Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 17–46
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

PLS generalised linear regression

Philippe Bastien[a], Vincenzo Esposito Vinzi[b,c,*], Michel Tenenhaus[c]

[a] *L'Oréal Recherche, Aulnay, France*
[b] *Department of Mathematics and Statistics University "Federico II", Naples, Italy*
[c] *HEC, School of Management Jouy-en-Josas, France*

Accepted 6 February 2004

Abstract

PLS univariate regression is a model linking a dependent variable $\mathbf{y}$ to a set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ of (numerical or categorical) explanatory variables. It can be obtained as a series of simple and multiple regressions. By taking advantage from the statistical tests associated with linear regression, it is feasible to select the significant explanatory variables to include in PLS regression and to choose the number of PLS components to retain. The principle of the presented algorithm may be similarly used in order to yield an extension of PLS regression to PLS generalised linear regression. The modifications to classical PLS regression, the case of PLS logistic regression and the application of PLS generalised linear regression to survival data are studied in detail. Some examples show the use of the proposed methods in real practice. As a matter of fact, classical PLS univariate regression is the result of an iterated use of ordinary least squares (OLS) where PLS stands for *partial least squares*. PLS generalised linear regression retains the rationale of PLS while the criterion optimised at each step is based on maximum likelihood. Nevertheless, the acronym PLS is kept as a reference to a general methodology for relating a response variable to a set of predictors. The approach proposed for PLS generalised linear regression is simple and easy to implement. Moreover, it can be easily generalised to any model that is linear at the level of the explanatory variables.

Keywords: Partial least squares; Stepwise regression; Variable selection; Modified PLS regression

* Corresponding author. Dipartimento di Matematica e Statistica, Università degli Sutdi di Napoli "Federico II", Via Cintia—Complesso Monte Sant'Angelo, 80126 Napoli, Italy. Tel.: +39-081-675112; fax: +39-081-675113.
E-mail address: vincenzo.espositovinzi@unina.it (V.E. Vinzi).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.02.005

1. PLS regression background

All variables $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p$ are assumed to be centred.

The PLS regression model with m components is written as

$$\mathbf{y} = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j \right) + \text{residual}, \tag{1}$$

with the constraint that the PLS components $\mathbf{t}_h = \sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j$ are orthogonal. We can consider that the parameters c_h and w_{hj}^{*} in model (1) are to be estimated. This is the nonlinear aspect of the model.

In the following, the same notation is used for the model parameters and their estimates. The context will clarify the actual meaning of the notation.

PLS regression (Wold et al., 1983; Tenenhaus, 1998; Garthwaite, 1994) is an algorithm for estimating the parameters of model (1). In the following, this algorithm is presented in a new version by linking each step to a simple or multiple OLS regression.

Computation of the first PLS component $\mathbf{t}_1$. The first component $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1^{*}$ is defined as

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}, \mathbf{x}_j)^2}} \sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}, \mathbf{x}_j) \mathbf{x}_j. \tag{2}$$

The weight $\operatorname{cov}(\mathbf{y}, \mathbf{x}_j)$ for the variable $\mathbf{x}_j$ may be also written as $\operatorname{cor}(\mathbf{y}, \mathbf{x}_j)^{*} s(y)^{*} s(\mathbf{x}_j)$ where $s(y)$ and $s(\mathbf{x}_j)$ are, respectively, the standard deviation of y and $\mathbf{x}_j$. As a consequence, in order for a variable $\mathbf{x}_j$ to be important in building up $\mathbf{t}_1$, it needs to be strongly correlated with $\mathbf{y}$ and to bear enough variability in terms of standard deviation.

The quantity $\operatorname{cov}(\mathbf{y}, \mathbf{x}_j)$ is also the regression coefficient a_{1j} in OLS simple regression between $\mathbf{y}$ and the modified explanatory variable $\mathbf{x}_j / \operatorname{var}(\mathbf{x}_j)$:

$$\mathbf{y} = a_{1j} (\mathbf{x}_j / \operatorname{var}(\mathbf{x}_j)) + \text{residual}. \tag{3}$$

Actually,

$$a_{1j} = \frac{\operatorname{cov}\left(\mathbf{y}, \dfrac{\mathbf{x}_j}{\operatorname{var}(\mathbf{x}_j)} \right)}{\operatorname{var}\left(\dfrac{\mathbf{x}_j}{\operatorname{var}(\mathbf{x}_j)} \right)} = \operatorname{cov}(\mathbf{y}, \mathbf{x}_j).$$

Thus, a test on the regression coefficient a_{1j} allows to assess the importance of the variable $\mathbf{x}_j$ in building $\mathbf{t}_1$. On this basis, the simple regression of $\mathbf{y}$ on $\mathbf{x}_j$ may be studied:

$$\mathbf{y} = a'_{1j} \mathbf{x}_j + \text{residual}. \tag{4}$$

As a matter of fact, there is an equivalence between testing whether a_{1j} or a'_{1j} are different from 0. In (2), each nonsignificant covariance may be replaced by a 0 so as to disregard the related explanatory variable.