Analytica Chimica Acta, 277 (1993) 267–271
Elsevier Science Publishers B.V., Amsterdam

Use of the mean quadratic error of prediction for the construction of biased linear models

Jiří Militký

Department of Textile Materials, Technical University, Liberec (Czech Republic)

Milan Meloun

Department of Analytical Chemistry, Technical University, Pardubice (Czech Republic)

(Received 30th June 1992; revised manuscript received 3rd August 1992)

Abstract

The main practical problems caused by multi-collinearity are reviewed. The biased estimators based on the generalization of principal components for avoiding multi-collinearity problems are described. The mean quadratic error of prediction criterion is used for the selection of suitable bias. Some advantages of biased regression are demonstrated on the problem of intercept estimation in a polynomial model.

Keywords: Optimization methods; Biased linear models; Mean quadratic error of prediction; Multi-collinearity

Multiple linear and non-linear chemical model building is among the most complex problems to be solved in chemometric practice. An interactive approach to model building can be divided into the following steps [1]: selection of provisional models; analysis of assumptions about the model, data and regression methods (regression diagnostics); extension and modification of the model, data and regression method; and testing the validity of the model, its predictive capability, etc.

An interactive strategy of multiple model building based on the above steps has been described [1]. Many problems in the realization of the second step are caused by strong multi-collinearity. Multi-collinearity in multiple linear regression analyses is defined as approximate linear dependences among the explanatory variables (columns of design matrix X).

Correspondence to: J. Militký, Department of Textile Materials, Technical University, Liberec (Czech Republic).

It is well known that under strong multi-collinearity the parameter estimates and hypotheses test are affected more by linear "links" between explanatory variables than by the regression model itself. The classical t-test of significance is highly inflated owing to large variances of regression coefficient estimates and the results of regression are often unacceptable.

A number of alternatives to the least-squares approach have been recommended to avoid multi-collinearity. The resulting estimators are biased, but may be preferable to classical least squares. The most popular of these are the ridge-type estimators proposed by Hoerl and Kennard [2] and several others [3].

In this paper the estimators based on generalized principal components are adopted. For suitable bias selection the criterion based on the mean quadratic error of prediction (MEP) is used. The proposed procedure of biased estimator construction is a part of the package CHEMSTAT for data analysis in chemometric practice.

SUMMARY OF LINEAR REGRESSION

The standard linear model with n observations of m explanatory variables is assumed. For an additive model of measurement errors the linear regression model has the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{1}$$

In Eqn. 1 the $n \times m$ matrix $\boldsymbol{X}$ contains the values of m explanatory (predictor) variables at each of n observations, $\boldsymbol{\alpha}$ is the $m \times 1$ vector of regression parameters and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of experimental errors; $\boldsymbol{y}$ is $n \times 1$ vector of observed values of the dependent variable.

The classical least-squares method is based on the following assumptions: regression parameters are not restricted; the regression model is linear in parameters and the additive model of measurements is valid (see Eqn. 1); design matrix $\boldsymbol{X}$ has a rank equal to n; and errors ϵ_i are independent identically distributed random variables with zero mean $E(\epsilon_i) = 0$ and diagonal covariance matrix $D(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{E}$, where $\sigma^2 < \infty$. For testing purposes it is assumed that errors ϵ_i have a normal distribution $N(0, \sigma^2)$. When these four assumptions are valid the parameter estimates $\boldsymbol{b}$ found by minimization of the least-squares criterion

$$S(\boldsymbol{b}) = \| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \| \tag{2}$$

are best linear unbiased estimators (BLUE). In Eqn. 2, $\| \ \|$ is the symbol for Euclidean norm.

The conventional least-squares estimator $\boldsymbol{b}$ has the form

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

The corresponding covariance is

$$D(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \tag{4}$$

From a geometrical point of view columns of design matrix $\boldsymbol{X}$ define an m-dimensional hyperplane L in n-dimensional Euclidean space E^n. The vector $\boldsymbol{X}\boldsymbol{\beta}$ and prediction vector

$$\boldsymbol{y}_{\mathrm{P}} = \boldsymbol{X}\boldsymbol{b} \tag{5}$$

lie in plane L. The prediction vector is an orthogonal projection of vector $\boldsymbol{y}$ to the plane L.

$$\boldsymbol{y}_{\mathrm{P}} = \boldsymbol{H}\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{6}$$

where $\boldsymbol{H}$ is the projection matrix. The residual vector

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{y}_{\mathrm{P}} \tag{7}$$

is orthogonal to plane L and has the minimum length. Vector $\boldsymbol{e}$ is related to projection matrix $\boldsymbol{H}$:

$$\boldsymbol{e} = (\boldsymbol{E} - \boldsymbol{H})\boldsymbol{y} \tag{8}$$

$\boldsymbol{E}$ denotes a unit matrix of order n. The variance matrix corresponding to prediction vector $\boldsymbol{y}_{\mathrm{P}}$ has the form

$$D(\boldsymbol{y}_{\mathrm{P}}) = \sigma^2 \boldsymbol{H} \tag{9}$$

and the variance matrix for residuals is

$$D(\boldsymbol{e}) = \sigma^2(\boldsymbol{E} - \boldsymbol{H}) \tag{10}$$

Statistical analysis related to least squares is based on normality of estimates $\boldsymbol{b}$.

MULTI-COLLINEARITY

Multi-collinearity does not mean a violation of assumptions about least-squares methods. It concerns an ill-conditioning of the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ which has two consequences: the determinant of matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is near zero and some eigenvalues of matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ are near zero. This problem arises in cases when one of columns x_j of matrix $\boldsymbol{X}$ is a near linear combination of several other columns.

Multi-collinearity causes many difficulties in the inverse of matrix $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$, i.e., numerical difficulties. In addition to numerical difficulties, multi-collinearity also leads to the following statistical difficulties: non-stability of estimates b caused by the great sensitivity of parameter estimates to small changes in the data vector $\boldsymbol{y}$, the estimates $\boldsymbol{b}$ often having unacceptable signs and magnitudes, which effects their chemometric interpretation; large variances $D(b_j)$ of individual estimates cause the t-test to indicate statistical insignificance of parameters β_j; and a strong correlation between elements of estimates $\boldsymbol{b}$ means that they cannot be interpreted separately.

On the other hand, in the case of multi-collinearity the determination coefficient (square of multiple correlation coefficient) is often high and

a regression model may fit the data fairly well. For the case of data smoothing by regression models, multi-collinearity does not cause any difficulties except numerical ones. Multi-collinearity exhibits serious problems especially in regression model building (selection of explanatory variables, etc.).

Sources of multi-collinearity can be categorized into following major groups: the over-estimated regression mode; inappropriate location of experimental points; and physical constraints in model or in data.

Techniques suitable for the detection of multi-collinearity and sources of multi-collinearity have been described previously [1].

GENERALIZED PRINCIPAL COMPONENT

For convenience it is assumed that the columns of matrix $\boldsymbol{X}$ are properly scaled so that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is equal to correlation matrix $\boldsymbol{R}$. As the matrix $\boldsymbol{R}$ is symmetrical it may be expressed as a sum of eigenvalues $\tau_1 \leqslant \tau_2 \leqslant \ldots \leqslant \tau_m$ and corresponding eigenvectors $\boldsymbol{P}_j$, $j = 1, \ldots, m$:

$$\boldsymbol{R} = \sum_{j=1}^{m} \tau_j \boldsymbol{P}_j \boldsymbol{P}_j^{\mathrm{T}} \tag{11}$$

The inverse matrix $\boldsymbol{R}^{-1}$ can be then expressed in the form

$$\boldsymbol{R}^{-1} = \sum_{j=1}^{m} \tau_j^{-1} \boldsymbol{P}_j \boldsymbol{P}_j^{\mathrm{T}} \tag{12}$$

The normalized estimates $\boldsymbol{b}_{\mathrm{N}}$ can be calculated by substituting Eqn. 12 into Eqn. 3:

$$\boldsymbol{b}_{\mathrm{N}} = \sum_{j=z}^{m} \left[\tau_j^{-1} \boldsymbol{P}_j \boldsymbol{P}_j^{\mathrm{T}}\right] \boldsymbol{r} \tag{13}$$

The vector $\boldsymbol{r}$ is the scaled version of the vector $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ containing paired correlation coefficients between dependent and exploratory variables.

The corresponding covariance matrix has the form

$$D(\boldsymbol{b}_{\mathrm{N}}) = \sigma_{\mathrm{N}}^2 \sum_{j=z}^{m} \tau_j^{-1} \boldsymbol{P}_j \boldsymbol{P}_j^{\mathrm{T}} \tag{14}$$

In the case of the least squares in Eqns. 13 and 14, the constant z is taken as $z = 1$. For the principal component regression the z can be equal to 1, 2, 3, ...

From both Eqns. 13 and 14 it follows that when the eigenvalues τ_j are small the estimates $\boldsymbol{b}_{\mathrm{N}}$ and their variances are high. When some of τ_j are equal to zero the $\boldsymbol{b}_{\mathrm{N}}$ and $D(\boldsymbol{b}_{\mathrm{N}})$ are infinite. One way to avoid these difficulties is the use of generalized principal component when the small eigenvalues τ_j (or its parts) are neglected [4].

Let us denote

$$W = \sum_{j=1}^{z} \tau_j \quad \text{and} \quad E = \sum_{j=1}^{m} \tau_j$$

The criterion for leaving out the parts with too small eigenvalues then has the form

$$\mathrm{abs}(W/E) = P \tag{15}$$

where P is a selected parameter (often equal to 10^{-5}). Equality 15 cannot be generally valid for an integral z and given P. In this instance the minimum value of z for which the inequality

$$\mathrm{abs}(W/E) > P$$

is valid is selected. The summation in Eqns. 13 and 14 is then made from $z - 1$ and the term corresponding to eigenvalue τ_{z-1} is "weighted" by the factor

$$U = \left[\mathrm{abs}(W) - \mathrm{abs}(E)P\right]/\tau_z \tag{16}$$

By using this procedure, the length of estimates $\| \boldsymbol{b}_{\mathrm{N}} \|$ with their variances may be continuously decreased in dependence on increasing parameter P. Parameter P then corresponds to bias caused by neglecting some terms in Eqns. 13 and 14.

A suitable magnitude of P can be determined from the requirement for a minimum of the mean quadratic error of prediction.

SELECTION OF SUITABLE P

One of the main properties of regression models is the good predictive ability. This predictive ability can be adopted also for selection of, in some sense optimum, parameter P.

TABLE 1

Experimental data

x	25	35	45	55	65	75	85	95	105	115
y	150	160	170	190	210	230	270	310	370	450

The predictive ability in a linear regression model can be characterized by the mean quadratic error of prediction (MEP), defined generally by

$$\mathrm{MEP} = \sum_{i=1}^{n} \left[y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{b}(i) \right]^2 / n \qquad (17)$$

where $\boldsymbol{b}_{(i)}$ is the estimate of regression model parameters when all points except the ith (ith row $\boldsymbol{x}_i$ of matrix $\boldsymbol{X}$) are used. The statistics MEP uses a prediction $y_{Pi} = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{b}_{(i)}$ which was constructed without information about the ith point.

The estimate $\boldsymbol{b}_{(i)}$ can be calculated from the least-squares estimate $\boldsymbol{b}$:

$$\boldsymbol{b}_{(i)} = \boldsymbol{b} - \left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \boldsymbol{x}_i e_i \right] / [1 - H_{ii}] \qquad (18)$$

where H_{ii} is the diagonal element of the projection matrix $\boldsymbol{H}$.

After substitution from Eqn. 18 into Eqn. 17, the following simple relation results:

$$\mathrm{MEP} = n^{-1} \sum_{i=1}^{n} e_i^2 / (1 - H_{ii})^2 \qquad (19)$$

For a selected P it is possible to calculate values of H_{ii} from Eqn. 6 and then the MEP criterion from Eqn. 19.

A suitable P corresponds to some minimum value of MEP. For the selection of this value of P a very simple strategy can be used: for $P \approx 10^{-30}$ the MEP_1 is calculated; for $z = 2, 3, \ldots$ the MEP_z are calculated until $\mathrm{MEP}_z < \mathrm{MEP}_{z-1}$; and in the interval $W_{z-1}/E \leqslant P \leqslant W_z/E$ the optimum P is selected by the interval halving method.

This procedure is very simple and requires only one decomposition of matrix $\boldsymbol{R}$. The calculated P do not correspond generally to a global minimum but parameter estimates and the statistical characteristics are greatly improved.

In the program package CHEMSTAT (TriloByte) the generalized principal component is used and the MEP criterion is computed. Then the trial-and-error procedure can be adopted for selecting a suitable P.

EXAMPLE

Many problems in chemometrics concern an approximation of instrumental data of convex (or concave) increasing (or decreasing) values by a polynomial so that this polynomial fulfils the condition of remaining the shape of the data. For solution of these types of problems the generalized principal component with optimum P minimizing the MEP can be used.

In connection with modelling a mechanical problem the intercept term was important. The

TABLE 2

Results for principal component regression

z	τ_z	P	MEP	b_7	s_7
1	9.30×10^{-10}	1.5×10^{-10}	380.20	195.5	355.9
2	4.53×10^{-7}	7.6×10^{-8}	15.60	152.3	81.2
3	7.38×10^{-5}	1.2×10^{-5}	9.12	138.7	27.6
4	5.65×10^{-3}	9.5×10^{-4}	8.85	128.3	8.7
5	2.35×10^{-1}	4.0×10^{-2}	7.63	131.2	3.8
6	5.76	1.0	10.29	136.1	2.4

TABLE 3

Regression results for least-squares (LS) and generalized principal component (GPC) regression

Method	P	MEP	D [a]	b_7	s_7
LS	10^{-30}	380.30	0.99960	195.5	355.9
GPC	0.28	6.436	0.99953	132.4	3.48

[a] Square of multiple correlation coefficient.

data are strictly convex (see Table 1) and the regression model was specified as a polynomial of the sixth degree (based on some formal and theoretical assumptions) [5]:

$$E(y) = \sum_{j=1}^{6} b_j x^j + b_7$$

The parameter b_7 is equal to the intercept. Table 2 gives results for the principal component regression.

Table 3 gives estimates b_7, standard deviations s_7 and determination coefficient D found by the classical least-squares procedure ($P = 10^{-30}$) and generalized principal components $P = 0.28$ for which the MEP criterion was the smallest.

From Tables 1 and 3 it is obvious that the intercepts from LS do not correspond to the experimental data. The estimate b_7 is higher than the values y_1, y_2, y_3 and y_4, which indicates that the proposed model has some minimum between the origin and point (x_1, y_1). The corresponding standard deviation s_7 is very high so that the estimate b_7 is very imprecise. The parameter b_7 calculated by generalized principal components is acceptable and precise.

Conclusion

The method of generalized principal components in combination with the MEP criterion is very attractive for constructing biased models. It can be also used for achieving such estimates which keep the model course corresponding to the data trend especially in polynomial-type models. This method is implemented in the software package CHEMSTAT (TriloByte) [6].

REFERENCES

1 M. Meloun, J. Militký and M. Forina, Chemometrics 2, Interactive Model Building and Testing on IBM PC, Horwood, Chichester, 1992.
2 A.E. Hoerl and P.W. Kennard, Technometrics, 12 (1970) 55.
3 R.R. Hocking, F.M. Speed and M.J. Lynn, Technometrics, 18 (1976) 425.
4 D.M. Marquardt, Technometrics, 12 (1970) 591.
5 K. Květoň, unpublished report, Czech Technical University, Prague, 1988.
6 CHEMSTAT Version 2.0, TriloByte, Pardubice, Czechoslovakia, 1991.