Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 86 (2007) 146–152
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Multilinear regression: Using SOL regression to explore the expectation space and using experimental designs to determine the best solution

Jacques Goupy

ReConFor, 24 avenue Perrichont, 75016 Paris, France

Received 8 January 2006; received in revised form 23 May 2006; accepted 7 June 2006
Available online 21 July 2006

Keywords: Multilinear regression; Sequential regression; PLS1 regression; Orthogonal regression; Experimental design; Optimization

1. Introduction

Classical multilinear regression [1–3] provides just one solution which is obtained by the least-squares method. PLS (Partial least squares) regression [4–6] has as many solutions as the number of coefficients in the mathematical model. Among these solutions it is then possible to choose the one which meets, at the best, a criterion of optimization. But the number of solutions is limited and the best solution could be between two PLS' solutions and the optimization is only partial. This is why Sequential Orthogonal Linear (or SOL) regression was proposed. This regression has an infinite number of solutions and the expectation space is entirely covered. It is then possible to find the best solution using experimental designs or statistical optimization tools as simplex, golden section or Fibonacci series.

2. Classical multilinear regression

Data are represented by a linear model:

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e} \tag{1}$$

Where

- $\mathbf{y}$ is the vector of random variables $(n,1)$ or response vector. This vector can be illustrated in the (n)dimensional orthogonal space, R^n, spanned by the n responses.
- $\mathbf{X}$ is the matrix (n,p) of the regressors. This matrix is assumed to be of full rank. According to Bates and Watts [2] the expectation space is the (p)dimensional subspace, R^p, of R^n defined by the p columns of $\mathbf{X}$. This space can also be named "regressor space". Each point of this subspace represents a set of coefficients which is a potential solution of the regression.
- $\mathbf{a}$ is the parameter vector $(p,1)$ of the linear model.
- $\mathbf{e}$ is the deviation vector $(n,1)$.

As there are n variables y, and p parameters, there are n equations and $n+p$ unknowns. The p missing equations are deduced from the least-squares criterion. The regression system is then:

$$\begin{cases} \mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e} \\ \dfrac{\partial\, {}^t\mathbf{e}\,\mathbf{e}}{\partial\, \mathbf{a}} = 0 \end{cases} \tag{2}$$

Under classical assumptions, the multilinear regression provides the solution:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \tag{3}$$

where

- $\hat{\mathbf{y}}$ is the expected response vector $(n,1)$ orthogonally projected onto the expectation space.
- $\mathbf{H}$ is the projection matrix or "hat" matrix, given by:

$$\mathbf{H} = \mathbf{X}[{}^t\mathbf{X}\mathbf{X}]^{-1}\,{}^t\mathbf{X} \tag{4}$$

The residual vector is equal to:

$$\hat{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{5}$$

And the least-squares estimate of the parameters is:

$$\hat{\mathbf{a}} = ({}^t\mathbf{X}\mathbf{X})^{-1}\,{}^t\mathbf{X}\,\hat{\mathbf{y}} \tag{6}$$

E-mail address: jacques@goupy.com.

0169-7439/$ - see front matter
doi:10.1016/j.chemolab.2006.06.010

3. SOL regression

3.1. General properties

Instead of projecting the response vector orthogonally onto the expectation space, it can be projected orthogonally on a specific direction Δ_1 (defined by the unit vector $\mathbf{t}_1$) of this space [7]. The system has n equations and $n+1$ unknowns. The least-squares criterion gives the missing unknown. Then, the regression system is:

$$\begin{cases} \mathbf{y} = \mathbf{X}\mathbf{b}_1 + \mathbf{e} \\ \dfrac{\partial\, {}^t\mathbf{e}\,\mathbf{e}}{\partial\, k} = 0 \\ \mathbf{b}_1 = k\, \mathbf{t}_1 \end{cases} \tag{7}$$

where

- $\mathbf{b}_1$ is the matrix $(p,1)$ of the parameters. These parameters are proportional to the direction parameters of $\mathbf{t}_1$.
- $\mathbf{t}_1$ is the matrix $(p,1)$ of the direction parameters of Δ_1. These parameters are known as they are chosen by the experimenter.
- k is a coefficient of proportionality. This is the unknown.

The solution is (Fig. 1):

$$\tilde{\mathbf{y}}_1 = \mathbf{H}_1\mathbf{y} \tag{8}$$

where

- $\tilde{\mathbf{y}}$ is the projection of the response vector $\mathbf{y}$ $(n,1)$ onto Δ_1.
- $\mathbf{H}_1$ is the projection matrix onto Δ_1, given by:

$$\mathbf{H}_1 = \mathbf{X}\mathbf{t}_1\left[{}^t\mathbf{t}_1\, {}^t\mathbf{X}\,\mathbf{X}\mathbf{t}_1\right]^{-1}\, {}^t\mathbf{t}_1\, {}^t\mathbf{X} \tag{9}$$

The residual vector is equal to:

$$\mathbf{e}_1 = (\mathbf{I}-\mathbf{H}_1)\mathbf{y} \tag{10}$$

And the least-squares estimate of the parameters corresponding to $\tilde{\mathbf{y}}_1$ is:

$$\tilde{\mathbf{b}}_1 = ({}^t\mathbf{X}\mathbf{X})^{-1}\,{}^t\mathbf{X}\ \tilde{\mathbf{y}}_1 \tag{11}$$

We have

$$\mathbf{y} = \tilde{\mathbf{y}}_1 + \mathbf{e}_1 \tag{12}$$

The orthogonal projection of the residual vector $\mathbf{e}_1$ onto the expectation space is BH (Fig. 1). The triangle "$\tilde{\mathbf{y}}_1$, BH, $\hat{\mathbf{y}}$" is a right-angled triangle, and B is on the hyper-sphere whose diameter is $\hat{\mathbf{y}}$. Then, the complete decomposition of $\mathbf{e}_1$ ends in $\hat{\mathbf{e}}$ and the regression solution is $\hat{\mathbf{y}}$.

The number of dimensions of the expectation space is equal to the number of model parameters: p.

As one direction has been used, the Δ_1 direction, the residual vector e_1 can be projected on the $(p-1)$dimensional subspace which is orthogonal to Δ_1. This subspace is defined by the columns of the $\mathbf{X}_{r_1}$ matrix which is obtained by the orthogonal projection:

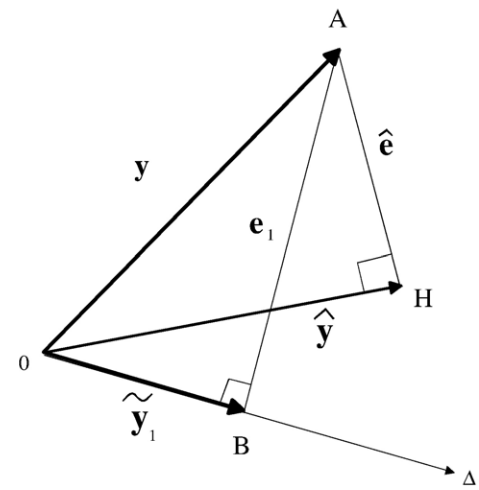

Fig. 1. Vector $\mathbf{y}$ is decomposed on the Δ_1 direction into a regressed vector $\tilde{\mathbf{y}}_1$ and a residual vector, $\mathbf{e}_1$.

$$\mathbf{X}_{r_1} = (\mathbf{I}-\mathbf{H}_1)\mathbf{X} \tag{13}$$

A new direction, Δ_2, is chosen in this new subspace and is defined by the vector $\mathbf{t}_2$. The residual vector $\mathbf{e}_1$ is projected onto this direction. The regressed vector, $\tilde{\mathbf{e}}_2$, and the corresponding residual vector, $\mathbf{e}_2$, are obtained (Fig. 2):

$$\tilde{\mathbf{e}}_2 = \mathbf{H}_2\mathbf{e}_1 \qquad \mathbf{e}_2 = (\mathbf{I}-\mathbf{H}_2)\mathbf{e}_1$$

with

$$\mathbf{H}_2 = \mathbf{X}_{r_1}\mathbf{t}_2\left[{}^t\mathbf{t}_2\, {}^t\mathbf{X}_{r_1}\,\mathbf{X}_{r_1}\mathbf{t}_2\right]^{-1}\, {}^t\mathbf{t}_2\, {}^t\mathbf{X}_{r_1}$$

The second solution of the regression is:

$$\tilde{\mathbf{y}}_2 = \tilde{\mathbf{y}}_1 + \tilde{\mathbf{e}}_2$$

The residual vector $\mathbf{e}_2$, corresponding to the projection of $\mathbf{e}_1$ onto Δ_2, can be projected on a Δ_3 direction belonging to the $(p-2)$ dimensional subspace which is orthogonal to Δ_1 and Δ_2. This subspace is defined by the $\mathbf{X}_{r_2}$ matrix:

$$\mathbf{X}_{r_2} = (\mathbf{I}-\mathbf{H}_2)\mathbf{X}_{r_1}$$

The residual vector, $\mathbf{e}_2$, being projected onto the Δ_3 direction, the regressed vector, $\tilde{\mathbf{e}}_3$, and the corresponding residual vector, $\mathbf{e}_3$, are obtained as follows:

$$\tilde{\mathbf{e}}_3 = \mathbf{H}_3\ \mathbf{e}_2 \qquad \mathbf{e}_3 = (\mathbf{I}-\mathbf{H}_3)\mathbf{e}_2$$

where

$$\mathbf{H}_3 = \mathbf{X}_{r_2}\ \mathbf{t}_3\left[{}^t\mathbf{t}_3\, {}^t\mathbf{X}_{r_2}\mathbf{X}_{r_2}\mathbf{t}_3\right]^{-1}\, {}^t\mathbf{t}_3\, {}^t\mathbf{X}_{r_2}$$

The third possible solution of the regression, $\tilde{\mathbf{y}}_3$, is (Fig. 3):

$$\tilde{\mathbf{y}}_3 = \tilde{\mathbf{y}}_2 + \tilde{\mathbf{e}}_3 = \tilde{\mathbf{y}}_1 + \tilde{\mathbf{e}}_2 + \tilde{\mathbf{e}}_3$$

The residual vectors are successively decomposed until the classical least-squares solution, $\hat{\mathbf{y}}$, is found.