabrics form some directly measurable ones (e.g. hand or comfort prediction).

In all above-mentioned cases the interdependencies between input and output variables are very complex and therefore data based models with good predictive capability are wanted. Due to limited range of some cotton fiber characteristics and strong multicollinearity is necessary to use careful statistical analysis for obtaining reliable results (Meloun and Militky, 1994).

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 create the mathematical form of model from information about system under investigation. In these cases the interactive approach to data based models building could be attractive.

The main aim of this contribution is proposal of regression models building strategy based on the graphically oriented methods for estimation of model correctness and identification of spurious data. Proposed methods are based on the special projections enabling the investigation of partial dependencies of response on the selected exploratory variable or creation of new orthogonal set of variables from principal component analysis (PCA). This strategy will be demonstrated on the prediction of skein break of cotton yarns from HVI data. This approach will be compared with nonparametric models based on the neural network (radial basis functions).

Experimental data

The enormous number of methods and their modifications has been proposed for characterization of cotton fibers. Traditional are fineness (expressed as micronaire value - MI or directly as Tex), length (expressed as mean length - LM), maturity (expressed as maturity index - IM from polarization microscope) and strength of individual fibers (SI) or bundle strength (expressed by Pressley index - PI). Classical low volume instruments (LVI) were for practical applications replaced by HVI (high volume instrument) testing systems, which are capable to provide rapid and cheap measurements of the basic cotton fiber properties as well. Fiber properties measured for example on Spinlab 900 include fineness (MI), length (UHM), length uniformity (UF), elongation (E), strength (S) and reflectance (RD). By using of the HVI is possible to obtain a large amount of fiber data. The main aims of interpretation of these data are:

- Prediction of influence of fiber properties on properties (strength) of yarns.
- Expression of fiber strength from fiber geometric characteristics and maturity.

Due to lack of theoretical knowledge, limited range of some fiber characteristics and their strong interdependence (multicollinearity) is necessary to use careful statistical analysis for obtaining reliable regression type predictive models (equation 1). For construction of predictive models the three main problems should be solved:

a) Inspection of data and screening of variables. Result is selection of important variables and possible discarding of spurious data.
b) Selection of the form of the regression model (linear, nonlinear, interaction). The graphically oriented methods are useful.
c) Avoiding multicollinearity due to strong mutual correlation between the fiber characteristics, and presence of highly influential points. The utilization of new orthogonal variables leads to simplification of problem.
d) Specification of the criterion for selection of the best predictive model.

It is shown below that for solving of the problems a) the PCA is suitable and for problems a) and b) the so-called partial regression graphs are very attractive. As a criterion for selection of predictive regression model the mean error of prediction MEP and predictive correlation coefficient PR are chosen.

The HVI data from El Moghahy (1989) are here used for prediction of the yarn strength (expressed as a skein break factor SB). The main goal is to create predictive regression model for prediction of SB from HVI data.

Predictive model building

Interactive approach to predictive model building can be divided into the following steps described by Meloun and Militky (1994):

a) Preliminary data analysis.
b) Selection of provisional models.
c) Analysis of assumptions about model, data and used regression methods (regression diagnostic).
d) Extension and modification of model, data and regression method.
e) Testing model validity, their prediction capability, etc.

Some interactive strategy of multiple model building based on the above steps is described in book Meloun and Militky (1994). 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_i cannot be used for model building.

Preliminary data analysis

The main aim of this analysis is identification of special features of data as non-random patterns, outliers, bad variables and dependencies. In the multivariate case is the most popular to use principal component analysis (PCA). It allows projection of multidimensional data onto few orthogonal features called principal components. These principal components are constructed as linear combination of original variables to maximize data variance. The input to PCA is data matrix **X** (Nxp) having N rows (samples) and p columns (variables). Output from PCA is matrix **Z** (Nxp), having transformed sample values N into p principal components. Columns of this matrix are mutually uncorrelated (orthogonal). Let is matrix **X** column centered. The output matrix columns are linear combinations of input matrix columns i.e. **Z = G*X.** Matrix **G** is orthogonal. Columns of matrix **Z** are called scores of principal components or principal axes. Because is matrix $\mathbf{X^TX}$ practically the covariance matrix it is valid that $\mathbf{Z^TZ=G^TX^TXG=L}$. Columns of matrix **G** are eigenvectors and diagonal matrix **L** contains eigenvalues of matrix $\mathbf{X^TX}$. Therefore the principal components can be obtained from eigenvectors of covariance matrix. They are a set of orthogonal axes to which data points may be referred and account for the data variance in a decreasing order of importance. The key property of PCA is that it attains the best linear map by minimizing the least squared errors of data reconstruction. The typical output from PCA is so called component graph. It is two-dimensional subspace spanned by first two principal axes. Digits indicate multivariate points and vectors the variables projections. Simple tool for summing of pair-wise correlation between variables including response variable is matrix plot. In this plot are correlations characterized by gray level.

Linear regression models

In sequel the standard linear model with n observations of m explanatory variables is assumed. For additive model of measurements errors the linear regression model has the form

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

In equation (1) the n x m matrix X contains the values of m explanatory (predictor) variables at each of n observations, b is the m x 1 vector of regression parameters and e is an n x 1 vector of experimental errors. The y is n x 1 vector of observed values of the dependent variable (response). The classical least squares is based on the following assumptions:

1. Regression parameters b *are not restricted,*
2. Regression model is linear in parameters and additive model of measurements is valid (equation 1).
3. Design matrix **X** has a rank equal to n,
4. Errors e_i are i.i.d. random variables with zero mean $E(e_i)=0$ and diagonal covariance matrix $D(e)= \sigma^2 \mathbf{E}$ where $\sigma^2 < \infty$.

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

$$\boxed{S(\mathbf{b}) = \| \mathbf{y} - \mathbf{X}\,\mathbf{b} \|_2} \qquad (2)$$

are best linear unbiased estimators (BLUE). In equation (2) the $\| \, . \, \|$ is symbol for Euclidean norm. The conventional least squares estimator **b** has the form

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

The corresponding covariance

$$\boxed{D(\mathbf{b}) = \sigma^2\ (\mathbf{X}^T\mathbf{X})^{-1}} \qquad (4)$$

From geometrical point of view columns of design matrix **X** define m-dimensional hyper-plane L in n-dimensional Euclidean space E^n. The vector **X** b and prediction vector

$$\boxed{\mathbf{y}_P = \mathbf{X}\,\mathbf{b}} \qquad (5)$$

lie in plane L. The prediction vector is orthogonal projection of vector **y** to the plane L.

$$\boxed{\mathbf{y}_P = \mathbf{H}\,\mathbf{y} = \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y}} \qquad (6)$$

where **H** is projection matrix. Residual vector

$$\mathbf{e} = \mathbf{y} - \mathbf{y}_P \qquad (7)$$

is orthogonal to plane L and has the minimal length. Vector **e** is related to projection matrix **H**

$$\mathbf{e} = (\mathbf{E} - \mathbf{H})\,\mathbf{y} \qquad (8)$$

Symbol **E** denotes unit matrix of order n. Variance matrix corresponding to prediction vector $\mathbf{y}_P$ has the form

$$D(\mathbf{y}_P) = \sigma^2\,\mathbf{H} \qquad (9)$$

and variance matrix for residuals is

$$D(\mathbf{e}) = \sigma^2\,(\mathbf{E} - \mathbf{H}) \qquad (10)$$

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

$$R^2 = 1 - S(\mathbf{b})/\sum_{(i)}(y_i - \sum_{(i)} y_i/n)^2 \qquad (11)$$

For model building the multiple correlation coefficient is not suitable. It is non-decreasing function of number of predictors and therefore the over-defined model results.

Graphs 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 nonlinearities due to function of predictors describing well the experimental data. A power type function of predictors is suitable when relation is monotone. Several diagnostic plot have been proposed for detection of curve between y and x_j. Some are described in the work of Berk and Booth (1995). Very useful for designed experiments without marked collinearities is partial regression plot. 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 standard part of modern statistical packages and can be constructed without recalculation of least squares. To discuss the properties of this type plot let we assume the regression model in the matrix notation

$$\mathbf{y} = \mathbf{X}_{(j)}\,\boldsymbol{\beta}^* + \mathbf{x}_j\,c + \boldsymbol{\varepsilon} \qquad (12)$$

where $\mathbf{X}_{(j)}$ is matrix formed by leaving out the j-th column $\mathbf{x}_j$ from matrix **X**, ß*is (n-1) x 1 parameter vector and c is regression parameter corresponding the j-th variable $\mathbf{x}_j$.

For investigation of partial linearity between y and j-th variable $\mathbf{x}_j$ the projection into space L orthogonal to space defined by columns of matrix $\mathbf{X}_{(j)}$ is used. Corresponding projection matrix into space L has the form

$$\mathbf{P}_{(j)} = \mathbf{E} - \mathbf{X}_{(j)}\,(\mathbf{X}_{(j)}{}^{\mathrm{T}}\,\mathbf{X}_{(j)})^{-1}\,\mathbf{X}_{(j)}{}^{\mathrm{T}} \qquad (13)$$

By using the projection $\mathbf{P}_{(j)}$ onto both sides of equation.(12) the following relation results

$$\mathbf{P}_{(j)}\,\mathbf{y} = \mathbf{P}_{(j)}\,\mathbf{x}_j\,c + \mathbf{P}_{(j)}\,\boldsymbol{\varepsilon} \qquad (14)$$

The product $\mathbf{P}_{(j)}\,\mathbf{X}_{(j)}\,b^*$ is equal to zero because the space spanned by $\mathbf{X}_{(j)}$ is orthogonal to residuals space.

From equation. (14) it follows 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 $\mathbf{X}_{(j)}$
- The term $\mathbf{u}_j = \mathbf{P}_{(j)}\,\mathbf{y}$ is the residual vector of regression of variable **y** on the other variables which form columns of the matrix $\mathbf{X}_{(j)}$

Partial regression graph is then dependence of vector $\mathbf{u}_j$ on the 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 $\mathbf{x}_j$ and random pattern shows unimportance of $\mathbf{x}_j$ for explaining the variability of **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 and intercept is equal to zero.
2. The correlation coefficient in PRL is equal to the partial correlation coefficient R_{yxj}.
3. Residuals corresponding to straight line in PRL are identical with residuals for a full model.
4. The influential points, nonlinearities and violations of least squares assumptions are markedly visualized.

Using of so-called catch matrix facilitates the construction of PLR. In some cases is useful to add low degree polynomial regression model into this plot.

One of main properties of regression models is the suitable prediction ability. This prediction ability can be adopted also for selection of optimal model. Prediction ability in linear regression model can be characterized by mean quadratic error of prediction (MEP) defined generally by relation:

$$MEP = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T b_{(i)})^2 / n \qquad (15)$$

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

$$\mathbf{b}_{(i)} = \mathbf{b} - [(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{x}_i\,e_i] / [\,1 - H_{ii}\,] \qquad (16)$$

where H_{ii} is diagonal element of projection matrix **H**. After substitution from equation (15) to the equation (12) the simple relation results

$$MEP = \frac{1}{n}\sum_{i=1}^{n} e_i^2\,/(1 - H_{ii})^2 \quad \mathrm{MEP} = \mathrm{n}^{-1}\,\Sigma\,\mathrm{e_i}^2/\,(1 - \mathrm{H_{ii}})^2 \qquad (17)$$

For selected model is possible to compute values

of H_{ii} *from equation (6) and then the MEP criterion from equation (17).*

Optimal model has minimal value of MEP. The MEP can be used for definition of the predicted multiple correlation coefficient PR

$$PR^2 = 1 - n * MEP / \sum_i (y_i - \sum_i y_i / n)^2 \quad (18)$$

The PR is attractive especially for empirical model building. It is closely connected with well-known method of cross validation or single leave out statistics.

Predictive model for yarn strength

Data for computation of predictive model describing the dependence of fiber yarn skein break factor y (SBF) on the fiber fineness X_1, length X_2, unevenness X_3, strength X_4, elongation X_5 and reflectance X_6 were published by El Moghazhy (1989). Because the data set is very small (17 points only) is analysis concentrated on the creation of model without detail inspection of data. The mutual dependences between explanatory variables (X_1,..X_6) and response (variable No. 7) are graphically summarized in correlation map shown on Figure 1. Level of correlation in this map corresponds to the level of gray scale. Perfect correlation is marked by white and no correlation is marked by black color. Correlation map shows relative high correlation between variable No2 and No3 with the rest of variables. Therefore the fiber length and strength correlates with the rest of fiber properties and yarn skein break factor. Column centered data projected to the space of first two principal components are given on the Figure 2. The vectors correspond to the individual variables projections. Small directional differences between vectors indicate the similarity between variables. It is visible that there are close connections between variables pairs No. 3-No.4 and No. 5-No.6. The first three singular values and loadings are given in the Table 1. The magnitude of each eigenvalue is related to fraction og of data variance explained by corresponding principal component. By inspection of loadings for first principal component it can be assumed that the important contributions to building of this component have variables No. 4 and No. 5 mainly. Because the principal components are mutually orthogonal the scatter plots of response on principal component reveal corresponding dependences. Based on the inspection of all scatter graphs the only first principal component has marked trend as is shown on Figure 3. On the other scatter graph was the trend hidden in the noise component. The detailed investigation shows that the nonlinear trend is created by the two various data clusters. Principal components are here not suitable for regression model building but still are useful for inspection of structures in data. Results of the linear least squares for the full linear regression model containing all variables are given in Tables 2 to 3. Partial regression graphs show that the partial dependencies for X1 (fineness) and X2 (length) are not significant. Therefore in the regression model were included only explanatory variables unevenness X3, strength X4, elongation X5 and reflectance X6. Partial regression graphs for variable No 6 is on Figure 4, for variable No 5 is on Figure 5 and for variable No 3 is on Figure 6. It is visible that the small nonlinearities are due to the high data scattering and the linear regression can be used for final model as well. Results of the linear least squares for this reduced model are given in the Tables 4 and 5. It is clear that this reduced model has better predictive ability in comparison wit full linear model or model selected by Mogahzy (1989). No marked multicollinearity has been detected and no influential points are presented.

Conclusion

It is clear that utilization of the partial regression graphs is very useful for creation of predictive type models from cotton fiber data. The MEP criterion can be used for selection of optimal sub-model. Above described methodology can be used for creation of predictive type regression models describing influence of fiber parameters on the parameters of yarn or fabrics or for expressing cotton fiber quality as well.

Acknowledgement

The Czech Ministry of Education Grant LN B090 supported this work

References

- Berk, K.N., Booth, D.E. (1995). Seeing a curve in Multiple Regression. *Technometrics*, **37**: 385.
- El Mogahzy, E. and Broughton, R.M. (1989). Diagnostic procedures for multicollinearity between HVI cotton fiber properties, *Text. Res. J.*, **59**: 440.
- 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.

Table 1. *Singular values and loadings for first three principal components.*

j	L_j	G_{1j}	G_{2j}	G_{3j}	G_{4j}	G_{5j}	G_{6j}
1	289.2729	-0.0613	-0.0319	-0.0275	-0.6756	-0.7328	0.0339
2	91.2887	-0.0019	-0.0046	0.0251	-0.0216	-0.0268	-0.9991
3	13.0395	0.1755	0.4161	-0.8905	-0.0366	0.0333	-0.0248

Table 2. *Parameter estimates.*

Parameter	Estimate	Standard deviation	Confidence level
B[0]	-6.267E+03	1.39E+03	0.001
B[1]	-1.127E+02	6.52E+01	0.115
B[2]	3.107E+02	4.41E+02	0.498
B[3]	6.661E+01	1.15E+01	0.000
B[4]	2.423E+01	7.25E+00	0.007
B[5]	1.238E+02	4.62E+01	0.023
B[6]	1.862E+01	3.01E+00	0.000

Table 3. *Statistical characteristics of regression.*

Multiple correlation coefficient, R	9.7198E-01
Determination coefficient, R2 :	9.4475E-01
Predicted correlation coefficient, PR 2 :	9.3197E-01
Mean quadratic error of prediction, MEP :	7.2693E+03

Table 4. *Parameter estimates for reduced model.*

Parameter	Estimate	Standard deviation	Confidence level
B[0]	-7.30E+03	1.00E+03	0.0001
B[3]	7.72E+01	1.71E+01	0.001
B[4]	2.00E+01	9.29E+00	0.052
B[5]	1.27E+02	5.39E+01	0.036
B[6]	2.20E+01	3.77E+00	0.0001

Table 5. *Statistical characteristics of regression for reduced model.*

Multiple correlation coefficient, R	9.6623E-01
Determination coefficient, R2 :	9.3361E-01
Predicted correlation coefficient, PR 2 :	9.4173E-01
Mean quadratic error of prediction, MEP :	6.2577E+03

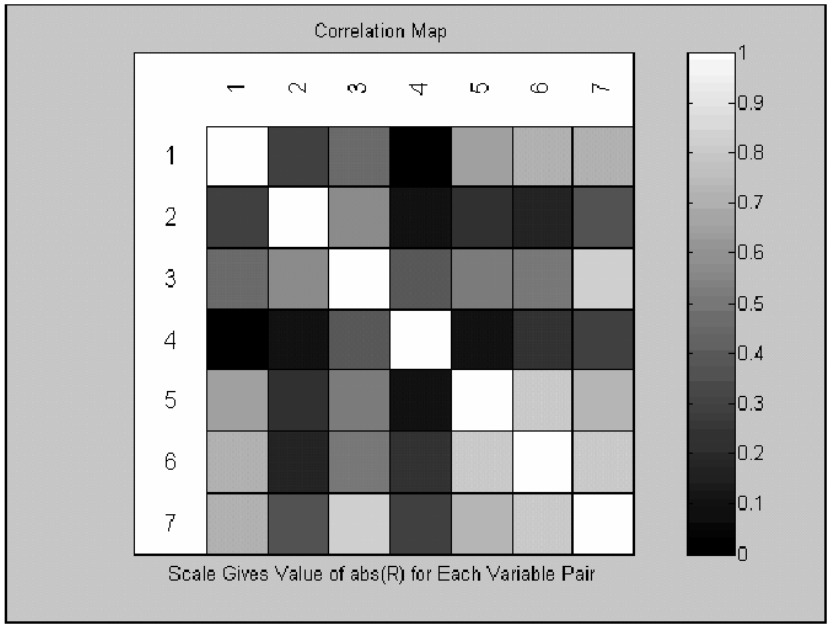

Figure 1.
Correlation map.

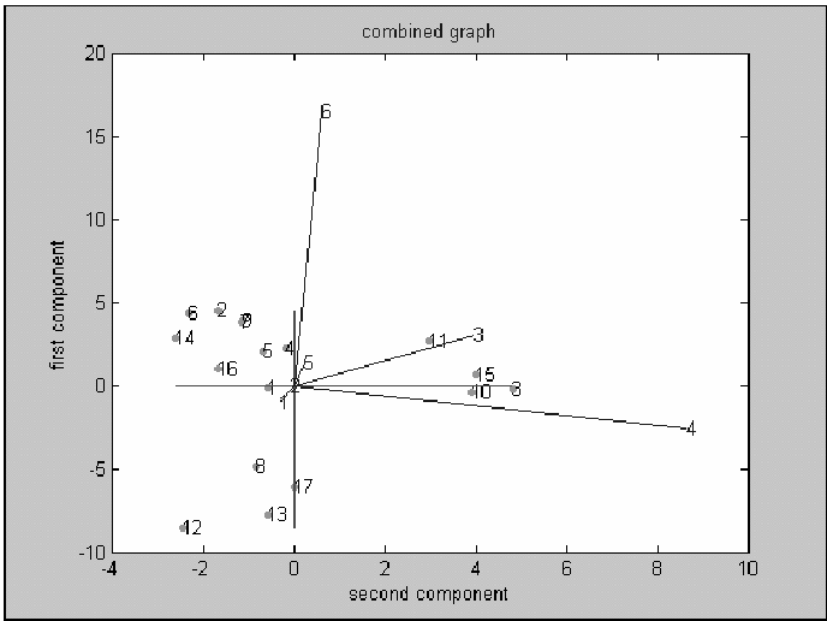

Figure 2.
Combined graph.

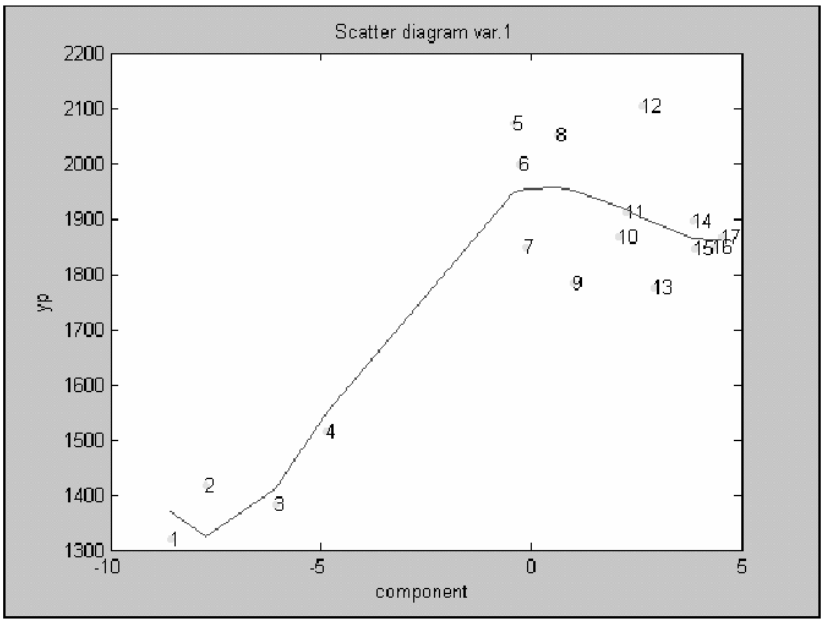

Figure 3.
PCA component regression (first component).

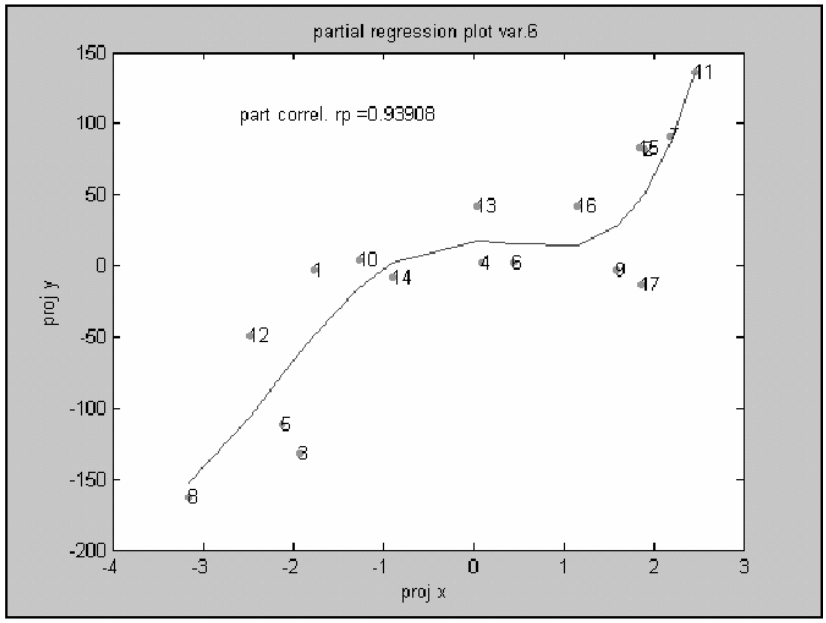

Figure 4.
Linear regression PLR (var6).

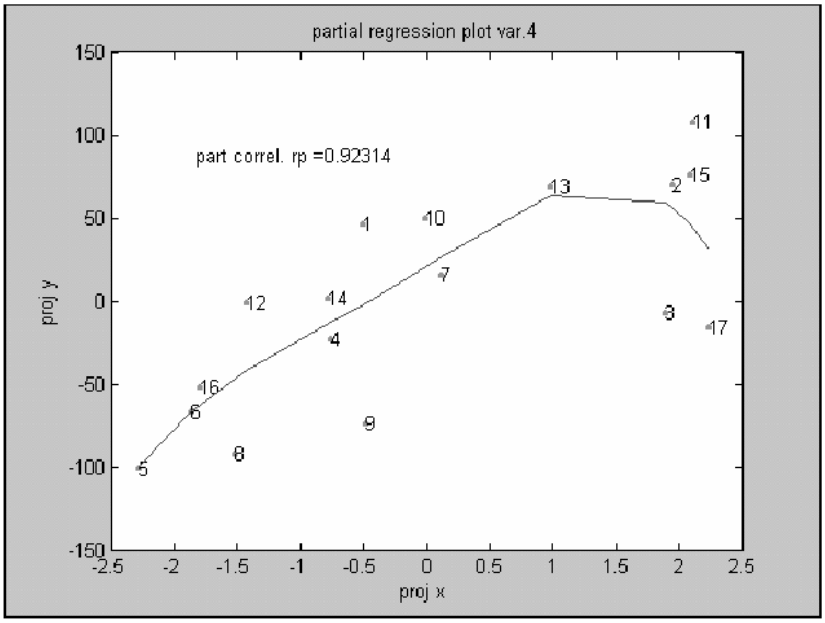

Figure 5.
Linear regression PLR (var4).

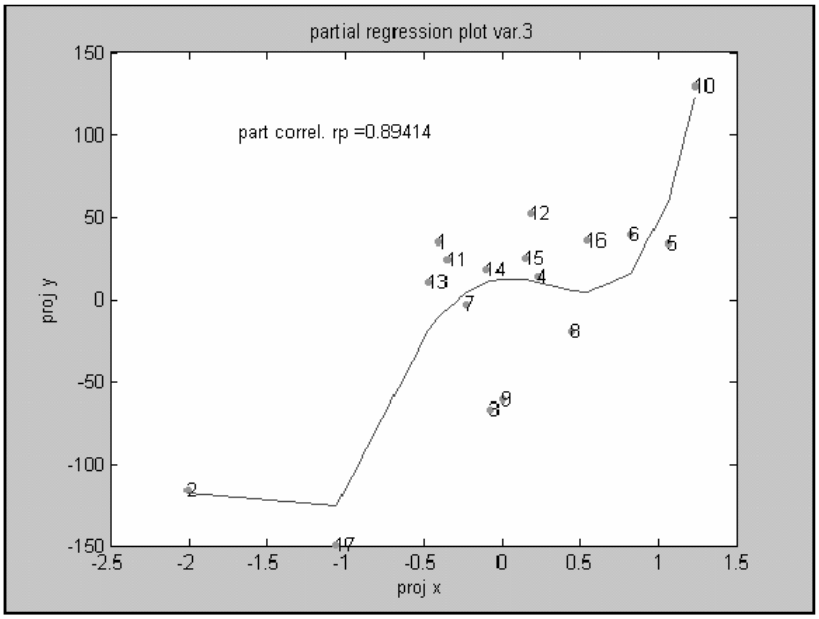

Figure 6.
Linear regression PLR (var3).