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.

Computation of the second PLS component $\mathbf{t}_2$. At first, the $p+1$ simple regressions of, respectively, $\mathbf{y}$ and each $\mathbf{x}_j$ on $\mathbf{t}_1$ are run:

$$\mathbf{y} = c_1\mathbf{t}_1 + \mathbf{y}_1, \tag{5}$$

$$\mathbf{x}_j = p_{1j}\mathbf{t}_1 + \mathbf{x}_{1j}. \tag{6}$$

The second PLS component $\mathbf{t}_2$ is defined as

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j})^2}} \sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j})\mathbf{x}_{1j}. \tag{7}$$

The quantity $\operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j})$ is also the regression coefficient a_{2j} in the OLS multiple regression of $\mathbf{y}$ on $\mathbf{t}_1$ and $\mathbf{x}_{1j}/\operatorname{var}(\mathbf{x}_{1j})$

$$\mathbf{y} = c_{1j}\mathbf{t}_1 + a_{2j}(\mathbf{x}_{1j}/\operatorname{var}(\mathbf{x}_{1j})) + \text{residual}. \tag{8}$$

This result is a consequence of the orthogonality between the residual $\mathbf{x}_{1j}$ and the component $\mathbf{t}_1$.

The partial correlation between $\mathbf{y}$ and $\mathbf{x}_j$ given $\mathbf{t}_1$ is defined as the correlation between the residuals $\mathbf{y}_1$ and $\mathbf{x}_{1j}$. Similarly, the partial covariance between $\mathbf{y}$ and $\mathbf{x}_j$ given $\mathbf{t}_1$ is defined as the covariance between the residuals $\mathbf{y}_1$ and $\mathbf{x}_{1j}$:

$$\operatorname{cov}(\mathbf{y}, \mathbf{x}_j \mid \mathbf{t}_1) = \operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j}). \tag{9}$$

Therefore, the second PLS component may be alternatively written as

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}, \mathbf{x}_j \mid \mathbf{t}_1)^2}} \sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}, \mathbf{x}_j \mid \mathbf{t}_1)\mathbf{x}_{1j}. \tag{10}$$

A test on the regression coefficient a_{2j} allows to assess the importance of $\mathbf{x}_{1j}$ in building $\mathbf{t}_2$. It is also possible to test the contribution of $\mathbf{x}_j$ to the building of the second PLS component by studying directly the regression of $\mathbf{y}$ on $\mathbf{t}_1$ and $\mathbf{x}_j$:

$$\mathbf{y} = c'_{1j}\mathbf{t}_1 + a'_{2j}\mathbf{x}_j + \text{residual}. \tag{11}$$

As a matter of fact, there is an equivalence between testing a_{2j} or a'_{2j} being different from 0 because the pairs of vectors $(\mathbf{t}_1, \mathbf{x}_{1j})$ and $(\mathbf{t}_1, \mathbf{x}_j)$ span the same space.

In (7), each nonsignificant covariance may be replaced by a 0 so as to disregard the related explanatory variable from contributing to the model.

For interpretation purposes, the component $\mathbf{t}_2$ is better expressed as a function of the raw variables $\mathbf{x}_j$. This is possible because the residuals $\mathbf{x}_{1j} = \mathbf{x}_j - p_{1j}\mathbf{t}_1$ are a function of $\mathbf{x}_j$.

When expressed in terms of $\mathbf{x}_j$, the component $\mathbf{t}_2$ is written as $\mathbf{t}_2 = \mathbf{X}\mathbf{w}_2^*$. From the operational point of view, this is achieved by replacing the residual $\mathbf{x}_{1j}$ in (10) by its expression $\mathbf{x}_j - p_{1j}\mathbf{t}_1$.

Computation of the next components and stopping rule. We follow the same procedure for computing the next components $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h^*$. The search of new components is stopped either in accordance with a cross-validation procedure (as implemented in SIMCA-P 10.5 Umetri, 2004) or when all partial covariances are not significant.

PLS regression formula. In (1) the coefficients c_h are estimated by multiple regression of $\mathbf{y}$ on the PLS components $\mathbf{t}_h$. The estimated regression equation may be then expressed in terms of the original variables $\mathbf{x}_j$'s:

$$\begin{aligned}\hat{\mathbf{y}} &= \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j \right), \\ &= \sum_{j=1}^{p} \left(\sum_{h=1}^{m} c_h w_{hj}^{*} \right) \mathbf{x}_j, \\ &= \sum_{j=1}^{p} b_j \mathbf{x}_j. \end{aligned} \tag{12}$$

2. PLS generalised linear regression (PLS-GLR)

It has been shown that PLS regression can be obtained by using simple and multiple regression. If these regressions are replaced by GLR, a new series of models is accessed: PLS-GLR.

PLS-GLR of $\mathbf{y}$ on $\mathbf{x}_1, \ldots, \mathbf{x}_p$ with m components is written as:

$$g(\theta) = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j \right), \tag{13}$$

where the parameter θ may be either the mean of a continuous variable $\mathbf{y}$, or the probability vector of the values taken by a discrete variable $\mathbf{y}$, or the hazard ratio in survival data for proportional hazard ratio models. The link function g is chosen by the user according to the probability distribution of $\mathbf{y}$ and the model goodness of fit to the data. The new PLS components $\mathbf{t}_h = \sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j$ are constrained to be orthogonal. In the following, for the sake of simplicity, the same notation as before is kept for all elements in the model related to PLS-GLR. This new approach is firstly described and then applied to different examples.

2.1. PLS-GLR algorithm

The algorithm consists of four steps:

(1) computation of the m PLS components $\mathbf{t}_h$ $(h = 1, \ldots, m)$;
(2) generalised linear regression of $\mathbf{y}$ on the m retained PLS components;
(3) expression of PLS-GLR in terms of the original explanatory variables;
(4) Bootstrap validation of coefficients in the final model of PLS-GLR.

Hereinafter, the first and the last steps are described in details while the middle two steps are shown directly in the examples as they are trivial.

Computation of the PLS components. Let $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_p\}$ be the matrix comprising the previously centred p explanatory variables $\mathbf{x}_j$'s. The objective is to search for m PLS orthogonal components $\mathbf{t}_h$'s defined as linear combinations of $\mathbf{x}_j$.

Computation of the first PLS component $\mathbf{t}_1$

Step 1: Compute the regression coefficient a_{1j} of $\mathbf{x}_j$ in the generalised linear regression of $\mathbf{y}$ on $\mathbf{x}_j$ for each variable $\mathbf{x}_j$, $j=1$ to p,

Step 2: Normalise the column vector $\mathbf{a}_1$ made by a_{1j}'s: $\mathbf{w}_1=\mathbf{a}_1/\|\mathbf{a}_1\|$,

Step 3: Compute the component $\mathbf{t}_1=\mathbf{X}\mathbf{w}_1/\mathbf{w}_1'\mathbf{w}_1$.

Computation of the second PLS component $\mathbf{t}_2$

Step 1: Compute the regression coefficient a_{2j} of $\mathbf{x}_j$ in the generalised linear regression of $\mathbf{y}$ on $\mathbf{t}_1$ and $\mathbf{x}_j$ for each variable $\mathbf{x}_j$, $j=1$ to p,

Step 2: Normalise the column vector $\mathbf{a}_2$ made by a_{2j}'s: $\mathbf{w}_2=\mathbf{a}_2/\|\mathbf{a}_2\|$,

Step 3: Compute the residual matrix $\mathbf{X}_1$ of the linear regression of $\mathbf{X}$ on $\mathbf{t}_1$,

Step 4: Compute the component $\mathbf{t}_2=\mathbf{X}_1\mathbf{w}_2/\mathbf{w}_2'\mathbf{w}_2$,

Step 5: Express the component $\mathbf{t}_2$ in terms of $\mathbf{X}$: $\mathbf{t}_2=\mathbf{X}\mathbf{w}_2^*$.

Computation of the hth PLS component $\mathbf{t}_h$. In the previous steps, the PLS components $\mathbf{t}_1,\ldots,\mathbf{t}_{h-1}$ have been yielded. The component $\mathbf{t}_h$ is obtained by iterating the search for the second component.

Step 1: Compute the regression coefficient a_{hj} of $\mathbf{x}_j$ in the generalised linear regression of $\mathbf{y}$ on $\mathbf{t}_1,\ldots,\mathbf{t}_{h-1}$ and $\mathbf{x}_j$ for each variable $\mathbf{x}_j$, $j=1$ to p,

Step 2: Normalise the column vector $\mathbf{a}_h$ made by a_{hj}'s: $\mathbf{w}_h=\mathbf{a}_h/\|\mathbf{a}_h\|$,

Step 3: Compute the residual matrix $\mathbf{X}_{h-1}$ of the linear regression of $\mathbf{X}$ on $\mathbf{t}_1,\ldots,\mathbf{t}_{h-1}$,

Step 4: Compute the component $\mathbf{t}_h=\mathbf{X}_{h-1}\mathbf{w}_h/\mathbf{w}_h'\mathbf{w}_h$,

Step 5: Express the component $\mathbf{t}_h$ in terms of $\mathbf{X}$: $\mathbf{t}_h=\mathbf{X}\mathbf{w}_h^*$.

Remarks

(1) Computation of the PLS component $\mathbf{t}_h$ can be simplified by setting to 0 those regression coefficients a_{hj} that show to be not significant. Only significant variables will then contribute to the computation of the PLS component.

(2) The number m of PLS components to be retained may be chosen by cross-validation on the predictive power of the model or by observing that the component $\mathbf{t}_{m+1}$ is not significant because none of the coefficients $a_{m+1,j}$ is significantly different from 0.

(3) The proposed algorithm may cope with missing data. Let $\mathbf{x}_{h-1,i}$ be the column vector obtained by transposing the ith row of $\mathbf{X}_{h-1}$. The value $\mathbf{t}_{hi}=\mathbf{x}_{h-1,i}'\mathbf{w}_h/\mathbf{w}_h'\mathbf{w}_h$ of the ith case on the component $\mathbf{t}_h$ represents the slope of the OLS line without constant term related to the cloud of points $(\mathbf{w}_h,\ \mathbf{x}_{h-1,i})$. This slope may be computed even when there are missing data. In such a case, in computing the hth PLS component, the denominator of Step 4 is computed only on the data available also for the numerator.

Bootstrap validation procedure. Once PLS model has been estimated and the proper number of components to retain chosen, it is important to ascertain statistical significance of explanatory variables. Hereinafter, this is done in a nonparametric

framework by means of a bootstrap procedure under the assumption that PLS components $\mathbf{t}_h$'s ($h=1,\ldots,m$) are given as they represent the most stable part in the model. Most commonly used software for PLS regression (SIMCA-P (Umetri, 2004) and The Unscrambler® (Camo, 2003)) provide with jackknife estimates of PLS regression coefficients. However, the algorithms behind the construction of such estimates are not clearly given. For this reason, the development of an ad hoc nonparametric procedure is preferred in the following with related steps clearly stated.

Let $\hat{F}_{(\mathbf{T}|\mathbf{y})}$ be the empirical distribution function given by the PLS components matrix $\mathbf{T}=\{\mathbf{t}_1,\ldots,\mathbf{t}_h,\ldots,\mathbf{t}_m\}$ and the response variable $\mathbf{y}$.

Step 1: Draw B random samples from $\hat{F}_{(\mathbf{T}|\mathbf{y})}$;

Step 2: For $b=1,\ldots,B$, compute:

$$\mathbf{c}^{(b)}=(\mathbf{T}^{(b)\prime}\mathbf{T}^{(b)})^{-1}\mathbf{T}^{(b)\prime}\mathbf{y}^{(b)} \quad \text{and} \quad \mathbf{b}^{(b)}=\mathbf{W}^{*}\mathbf{c}^{(b)},$$

where $[\mathbf{T}^{(b)},\mathbf{y}^{(b)}]$ is the bth bootstrap sample, $\mathbf{c}^{(b)}$ is the vector of the corresponding estimates of PLS coefficients linking components to $\mathbf{y}$, $\mathbf{b}^{(b)}$ is the vector of p regression coefficients linking original explanatory variables to $\mathbf{y}$ (see Eq. (12)) for the bth sample, and $\mathbf{W}^{*}$ is the unchanged matrix of the original explanatory variables weights in constructing the m PLS components,

Step 3: For each $j=1,\ldots,p$ Φ_{b_j} is the Monte Carlo approximation of the bootstrap cumulative distribution function (*cdf*) of b_j.

For each b_j, box-plots and confidence intervals can be derived by means of the percentiles of Φ_{b_j}. The confidence interval is defined as $I_j(\alpha)=[\Phi_{b_j}^{-1}(\alpha),\Phi_{b_j}^{-1}(1-\alpha)]$ where $\Phi_{b_j}^{-1}(\alpha)$ and $\Phi_{b_j}^{-1}(1-\alpha)$ are the values from the bootstrap *cdf* such that a nominal $100(1-2\alpha)\%$ confidence interval is yielded. In order to improve the coverage accuracy (i.e. the capability of $I_j(\alpha)$ to give accurate coverage probabilities) of the above intervals, the so-called BC_a (*Bias-Correction and accelerated*) intervals, proposed in Efron and Tibshirani (1993), are considered. A BC_a interval is defined as

$$I_j^{\mathrm{BC}_a}(\alpha)=[\Phi_{b_j}^{-1}(\alpha_{j,1}),\Phi_{b_j}^{-1}(1-\alpha_{j,2})],$$

where

$$\alpha_{j,1}=\Theta\left(\hat{z}_0+\frac{\hat{z}_0+z(\alpha)}{1-\hat{a}[\hat{z}_0+z(\alpha)]}\right) \quad \text{and}$$

$$\alpha_{j,2}=\Theta\left(\hat{z}_0+\frac{\hat{z}_0+z(1-\alpha)}{1-\hat{a}[\hat{z}_0+z(1-\alpha)]}\right).$$

Here, $\Theta(\cdot)$ is the standard normal *cdf* and $z(\alpha)$ is the 100αth percentile of a standard normal distribution. The bias correction factor $\hat{z}_0$ and the acceleration factor $\hat{a}$ can be obtained, respectively, as

$$\hat{z}_0=\Theta^{-1}(\Phi_{b_j}(b_j)),$$

where $\Theta^{-1}(\cdot)$ is the inverse function of the standard normal *cdf*, and

$$\hat{a}=\frac{\sum_{i=1}^{n}(b_{j(\cdot)}-b_{j(i)})^3}{6\left[\sum_{i=1}^{n}(b_{j(\cdot)}-b_{j(i)})^3\right]^{3/2}},$$

where $b_{j(\cdot)}$ is the jack-knife estimate of b_j and $b_{j(i)}$ is the estimate of b_j when the ith observation is dropped out of the sample. If both $\hat{z}_0$ and $\hat{a}$ are equal to zero, then $I_j(\alpha)$ and $I_j^{\mathrm{BC}_a}(\alpha)$ coincide.

One of the advantages of BC_a intervals concerns their accuracy, i.e. a $100(1-2\alpha)\%$ confidence interval is supposed to have an overall probability 2α of not covering. Moreover, it can be shown that intervals $I_j^{\mathrm{BC}_a}(\alpha)$'s converge toward unbiasedness with a higher speed (relative to the number B of bootstrap samples) than intervals $I_j(\alpha)$'s. As a matter of fact, they are second order accurate intervals while simple bootstrap intervals are only first order accurate. Roughly speaking, the performance improvement of BC_a intervals is due to the fact that simple bootstrap confidence intervals implicitly assume symmetry of the bootstrap distribution of the estimator while BC_a intervals automatically take into account the shape (in terms of skewness and kurtosis) and the bias of the distribution.

It is worth noticing that the above-mentioned confidence intervals are not simultaneous (Shao and Tu, 1995). Therefore, they are not meant for pairwise or multiple comparisons but can only be interpreted one at a time.

2.2. The case of multiple regression

The application of the general algorithm proposed above to the specific case of multiple regression yields a natural modification of PLS regression that was earlier proposed in (Shenk and Westerhaus, 1991) with the name of "modified PLS regression". In this case, response variable $\mathbf{y}$ is standardised while explanatory variables $\mathbf{x}_j$'s are simply centred.

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

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{1j}^2}} \sum_{j=1}^{p} a_{1j}\mathbf{x}_j, \tag{14}$$

where a_{1j} is the regression coefficient of the centred variable $\mathbf{x}_j$ in the regression of $\mathbf{y}$ on each $\mathbf{x}_j$. By replacing a_{1j} by its value $\mathrm{Cov}(\mathbf{x}_j,\mathbf{y})/\mathrm{Var}(\mathbf{x}_j)$ expression (14) becomes

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \left[\frac{\mathrm{Cov}(\mathbf{x}_j,\mathbf{y})}{\mathrm{Var}(\mathbf{x}_j)}\right]^2}} \sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_j,\mathbf{y})\mathbf{x}_j^*, \tag{15}$$

where $\mathbf{x}_j^*$ shows that $\mathbf{x}_j$ is standardised. It is worth noticing that the explanatory variables are automatically reduced when changing from Eq. (14) to Eq. (15). It is certainly preferable to change the normalisation so as to obtain

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_j,\mathbf{y})^2}} \sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_j,\mathbf{y})\mathbf{x}_j^*. \tag{16}$$

Only those explanatory variables that are significantly correlated with $\mathbf{y}$ contribute to the computation of $\mathbf{t}_1$. There is no difference at all between this new component and the usual PLS component computed on standardised data.

Computation of the second PLS component $\mathbf{t}_2$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the explanation of $\mathbf{y}$ in addition to $\mathbf{t}_1$ are looked for. Therefore, the multiple regressions of $\mathbf{y}$ on $\mathbf{t}_1$ and each $\mathbf{x}_j$ are built

$$\mathbf{y} = c_1\mathbf{t}_1 + a_{2j}\mathbf{x}_j + \text{residual}. \tag{17}$$

In order to obtain a second component $\mathbf{t}_2$ that is orthogonal to $\mathbf{t}_1$, it is necessary to build it starting from the residuals $\mathbf{x}_{1j}$ of the regressions of each $\mathbf{x}_j$ on $\mathbf{t}_1$

$$\mathbf{x}_j = p_{1j}\mathbf{t}_1 + \mathbf{x}_{1j}. \tag{18}$$

Thus, the second component $\mathbf{t}_2$ is defined as

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{2j}^2}} \sum_{j=1}^{p} a_{2j}\mathbf{x}_{1j}, \tag{19}$$

where a_{2j} is also the regression coefficient of $\mathbf{x}_{1j}$ in the regression of $\mathbf{y}$ on $\mathbf{t}_1$ and $\mathbf{x}_{1j}$

$$\begin{aligned}\mathbf{y} &= c_1\mathbf{t}_1 + a_{2j}(p_{1j}\mathbf{t}_1 + \mathbf{x}_{1j}) + \text{residual}\\ &= (c_1 + a_{2j}p_{1j})\mathbf{t}_1 + a_{2j}\mathbf{x}_{1j} + \text{residual}.\end{aligned} \tag{20}$$

The orthogonality between $\mathbf{x}_{1j}$ and $\mathbf{t}_1$, by modifying the normalisation, leads to

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_{1j}, \mathbf{y})^2}} \sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_{1j}, \mathbf{y})\mathbf{x}_{1j}^*, \tag{21}$$

where $\mathbf{x}_{1j}^*$ represents the standardised version of the residual vector $\mathbf{x}_{1j}$.

Computation of the hth PLS component $\mathbf{t}_h$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the explanation of $\mathbf{y}$ in addition to $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$ are looked for. Therefore, the multiple regressions of $\mathbf{y}$ on $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$ and each $\mathbf{x}_j$ are built

$$\mathbf{y} = c_1\mathbf{t}_1 + c_2\mathbf{t}_2 + \cdots + c_{h-1}\mathbf{t}_{h-1} + a_{hj}\mathbf{x}_j + \text{residual}. \tag{22}$$

In order to obtain a hth component $\mathbf{t}_h$ that is orthogonal to $\mathbf{t}_{h-1}$, it is necessary to build it starting from the residuals $\mathbf{x}_{(h-1)j}$ of the regressions of each $\mathbf{x}_j$ on $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$

$$\mathbf{x}_j = p_{1j}\mathbf{t}_1 + p_{2j}\mathbf{t}_2 + \cdots + p_{(h-1)j}\mathbf{t}_{h-1} + \mathbf{x}_{(h-1)j}. \tag{23}$$

The component $\mathbf{t}_h$ is then defined as

$$\mathbf{t}_h = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{hj}^2}} \sum_{j=1}^{p} a_{hj}\mathbf{x}_{(h-1)j}, \tag{24}$$

where a_{hj} is also the regression coefficient of $\mathbf{x}_{(h-1)j}$ in the regression of $\mathbf{y}$ on $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$ and $\mathbf{x}_{(h-1)j}$

$$\begin{aligned}\mathbf{y} &= c_1\mathbf{t}_1 + c_2\mathbf{t}_2 + \cdots + c_{h-1}\mathbf{t}_{h-1} + a_{hj}(p_{1j}\mathbf{t}_1 + p_{2j}\mathbf{t}_2 + \cdots\\ &\quad + p_{(h-1)j}\mathbf{t}_{(h-1)} + \mathbf{x}_{(h-1)j}) + \text{residual}\\ &= (c_1 + a_{hj}p_{1j})\mathbf{t}_1 + (c_2 + a_{hj}p_{2j})\mathbf{t}_2 + \cdots + (c_{h-1} + a_{hj}p_{(h-1)j})\mathbf{t}_{h-1}\\ &\quad + a_{hj}\mathbf{x}_{(h-1)j} + \text{residual}.\end{aligned} \tag{25}$$

The orthogonality between $\mathbf{x}_{(h-1)j}$ and $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$, by modifying the normalisation, leads to

$$\mathbf{t}_h = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_{(h-1)j}, \mathbf{y})^2}} \sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_{(h-1)j}, \mathbf{y}) \mathbf{x}^*_{(h-1)j}, \tag{26}$$

where $\mathbf{x}^*_{(h-1)j}$ represents the standardised version of the residual vector $\mathbf{x}_{(h-1)j}$.

Stopping rule. The search for new components stops when all regression coefficients are not significant.

A case study: Cornell data on gasoline mixture. The proposed algorithm is now applied to a multiple regression problem related to the Cornell data (Kettaneh-Wold, 1992; Tenenhaus, 1998). Only the significant variables are retained for computing the PLS components. The dataset consists of 12 gasoline samples whose mixture is known in terms of proportions (Table 1). he influence of the mixture on the octane ratio $\mathbf{y}$ needs to be determined. Correlations between variables are given in Table 2.

Variables:
$\mathbf{y}$ = motor octane rating
$\mathbf{x}_1$ = straightrun (between 0 and 0.21)
$\mathbf{x}_2$ = reformate (between 0 and 0.62)
$\mathbf{x}_3$ = thermally cracked naphta (between 0 and 0.12)
$\mathbf{x}_4$ = catalytically cracked naphta (between 0 and 0.62)
$\mathbf{x}_5$ = polymer (between 0 and 0.12)
$\mathbf{x}_6$ = alkylate (between 0 and 0.74)
$\mathbf{x}_7$ = natural gasoline (between 0 and 0.08)

Table 1
Cornell data

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{y}$
0.00	0.23	0.00	0.00	0.00	0.74	0.03	98.70
0.00	0.10	0.00	0.00	0.12	0.74	0.04	97.80
0.00	0.00	0.00	0.10	0.12	0.74	0.04	96.60
0.00	0.49	0.00	0.00	0.12	0.37	0.02	92.00
0.00	0.00	0.00	0.62	0.12	0.18	0.08	86.60
0.00	0.62	0.00	0.00	0.00	0.37	0.01	91.20
0.17	0.27	0.10	0.38	0.00	0.00	0.08	81.90
0.17	0.19	0.10	0.38	0.02	0.06	0.08	83.10
0.17	0.21	0.10	0.38	0.00	0.06	0.08	82.40
0.17	0.15	0.10	0.38	0.02	0.10	0.08	83.20
0.21	0.36	0.12	0.25	0.00	0.00	0.06	81.40
0.00	0.00	0.00	0.55	0.00	0.37	0.08	88.10

The sum of all $\mathbf{x}_j$'s for a given sample is equal to 1.

Multiple regression. This dataset causes some problems when analysed with multiple regression. As a matter of fact, the output from Proc REG of SAS on a model without the constant term (Table 3) shows results that are uncoherent with the observed

Table 2
Correlation matrix

Pearson correlation coefficients, $N = 12$							
	Y	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$
Y	**1.000**						
X1	**−0.837**	1.000					
X2	−0.071	0.104	1.000				
X3	**−0.838**	**0.999**	0.101	1.000			
X4	**−0.707**	0.371	−0.537	0.374	1.000		
X5	0.494	−0.548	−0.293	−0.548	−0.211	1.000	
X6	**0.985**	−0.805	−0.191	−0.805	−0.646	0.463	1.000
X7	**−0.741**	0.603	−0.590	0.607	**0.916**	−0.274	−0.656

In bold, significant correlations with $p < 0.0001$.

correlations for variables $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5$ and $\mathbf{x}_7$. For variables $\mathbf{x}_1$, $\mathbf{x}_3$ and $\mathbf{x}_7$ this is due to a multicollinearity problem. For variables $\mathbf{x}_2$ and $\mathbf{x}_5$ an explication may be found by using PLS regression or, as it is reported in the following, by running a modified PLS regression.

Table 3
Multiple regression: SAS output

Dependent variable: Y
NOTE: No intercept in model. *R*-Square is redefined.
Root MSE 0.83619 R-Square 1.0000
Dependent Mean 88.58333 Adj R-Sq 0.9999
Coeff Var 0.94396

Variable	DF	Parameter estimate	Standard error	t Value	$\Pr > \lvert t \rvert$
$\mathbf{X}_1$	1	34.32023	209.12698	0.16	0.8761
$\mathbf{X}_2$	1	85.92283	1.24820	68.84	<0.0001
$\mathbf{X}_3$	1	141.25193	375.32840	0.38	0.7221
$\mathbf{X}_4$	1	77.18010	9.21351	8.38	0.0004
$\mathbf{X}_5$	1	87.75022	5.81572	15.09	<0.0001
$\mathbf{X}_6$	1	100.30083	3.47382	28.87	<0.0001
$\mathbf{X}_7$	1	116.92128	81.09563	1.44	0.2089

Forward stepwise modified PLS regression. The response variable $\mathbf{y}$ is taken to be standardised ($\mathbf{y}^*$).

Computation of $\mathbf{t}_1$. Component $\mathbf{t}_1$ is computed only with those variables $\mathbf{x}_j$'s that are significantly correlated to $\mathbf{y}$.

Then, formula (16) yields

$$\mathbf{t}_1 = \frac{-0.8373\mathbf{x}_1^* - 0.83796\mathbf{x}_3^* - 0.70671\mathbf{x}_4^* + 0.98507\mathbf{x}_6^* - 0.74112\mathbf{x}_7^*}{\sqrt{0.8373^2 + 0.83796^2 + 0.70671^2 + 0.98507^2 + 0.74112^2}}$$

$$= 0.4526\mathbf{x}_1^* - 0.4530\mathbf{x}_3^* - 0.3820\mathbf{x}_4^* + 0.5325\mathbf{x}_6^* - 0.4006\mathbf{x}_7^*.$$

Computation of $\mathbf{t}_2$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the construction of $\mathbf{t}_2$ are looked for. Therefore, 7 multiple regressions of $\mathbf{y}^*$ on $\mathbf{t}_1$ and each $\mathbf{x}_j$ are run. Significance levels (p-values) are reported in Table 4.

Table 4
Results from multiple regression of $\mathbf{y}^*$ on $\mathbf{t}_1$ and each $\mathbf{x}_j$

Variable	p-values
$\mathbf{x}_1$	0.6225
$\mathbf{x}_2$	**0.0101**
$\mathbf{x}_3$	0.6016
$\mathbf{x}_4$	0.9055
$\mathbf{x}_5$	0.7221
$\mathbf{x}_6$	$<$**0.0001**
$\mathbf{x}_7$	0.0532

Only variables $\mathbf{x}_2$ and $\mathbf{x}_6$ are significant at a risk level $\alpha = 0.05$. The residual vectors $\mathbf{x}_{12}$, $\mathbf{x}_{16}$ from the regressions of $\mathbf{x}_2$, $\mathbf{x}_6$ on $\mathbf{t}_1$ are computed. They are orthogonal to $\mathbf{t}_1$ and linear combinations of $\mathbf{x}_j$. As a consequence, any linear combination of $\mathbf{x}_{12}$ and $\mathbf{x}_{16}$ is also a linear combination of $\mathbf{x}_j$ and orthogonal to $\mathbf{t}_1$. The application of (21) leads to

$$\mathbf{t}_2 = \frac{1}{\sqrt{0.19565^2 + 0.25762^2}}(-0.19565\mathbf{x}_{12}^* + 0.25762\mathbf{x}_{16}^*),$$

$$= 0.4211\mathbf{x}_1^* - 0.6098\mathbf{x}_2^* + 0.4214\mathbf{x}_3^* + 0.3554\mathbf{x}_4^* + 1.5363\mathbf{x}_6^* + 0.3727\mathbf{x}_7^*.$$

It is worth noticing that variable $\mathbf{x}_2$ is not at all significant in terms of correlation with $\mathbf{y}$. Nevertheless, its partial correlation with $\mathbf{y}$ conditionally to $\mathbf{t}_1$ is significant.

Computation of $\mathbf{t}_3$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the construction of $\mathbf{t}_3$ are looked for. Therefore, 7 multiple regressions of $\mathbf{y}^*$ on $\mathbf{t}_1$, $\mathbf{t}_2$ and each $\mathbf{x}_j$ are run. Table 5 reports the p-values.

Table 5
Results from multiple regression of $\mathbf{y}^*$ on $\mathbf{t}_1$, $\mathbf{t}_2$ and each $\mathbf{x}_j$

Variables	p-values
$\mathbf{x}_1$	**0.0289**
$\mathbf{x}_2$	**0.0294**
$\mathbf{x}_3$	**0.0258**
$\mathbf{x}_4$	**0.0177**
$\mathbf{x}_5$	0.6356
$\mathbf{x}_6$	**0.0294**
$\mathbf{x}_7$	0.0922

Conclusion: Variables $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{x}_6$ are significant at a risk level $\alpha = 0.05$; a third PLS component $\mathbf{t}_3$ may be then computed.

The third PLS component $\mathbf{t}_3$ is then computed on the standardised residual vectors $\mathbf{x}_{21}^*$, $\mathbf{x}_{22}^*$, $\mathbf{x}_{23}^*$, $\mathbf{x}_{24}^*$, $\mathbf{x}_{26}^*$ from the multiple regressions of the significant explanatory variables on the components $\mathbf{t}_1$ and $\mathbf{t}_2$

$$\mathbf{t}_3 = \frac{0.0859\mathbf{x}_{21}^* + 0.0857\mathbf{x}_{22}^* + 0.0872\mathbf{x}_{23}^* - 0.0912\mathbf{x}_{24}^* + 0.0857\mathbf{x}_{26}^*}{\sqrt{0.0859^2 + 0.0857^2 + 0.0872^2 + 0.0912^2 + 0.0857^2}}$$

$$= 1.600\mathbf{x}_1^* + 1.248\mathbf{x}_2^* + 1.622\mathbf{x}_3^* - 0.243\mathbf{x}_4^* + 3.260\mathbf{x}_6^* + 0.541\mathbf{x}_7^*.$$

Computation of $\mathbf{t}_4$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the construction of $\mathbf{t}_4$ are looked for by studying the regressions of $\mathbf{y}^*$ on $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$ and each $\mathbf{x}_j$. Table 6 reports the p-values.

Table 6
Results from multiple regression of $\mathbf{y}^*$ on $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$ and each $\mathbf{x}_j$

Variable	p-values
$\mathbf{x}_1$	0.7096
$\mathbf{x}_2$	0.9378
$\mathbf{x}_3$	0.8517
$\mathbf{x}_4$	0.5711
$\mathbf{x}_5$	0.6867
$\mathbf{x}_6$	0.9378
$\mathbf{x}_7$	0.3351

Conclusion: There is no significant explanatory variable at the risk level $\alpha = 0.05$; only the first three PLS components are retained.

PLS regression equation with three components. Regression of the original response variable $\mathbf{y}$ on the three components $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$ leads to the equation

$$\hat{\mathbf{y}} = 88.58 + 3.25\mathbf{t}_1 + 2.53\mathbf{t}_2 + 1.20\mathbf{t}_3,$$

that can be also written in terms of the original explanatory variables so as to show the related regression coefficients:

$$\hat{\mathbf{y}} = 87.682 - 5.920\mathbf{x}_1 - 2.034\mathbf{x}_2 - 10.060\mathbf{x}_3 - 3.892\mathbf{x}_4 + 15.133\mathbf{x}_6 - 26.429\mathbf{x}_7$$

$$= 87.682\left(\sum_{j=1}^{7}\mathbf{x}_j\right) - 5.920\mathbf{x}_1 - 2.034\mathbf{x}_2 - 10.060\mathbf{x}_3$$

$$-3.892\mathbf{x}_4 + 15.133\mathbf{x}_6 - 26.429\mathbf{x}_7$$

$$= 81.762\mathbf{x}_1 + 85.648\mathbf{x}_2 + 77.622\mathbf{x}_3 + 83.790\mathbf{x}_4 + 87.682\mathbf{x}_5$$

$$+ 102.815\mathbf{x}_6 + 61.253\mathbf{x}_7.$$

The maximum octane ratio is obtained by setting $\mathbf{x}_2 = 0.14$, $\mathbf{x}_5 = 0.12$, $\mathbf{x}_6 = 0.74$ such that

$$\hat{\mathbf{y}} = 85.648 \times 0.14 + 87.682 \times 0.12 + 102.815 \times 0.74 = 98.59.$$

Use of classical PLS regression and bootstrap validation. PLS regression has been also run on Cornell data. Cross-validation leads to 3 PLS components. Fig. 1 shows a multiple box-plot representation of bootstrap distributions (with $B = 1000$) for the standardised regression coefficients related to the seven explanatory variables. All coefficients, except the one related to $\mathbf{x}_5$, have their bootstrap distributions well below or above zero and thus can be considered statistically significant. However, the graphical representation clearly shows asymmetry for almost all bootstrap distributions. Namely, the medians of the distributions for standardised b_1, b_3 and b_7 are closer to the respective third quartiles while the medians of the distributions for standardised b_2 and b_4 are close to the first quartiles; the distributions of standardised b_2 and b_7 show right heavy tails.

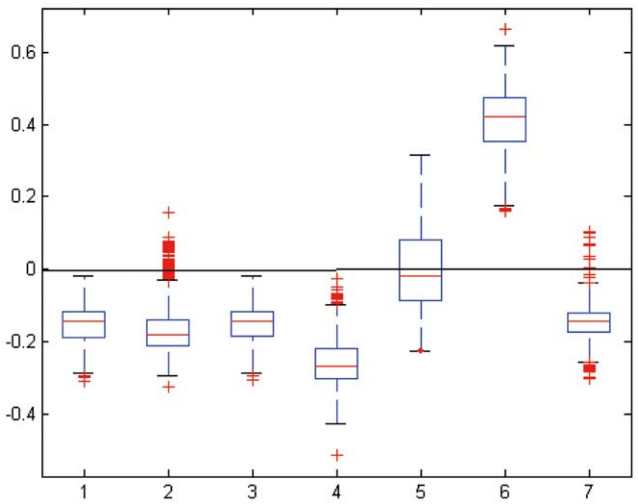

Fig. 1. Bootstrap distributions of standardised b_j ($j = 1, \ldots, 7$).

Fig. 2 shows 95% confidence intervals in the case of both classical Bootstrap (continuous lines) and BC_a (long dotted lines) intervals. By comparing (Table 7) the values of $\alpha_{j,1}$ and $\alpha_{j,2}$ with 0.025 and 0.975 (nominal confidence bounds), it can be seen that the bias-correction and the acceleration factor shift the intervals and may even change the proportion of empirical (bootstrap) distribution comprised in between. Actually, the shift takes into account shape and bias of the empirical distributions and ensures a 95% confidence level.

BC_a confidence interval related to variable $\mathbf{x}_5$, polymer, comprises zero and thus the related coefficient is considered to be not statistically significant.

BC_a confidence intervals after removing variable $\mathbf{x}_5$ (short dotted lines in Fig. 2) show how all remaining variables are significant. Also in this case, both $\alpha_{j,1}$ and $\alpha_{j,2}$ are shifted (Table 7) and the related confidence intervals are by far more accurate than the previous ones.

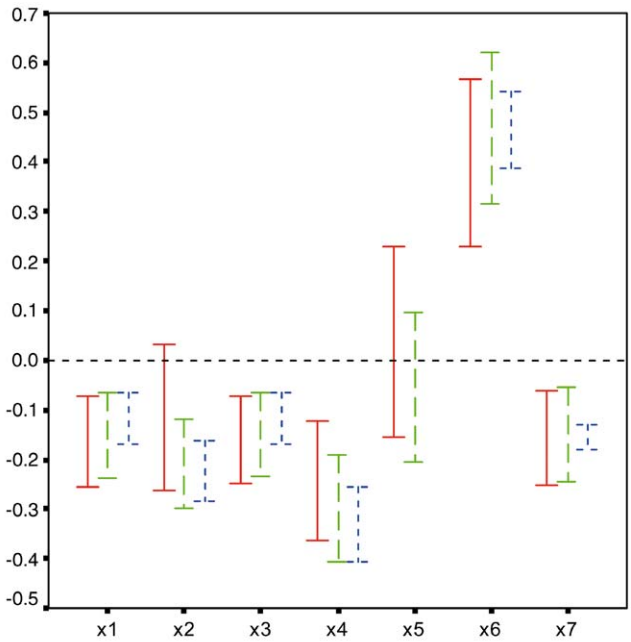

Fig. 2. 95% bootstrap confidence intervals: continuous lines indicate $I_j(\alpha)$'s, long dotted lines indicate $I_j^{\mathrm{BC}_a}(\alpha)$'s, short dotted lines indicate $I_j^{\mathrm{BC}_a}(\alpha)$'s after removing variable $\mathbf{x}_5$.

Table 7
Order of percentiles for BC_a confidence intervals (with and without $\mathbf{x}_5$)

All variables included				After removing $\mathbf{x}_5$		
	$\alpha_{j,1}$,	$\alpha_{j,2}$	Covered proportion of bootstrap distributions	$\alpha_{j,1}$,	$\alpha_{j,2}$	Covered proportion of bootstrap distributions
$\mathbf{x}_1$	0.0397	0.9840	0.9443	0.037	0.983	0.946
$\mathbf{x}_2$	0.0012	0.8077	0.8065	0.011	0.951	0.940
$\mathbf{x}_3$	0.0402	0.9842	0.9440	0.037	0.983	0.946
$\mathbf{x}_4$	0.0027	0.8794	0.8767	0.014	0.959	0.944
$\mathbf{x}_5$	0.0034	0.7832	0.7798		Removed	
$\mathbf{x}_6$	0.1539	0.9982	0.8443	0.057	0.992	0.935
$\mathbf{x}_7$	0.0280	0.9778	0.9498	0.028	0.977	0.949

2.3. The case of logistic regression: Bordeaux wine data

The proposed algorithm for generalised linear regression is now applied to logistic regression by referring to a dataset on Bordeaux wines.

Table 8
Bordeaux wine data

OBS	Year	Temperature	Sunshine	Heat	Rain	Quality
1	1924	3064	1201	10	361	2
2	1925	3000	1053	11	338	3
3	1926	3155	1133	19	393	2
4	1927	3085	970	4	467	3
5	1928	3245	1258	36	294	1
6	1929	3267	1386	35	225	1
7	1930	3080	966	13	417	3
8	1931	2974	1189	12	488	3
9	1932	3038	1103	14	677	3
10	1933	3318	1310	29	427	2
11	1934	3317	1362	25	326	1
12	1935	3182	1171	28	326	3
13	1936	2998	1102	9	349	3
14	1937	3221	1424	21	382	1
15	1938	3019	1230	16	275	2
16	1939	3022	1285	9	303	2
17	1940	3094	1329	11	339	2
18	1941	3009	1210	15	536	3
19	1942	3227	1331	21	414	2
20	1943	3308	1366	24	282	1
21	1944	3212	1289	17	302	2
22	1945	3361	1444	25	253	1
23	1946	3061	1175	12	261	2
24	1947	3478	1317	42	259	1
25	1948	3126	1248	11	315	2
26	1949	3458	1508	43	286	1
27	1950	3252	1361	26	346	2
28	1951	3052	1186	14	443	3
29	1952	3270	1399	24	306	1
30	1953	3198	1259	20	367	1
31	1954	2904	1164	6	311	3
32	1955	3247	1277	19	375	1
33	1956	3083	1195	5	441	3
34	1957	3043	1208	14	371	3

The following variables (Table 8) were measured in 34 years (1924–1957):

Temperature:	Sum of average day temperatures (°C)
Sunshine:	Duration of sunshine (h)
Heat:	Number of very warm days
Rain:	Rain height (mm)
Wine quality:	$1 =$ good, $2 =$ average, $3 =$ poor

Ordinal logistic regression. Ordinal logistic regression of quality on the four *standardised* predictors corresponds to the following model

$$\text{Prob}(\mathbf{y} \leqslant k) = \frac{e^{\alpha_k+\beta_1\text{Temperature}+\beta_2\text{Sunshine}+\beta_3\text{Heat}+\beta_4\text{Rain}}}{1+e^{\alpha_k+\beta_1\text{Temperature}+\beta_2\text{Sunshine}+\beta_3\text{Heat}+\beta_4\text{Rain}}}, \tag{27}$$

where $k = 1, 2$, relates to the first two levels of quality.

It is a proportional odds ratio model that is here accepted by means of a Score test given in Table 9. This table shows also the results from a Proc Logistic run in SAS on the standardised variables. The p-values yielded by the Wald test on the four logistic regression coefficients are, respectively, $0.0573, 0.1046, 0.4568, 0.0361$. Only Temperature and Rain are significant variables at the 10% risk level. The estimation of model (27) allows to compute the probability for a year to be of good, average or poor quality. After assigning a year to the most probable quality, Table 10 is obtained by crossing observed quality with the predicted one. There are seven misclassified years.

Table 9
Logistic regression of quality on the standardised meteorological variables

Score test for the proportional odds assumption Chi-square = 2.9159 with 4 DF ($p = 0.5720$)					
Analysis of maximum likelihood estimates					
Variable	DF	Parameter estimate	Standard error	Wald chi-square	Pr > chi-square
INTERCP1	1	−2.6638	0.9266	8.2641	0.0040
INTERCP2	1	2.2941	0.9782	5.4998	0.0190
TEMPERA	1	3.4268	1.8029	3.6125	0.0573
SUNSHINE	1	1.7462	1.0760	2.6335	0.1046
HEAT	1	−0.8891	1.1949	0.5536	0.4568
RAIN	1	−2.3668	1.1292	4.3931	0.0361

Table 10
Prediction quality of model (27) by using classical logistic regression

Quality observed counts	Predicted 1	2	3	Total
1	8	3	0	11
2	2	8	1	11
3	0	1	11	12
Total	10	12	12	34

PLS ordinal logistic regression. In the Bordeaux wine data, the multicollinearity between predictors leads to two difficulties: first, influent variables such as Sunshine and Heat become nonsignificant in model (27) even though they are significant when taken one at a time; second, the variable Heat shows a negative coefficient in the model while it has a positive influence on quality in reality.

Generally, PLS logistic regression allows to obtain a coherent model at the level of coefficients while retaining all predictors. Moreover, it is equally feasible when predictors have missing data. Separate logistic regressions of quality on each standardised predictor yields the coefficients a_{1j} of Temperature, Sunshine, Heat and Rain

equal to, respectively, 3.0117 (0.0002), 3.3401 (0.0002), 2.1445 (0.0004) and −1.7906 (0.0016), with the p-values given in parentheses. These coefficients are all significant with coherent signs.

After normalizing the coefficients, the first PLS component is defined as

$$\mathbf{t}_1 = \frac{3.0117 \times \text{Temperature} + 3.3401 \times \text{Sunshine} + 2.1445 \times \text{Heat} - 1.7906 \times \text{Rain}}{\sqrt{(3.0117)^2 + (3.3401)^2 + (2.1445)^2 + (-1.7906)^2}},$$

$$= 0.5688 \times \text{Temperature} + 0.6309 \times \text{Sunshine} + 0.4050 \times \text{Heat} - 0.3382 \times \text{Rain}.$$

The results from logistic regression of quality on component $\mathbf{t}_1$ are given in Table 11. It is satisfactory to notice that there are only 6 misclassified years, i.e. one misclassified year less than in classical logistic regression.

Table 11
Results from logistic regression of quality on component $\mathbf{t}_1$

Analysis of parameter	Maximum DF	Likelihood estimate	Estimates std.error	Chi-square	Pr > Chi sq
Intercept	1	−2.2650	0.8644	6.8662	0.0088
Intercept2	1	2.2991	0.8480	7.3497	0.0067
$t1$	1	2.6900	0.7155	14.1336	0.0002

Cross-table of observed and predicted quality

Quality observed Counts	Predicted 1	2	3	Total
1	9	2	0	11
2	2	8	1	11
3	0	1	11	12
Total	11	11	12	34

In order to search for the variables whose contribution to the second component $\mathbf{t}_2$ is significant, logistic regressions of quality on $\mathbf{t}_1$ and each standardised predictor $\mathbf{x}_j^*$ are built

$$\text{Prob}(\mathbf{y} \leqslant k) = \frac{e^{\alpha_k + \beta_1 \mathbf{t}_1 + \beta_{2j} \mathbf{x}_j^*}}{1 + e^{\alpha_k + \beta_1 \mathbf{t}_1 + \beta_{2j} \mathbf{x}_j^*}}. \tag{28}$$

The following p-values are obtained for the predictors coefficients: 0.6765, 0.6027, 0.0983, 0.2544. Therefore, the second PLS component is not significant and, as a consequence, the model with only one component is retained.

By expressing the component $\mathbf{t}_1$ in terms of the standardized variables Temperature, Sunshine, Heat and Rain, the final estimates of the parameters in model (27) are more coherent than the ones earlier obtained by classical logistic regression

$$\text{Prob}(\mathbf{y} = 1) = \frac{e^{-2.265 + 1.53 \times \text{Temperature} + 1.70 \times \text{Sunshine} + 1.09 \times \text{Heat} - 0.91 \times \text{Rain}}}{1 + e^{-2.5265 + 1.53 \times \text{Temperature} + 1.70 \times \text{Sunshine} + 1.09 \times \text{Heat} - 0.91 \times \text{Rain}}},$$

and

$$\text{Prob}(\mathbf{y} \leqslant 2) = \frac{e^{2.2991+1.53\times\text{Temperature}+1.70\times\text{Sunshine}+1.09\times\text{Heat}-0.91\times\text{Rain}}}{1+e^{2.2991+1.53\times\text{Temperature}+1.70\times\text{Sunshine}+1.09\times\text{Rain}-0.91\times\text{Rain}}}.$$

Validation issues for PLS logistic regression: *randomised and balanced bootstrap*. In order to derive an empirical distribution for the parameters of PLS Logistic Regression, a simpler variant of bootstrap is adopted with the name of randomised bootstrap. It was earlier proposed in (Efron, 1983) for the case of a binary ($K = 2$) categorical response variable $\mathbf{y}$.

Let $\hat{F}$ be the empirical distribution of $(y_i, \mathbf{x}_i)$; $\hat{F}$ puts equal mass $1/n$ on each observed unit. Efron (1983) suggests two possible modifications of $\hat{F}$:

case 1: $\pi_i(\bar{y}_i, \mathbf{x}_i) = 0.1$ (empirical assessment of the probability) where $\pi_i(\bar{y}_i, \mathbf{x}_i)$ represents the probability of the complementary units to $(y_i, \mathbf{x}_i)$;

case 2: $\pi_i(y_i, \mathbf{x}_i) = \frac{e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}$ (probability based on predictions from logistic regression).

Let $\hat{F}^{(\text{RAND})}$ be the modified empirical distribution. The bootstrap estimates are then based on the Monte Carlo approximation of the parameters bootstrap distributions.

An extension of "randomised" bootstrap to PLS Logistic Regression (generalised to the case where $K > 2$) is the following.

Let $\pi_i(y_i = k \mid \mathbf{x}_i)$ be the probability to observe $y_i = k$ $(k = 1, 2, \ldots, K)$ given a row vector $\mathbf{x}_i$ for the explanatory variables. These probabilities can be then derived in two different alternative ways: take the conditional probability estimates from classical or PLS logistic regression model or refer to the empirical distribution of $\mathbf{y}$.

Let B be the number of resamples, "randomised" boostrap performs the following steps:

Step 1: For each $b = 1, 2, \ldots, B$, let $X^{(b)}$ be the bth resample made by n i.i.d. random variables drawn from the explanatory variables part of $\hat{F}$. In order to properly assign a value $y_i^{(b)} = k$ $(k = 1, 2, \ldots, K)$ to each vector $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$, a pseudo random number is drawn from $\pi_i(y_i = k \mid \mathbf{x}_i)$,

Step 2: For each resample $(y_i^{(b)}, x_i^{(b)})$, the estimate $\hat{\beta}^{(b)}$ of PLS logistic regression coefficients is yielded,

Step 3: For each explanatory variable $\mathbf{x}_j$ $(j = 1, \ldots, 4)$ the Monte Carlo approximation $\Phi^{(B)}_{\hat{F}^{(\text{RAND})}, \hat{\beta}_j}$ of the bootstrap distribution of $\hat{\beta}_j$ is obtained.

Finally, the percentiles of $\Phi^{(\text{B})}_{\hat{F}^{(\text{RAND})}, \hat{\beta}_j}$ may be used for building up confidence intervals for β_j.

It is worth noticing that efficiency of bootstrap computation can be easily improved by means of balanced bootstrap (or permutation bootstrap (Efron and Tibshirani, 1993)), a simple modification of the ordinary bootstrap sampling schema. In the procedure of randomised bootstrap for PLS logistic regression, the balanced bootstrap (Tenenhaus, 2002) assures that each row $\mathbf{x}_i$ appears B times in the set of nB resampled rows. This is easily accomplished by running the following procedure:

(1) Stack B copies of $\mathbf{X}$ into a matrix, say $\boldsymbol{\Upsilon}$ with nB rows,
(2) Take a random permutation of $\boldsymbol{\Upsilon}$,

(3) The first bootstrap sample consists in the first n rows of Υ; the second bootstrap sample in the second set of n rows of Υ and so on until the Bth set coinciding with the last n rows of Υ.

The nonparametric validation procedure suggested above for PLS Logistic Regression has been applied to Bordeaux wine data (Fig. 3) with $B = 1000$. Actually, 12 resamples out of 1000 were unuseful due to divergence toward infinite of the coefficient estimates. This is a very well-known problem in literature and usually appears when a complete (or almost complete) separation of data is achieved in the bootstrap resample. In Moulton and Zager (1991) a modified estimation algorithm for logistic regression is proposed in order to overcome the problem. However, this algorithm was not worth using here because of the very small number of problematic samples. Finally, due to the strong significance enhanced by confidence intervals in Fig. 3, it was not even worth computing the BC_a intervals discussed earlier.

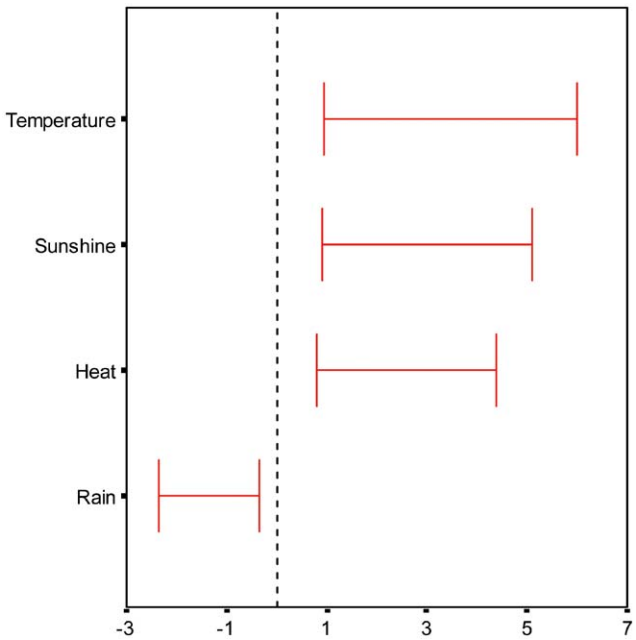

Fig. 3. 95% randomized/balanced bootstrap confidence intervals for Bordeaux wine data.

2.4. The case of censored data: determining risk factors associated with the age when first white hair appears in males

2.4.1. Data presentation

Started in 1994 by Prof. Serge Hercberg, the epidemiological study SU.VI.MAX (*SUpplémentation en VItamines et Minéraux Antioxidants*) aimed to evaluate the

Table 12
Data pre-treatment and construction of the response function

I	t	$\mathbf{y}_{it}$
Case 1	$\leqslant 30$	0
Case 1	31–35	0
Case 1	36–40	1
Case 2	$\leqslant 30$	0
Case 2	31–35	0
Case 2	36–40	0
Case 2	41–45	0

nutritional state of the French population and study the influence of an oral supplementation of vitamins and anti-oxidant minerals on a set of health indicators: heart attack, cardiovascular illnesses, cancers, where free radicals are strongly involved (Hercberg et al., 1998). This study was meant to finish after 8 years (in 2003) and comprises more than 12,000 volunteers between 25 and 65 years old, constituting a representative sample of the French population. Half of them were given an anti-oxidant treatment, the others a placebo.

Starting from this cohort, L'Oréal (a French cosmetics company) run a study on the health state of hair and fingernails on 10,323 cases (4057 males and 6266 females). On the basis of responses to a questionnaire covering more than 150 items, an attempt was made to demonstrate the risk factors associated with the premature onset of greying hair in men.

2.4.2. Variables description

The age of appearance of the first white hair in males is the response variable. It is a discrete variable with 6 categories: up to 30 years old, 31–35 years, 36–40 years, 41–45 years, 45–50 years, more than 50 years old.

The needed background for analysing this type of data is given in Allison (1995). Each case, say i, is duplicated m_i times where m_i is the number of his/her age class at the time of appearance of the first white hair or his/her age class at present if no white hair has appeared yet. The response variable $\mathbf{y}_{it}$ is then defined. It is equal to 0 if the ith case has no white hair at age t and 1 otherwise. For the last record in the list ($t = m_i$), $\mathbf{y}_{it}$ is equal to 1 if the event occurred, and 0 otherwise.

For instance, Table 12 shows two cases: case 1 is a 45-year-old male whose first white hair appeared at the age of 38 (3 records for this case); case 2 is a 45-year-old male with no white hair (4 records for this case).

The risk factors that have been selected from the SUVIMAX database are the following:

Previous appearances in the family (*dummy variables*: 1 = *yes*, 0 = *no*):

- The mother had white hair before being 30 years old ($\mathbf{x}_1$)
- The sister had white hair before being 30 years old ($\mathbf{x}_2$)
- The mother had no white hair after being 60 years old ($\mathbf{x}_3$)

- The father had white hair before being 30 years old ($\mathbf{x}_4$)
- The father had no white hair after being 60 years old ($\mathbf{x}_5$)

Hair features

- Thickness (very thin, thin, medium, thick)
- Natural color (red, blond, light brown, brown, dark brown, black)

Only significant factors have been retained for the application. Moreover, for a comparison between classical generalised linear model (McCullagh and Nelder, 1989) and PLS-GLR, only the cases with no missing data on the selected risk factors are considered.

2.4.3. The model

Let P_{it} be the probability that the first white hair for case i appeared at time t knowing that it has not appeared at previous times $1, 2, \ldots, t-1$. Then

$$P_{it} = \text{Prob}(\mathbf{y}_{it} = 1/\mathbf{y}_{i1} = 0, \ldots, \mathbf{y}_{i,t-1} = 0) \tag{29}$$

is the conditional probability of observing "$\mathbf{y}_{it} = 1$" knowing that "$\mathbf{y}_{i1} = 0, \ldots, \mathbf{y}_{i,t-1} = 0$".

Let us suppose that the underlying continuous model is a Cox model (Cox, 1972) with proportional risks

$$\begin{aligned}\text{Log}[h_i(t)/h_0(t)] &= \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \beta_5\mathbf{x}_5 \\ &\quad + \beta_6\text{Red} + \beta_7\text{Blond} + \beta_8\text{Light Brown} + \beta_9\text{Brown} \\ &\quad + \beta_{10}\text{Dark Brown} + \beta_{11}\text{Very Thin} + \beta_{12}\text{Thin} + \beta_{13}\text{Medium} \\ &= \sum_{j=1}^{13} \beta_j\mathbf{v}_j,\end{aligned} \tag{30}$$

where $\mathbf{v}_j$ relates to any predictor indiscriminately.

A generalised linear model with the "Complementary Log Log" (Prentice and Gloeckler, 1978) link function may be used for modeling the response variable in function of the predictors

$$\text{Log}[-\text{Log}(1 - P_{it})] = \alpha_t + \sum_j \beta_j\mathbf{v}_j. \tag{31}$$

The interpretation of the regression coefficients in (31) in terms of hazard ratios coincides with the one related to the underlying model (30). For instance, appearance of the first white hair in the mother before being 30 years old increases the risk by $100(\exp(\beta_1) - 1)\%$, once the other variables are fixed.

The particular form of the response vector associated to the same case (all values are zero if the case is censored, all values are zero but the last one that is equal to 1 if the case is not censored) allows a factorisation of the likelihood by means of P_{it}

conditional probabilities:

$$
\begin{aligned}
L &= \prod_{i=1}^{N} \text{prob}(\mathbf{y}_{i1}=0,\mathbf{y}_{i2}=0,\ldots,\mathbf{y}_{it_i-1}=0,\mathbf{y}_{it_i}=1)\\
&= \prod_{i=1}^{N} \text{prob}(\mathbf{y}_{it_i}=1/\mathbf{y}_{i1}=0,\ldots,\mathbf{y}_{it_i-1}=0)\times \text{prob}(\mathbf{y}_{i1}=0,\ldots,\mathbf{y}_{it_i-1}=0)\\
&= \prod_{i=1}^{N} P_{it_i}(1-P_{it_i-1})\times\cdots\times(1-P_{i1})\\
&= \prod_{i=1}^{N}\prod_{l\leqslant t_i} P_{il}^{\mathbf{y}_{il}}(1-P_{il})^{1-\mathbf{y}_{il}}. \qquad (32)
\end{aligned}
$$

It is all set as if the responses $\mathbf{y}_{il}$ were independently distributed as Bernoulli probability distributions with parameter P_{il}. Model (31) is a specific generalised linear model that is going to be studied by means of Proc GENMOD in SAS and PLS-GLR.

2.4.4. Generalised linear model

Table 13 reports the results from Proc GENMOD in SAS version 8.2 applied to model (31).

Remarks

(1) The ratio between deviance and degrees of freedom is less than 1 and tends to validate the model goodness of fit to data.
(2) Conditional probability of first white hair appearance increases with age. This is the statistical translation of what is a natural phenomenon in the studied population.
(3) Results concerning the family history were as expected.
(4) The risk of appearance of white hair increases with hair thickness.
(5) The risk of appearance of white hair increases with the darkness of the natural hair colour.
(6) The red colour is located between light brown and brown. This result confirms the peculiarity of this colour from the bio-chemical point of view and underlines the need to treat this factor as categorical.

2.4.5. PLS generalised linear model

In Marx (1996) the parameters of a generalised linear regression are estimated by introducing PLS regression in the iterated weighted least-squares algorithm used for likelihood maximisation. The approach proposed in the following is much simpler, its implementation needs a little and easy programming and it can be easily generalised to all models that are linear at the level of explanatory variables.

2.4.5.1. Computation of the first PLS component. A generalised linear model is built with age and each predictor separately taken. Categorical variables are considered

Table 13
GENMOD output

Criteria for assessing goodness of fit

Criterion	DF	Value	Value/DF
Deviance	6147	5161.5374	0.8397
Scaled Deviance	6147	5161.5374	0.8397
Pearson Chi-Square	6147	5761.4509	0.9373
Scaled Pearson X2	6147	5761.4509	0.9373
Log Likelihood		−2580.7687	

LR statistics for type 3 analysis

Source	DF	Chi-square	Pr > Chi sq
Age	5	1448.91	<.0001
X1	1	14.61	0.0001
X2	1	19.97	<.0001
X3	1	30.20	<.0001
X4	1	45.62	<.0001
X5	1	23.95	<.0001
Thickness	3	16.75	0.0008
Color	5	31.12	<.0001

Analysis of parameter estimates

Parameter	DF	Estimate	Standard error	Wald 95% confidence limits		Chi-square	Pr > Chi Sq
Intercept	1	1.3745	0.1434	1.0935	1.6556	91.89	<.0001
<= 30	1	−3.6202	0.1345	−3.8837	−3.3566	724.63	<.0001
31–35	1	−3.1961	0.1233	−3.4378	−2.9544	671.75	<.0001
36–40	1	−2.2025	0.1039	−2.4061	−1.9988	449.09	<.0001
41–45	1	−1.3685	0.0974	−1.5595	−1.1776	197.30	<.0001
46–50	1	−0.9598	0.1014	−1.1586	−0.7610	89.54	<.0001
> 50	1	0.0000	0.0000	0.0000	0.0000	—	—
X1	1	0.4977	0.1244	0.2539	0.7415	16.01	<.0001
X2	1	0.6281	0.1330	0.3675	0.8887	22.31	<.0001
X3	1	−0.4532	0.0865	−0.6227	−0.2837	27.46	<.0001
X4	1	0.8153	0.1102	0.5994	1.0312	54.77	<.0001
X5	1	−0.4323	0.0928	−0.6141	−0.2505	21.72	<.0001
Very thin	1	−0.4598	0.1157	−0.6865	−0.2330	15.80	<.0001
Thin	1	−0.3518	0.0982	−0.5442	−0.1594	12.84	0.0003
Medium	1	−0.2791	0.0975	−0.4701	−0.0881	8.20	0.0042
Thick	0	0.0000	0.0000	0.0000	0.0000	—	—
Red	1	−0.2751	0.3141	−0.8908	0.3405	0.77	0.3811
Blond	1	−0.6806	0.1452	−0.9652	−0.3959	21.96	<.0001
Light brown	1	−0.3959	0.1055	−0.6027	−0.1891	14.07	0.0002
Brown	1	−0.3141	0.1032	−0.5165	−0.1118	9.26	0.0023
Dark brown	1	−0.1310	0.1120	−0.3506	0.0886	1.37	0.2422
Black	0	0.0000	0.0000	0.0000	0.0000	—	—

as a whole in the model. All variables are significant (Table 14). This result is not surprising as only significant risk factors have been retained for this application. Therefore, all variables contribute to the construction of the first component $\mathbf{t}_1$ (Table 15).

Table 14
Generalised linear regression of **y** on age and each predictor

Variables	Wald		
	DoF	Chi-square	p-value
$\mathbf{x}_1$	1	42.53	<0.0001
$\mathbf{x}_2$	1	59.76	<0.0001
$\mathbf{x}_3$	1	35.16	<0.0001
$\mathbf{x}_4$	1	74.05	<0.0001
$\mathbf{x}_5$	1	31.93	<0.0001
Thickness	3	27.87	<0.0001
Colour	5	46.93	<0.0001

Table 15
Generalised linear regression of **y** on age and each dummy or categorical predictor

Variables ($\mathbf{v}_j$)	Regression coefficients (a_{1j})
$\mathbf{x}_1$	0.73
$\mathbf{x}_2$	0.92
$\mathbf{x}_3$	−0.50
$\mathbf{x}_4$	0.93
$\mathbf{x}_5$	−0.51
Very thin	−0.56
Thin	−0.38
Medium	−0.28
Thick	0.00
Red	−0.41
Blond	−0.80
Light brown	−0.49
Brown	−0.40
Dark brown	−0.19
Black	0.00

2.4.5.2. Computation of the second PLS component. A generalised linear model is built with age, $\mathbf{t}_1$, and each predictor separately taken. Table 16 shows how all variables are not significant. Only the first PLS component $\mathbf{t}_1$ is retained.

2.4.5.3. Generalised linear model on $\mathbf{t}_1$. Table 17 reports the results from generalised linear regression McCullagh and Nelder (1989) of **y** on age and $\mathbf{t}_1$.

Remark

Table 18 shows how the goodness of fit criteria for the cloglog and the PLS cloglog models are really similar. They lead to accept these models as the values of deviance and Pearson's Chi-square are close to the degrees of freedom.

Table 16
Generalised linear regression of $\mathbf{y}$ on age, $\mathbf{t}_1$ and each predictor

Variables	Wald		
	DoF	Chi-square	p-value
$\mathbf{x}_1$	1	1.23	0.27
$\mathbf{x}_2$	1	1.63	0.20
$\mathbf{x}_3$	1	0.41	0.52
$\mathbf{x}_4$	1	0.48	0.49
$\mathbf{x}_5$	1	0.26	0.61
Thickness	3	0.41	0.94
Colour	5	0.26	0.99

Table 17
Generalised linear regression of $\mathbf{y}$ on age and $\mathbf{t}_1$

Criteria for assessing goodness of fit			
Criterion	DF	Value	Value/DF
Deviance	6159	5164.8371	0.8386
Scaled deviance	6159	5164.8371	0.8386
Pearson chi-square	6159	5779.5776	0.9384
Scaled pearson X2	6159	5779.5776	0.9384
Log likelihood		−2582.4185	

Analysis of parameter estimates							
Parameter	DF	Estimate	Standard error	Wald 95% confidence limits		Chi-square	Pr > Chi Sq
< = 30	1	−2.9125	0.1041	−3.1165	−2.7085	783.28	<0.0001
31–35	1	−2.4855	0.0897	−2.6614	−2.3096	767.06	<0.0001
36–40	1	−1.4883	0.0610	−1.6078	−1.3688	595.66	<0.0001
41–45	1	−0.6544	0.0498	−0.7521	−0.5567	172.38	<0.0001
46–50	1	−0.2500	0.0580	−0.3637	−0.1363	18.56	<0.0001
> 50	1	0.7060	0.0829	0.5436	0.8684	72.59	<0.0001
$\mathbf{t}_1$	1	1.6864	0.1036	1.4833	1.8894	264.98	<0.0001

Table 18
Cloglog and PLS cloglog models goodness of fit

	Cloglog model			PLS cloglog model		
	Value	DoF	Value/DoF	Value	DoF	Value/DoF
Deviance	5162	6147	0.84	5165	6159	0.84
Pearson's Chi-square	5761	6147	0.94	5780	9159	0.94

2.4.5.4. PLS model in terms of the original variables. PLS generalised linear model may be expressed in function of the original variables by using the expression $\mathbf{t}_1$. Estimates of coefficients in cloglog and PLS cloglog models, as well as their 95% confidence intervals (Wald for classical model and bootstrap for PLS model) are very close to each other (Table 19).

Table 19
Coefficients of cloglog model and PLS cloglog model with 1 component in terms of the original variables

Variables	Cloglog				Cloglog PLS			
	Parameter estimates	Standard error	Hazard ratio	Wald 95% CI	Parameter estimates	Standard error	Hazard ratio	Bootstrap 95% CI
$\leqslant 30$	−3.62	0.13	0.03	0.02–0.03	−3.65	0.15	0.03	0.02–0.03
31–35	−3.20	0.12	0.04	0.03–0.05	−3.22	0.13	0.04	0.03–0.05
36–40	−2.20	0.10	0.11	0.09–0.14	−2.21	0.11	0.11	0.09–0.13
41–45	−1.37	0.10	0.25	0.21–0.31	−1.38	0.09	0.25	0.21–0.30
46–50	−0.96	0.10	0.38	0.31–0.47	−0.97	0.10	0.38	0.30–0.46
> 50	0.00	—	—	—	0.00	0.00	—	—
$\mathbf{x}_1$	0.50	0.12	1.65	1.28–2.10	0.59	0.12	1.80	1.45–2.22
$\mathbf{x}_2$	0.63	0.13	1.88	1.45–2.44	0.74	0.12	2.09	1.72–2.59
$\mathbf{x}_3$	−0.45	0.09	0.64	0.54–0.76	−0.39	0.06	0.67	0.58–0.75
$\mathbf{x}_4$	0.82	0.11	2.27	1.82–2.80	0.74	0.12	2.11	1.79–2.97
$\mathbf{x}_5$	−0.43	0.09	0.65	0.54–0.78	−0.40	0.07	0.67	0.58–0.77
Very thin	−0.46	0.12	0.63	0.50–0.79	−0.44	0.09	0.64	0.54–0.75
Thin	−0.35	0.10	0.70	0.58–0.85	−0.30	0.08	0.74	0.63–0.87
Medium	−0.28	0.10	0.76	0.63–0.91	−0.22	0.08	0.80	0.70–0.93
Thick	0.00	—	—	—	0.00	0.00	—	—
Red	−0.28	0.31	0.76	0.41–1.40	−0.30	0.24	0.74	0.49–1.19
Blond	−0.68	0.15	0.51	0.38–0.67	−0.64	0.13	0.53	0.42–0.64
Light brown	−0.40	0.11	0.67	0.55–0.83	−0.39	0.09	0.68	0.56–0.79
Brown	−0.31	0.10	0.73	0.59–0.90	−0.32	0.10	0.73	0.62–0.84
Dark brown	−0.13	0.11	0.88	0.70–1.09	−0.15	0.11	0.86	0.72–1.03
Black	0.00	—	—	—	0.00	—	—	—

Remarks

(1) Confidence intervals are obtained by means of the "Balanced bootstrap" resampling method (Davison et al., 1986) as it performs better than classical "Uniform bootstrap" in reducing the simulation error.
(2) Confidence intervals are estimated by referring to the percentiles of the bootstrap empirical distribution, however, the use of the bootstrap estimate for the coefficients standard deviation would yield very comparable results.
(3) A PLS model with two components yields the same estimates as the classical model with a precision up to the second decimal digit.

2.4.6. PLS Cox model

The Cox model (30) could also be used with discrete data (here, the age when white hair appears takes values $t = 1$–6). Efron approximation is used for taking into

account ties. As for model (31) on complementary log log, only one PLS component is retained. Table 20 reports results from classical Cox model and PLS Cox.

Table 20
Coefficients of classical Cox model and PLS Cox model with 1 component in terms of the original variables

Variables	Cox				PLS Cox			
	Parameter estimates	Standard error	Hazard ratio	Wald 95% CI	Parameter estimates	Standard error	Hazard ratio	Bootstrap 95% CI
$\mathbf{x}_1$	0.47	0.12	1.60	1.27–2.02	0.55	0.11	1.72	1.44–2.16
$\mathbf{x}_2$	0.58	0.13	1.78	1.39–2.29	0.69	0.11	1.99	1.62–2.36
$\mathbf{x}_3$	−0.43	0.08	0.65	0.55–0.77	−0.37	0.06	0.69	0.61–0.77
$\mathbf{x}_4$	0.76	0.11	2.15	1.75–2.64	0.69	0.10	2.01	1.64–2.46
$\mathbf{x}_5$	−0.41	0.09	0.67	0.56–0.79	−0.39	0.07	0.68	0.61–0.77
Very thin	−0.43	0.11	0.65	0.52–0.81	−0.42	0.08	0.66	0.57–0.79
Thin	−0.33	0.10	0.72	0.60–0.87	−0.29	0.07	0.75	0.66–0.85
Medium	−0.26	0.09	0.77	0.64–0.93	−0.21	0.07	0.81	0.70–0.95
Thick	0.00	—	—	—	0.00	—	—	—
Red	−0.26	0.30	0.77	0.42–1.39	−0.30	0.21	0.74	0.46–1.11
Blond	−0.65	0.14	0.52	0.40–0.69	−0.61	0.12	0.54	0.43–0.69
Light brown	−0.38	0.10	0.69	0.56–0.84	−0.37	0.09	0.69	0.56–0.81
Brown	−0.30	0.10	0.74	0.61–0.90	−0.30	0.09	0.74	0.62–0.89
Dark brown	−0.13	0.11	0.88	0.71–1.09	−0.15	0.10	0.86	0.71–1.06
Black	0.00	—	—	—	0.00	—	—	—

Remarks

(1) As it could be expected, results from Cox model and from cloglog model are very similar.
(2) Here, the same remarks previously done in comparing cloglog model with PLS cloglog apply.
(3) Coefficients are estimated by maximising partial likelihood (Cox, 1972).
(4) The Cox model is not strictly speaking a generalised linear model McCullagh and Nelder (1989). PLS generalised linear regression can also be extended beyond the exponential family.
(5) The Cox model does not need the duplication of patient data as for the cloglog model.

In Nguyen and Rocke (2002), the so-called Partial Least Squares Proportional Hazard Regression is proposed for the application to gene expression data from DNA microarrays. Their proposal actually consists of a two-stage strategy of analysis: PLS regression at the first stage in order to reduce data dimensionality and extract PLS components; Proportional Hazard regression model at the second stage in order to estimate survival distribution. This two-stage strategy does not take into account the censoring information in the estimation of PLS components, thus inducing bias in their

estimates. The approach proposed above goes further than that as it is meant to use PLS criterion inside generalised linear regression model, so as to enrich its interpretation and enlarge its feasibility to peculiar and difficult situations often encountered in real practice.

3. Conclusion

The PLS regression algorithm that has been re-formulated above shows several advantages:

(1) Classical PLS regression is directly linked to the usual procedures for simple and multiple regressions and is therefore enriched by the classical testing procedures of such methods. In this framework, the statistical tests aim at identifying those explanatory variables that do not significantly contribute to the construction of PLS components and, consequently, with a low explanatory power on the response variable. A PLS components is judged as not significant when there is no explanatory variable with a significant weight. In its construction. In the example of Section 2, the proposed approach led to the same variable selection as backward stepwise classical PLS regression. The approach might be further validated on a wider variety of examples also in comparison with other approaches to variable selection in PLS recently proposed in Forina et al. (1999), Gauchi and Chagnon (2001), Höskuldsson (2001), Lazraq et al. (2003), Lingren et al. (1994) and Sarabia et al. (2001).
(2) In practice, when a strong degree of multicollinearity shows up, stepwise multiple regression is commonly used. The inconvenience of this method is the elimination of explanatory variables strongly correlated to the response variable and thus important for the user. On the contrary, PLS regression allows to retain in the model all variables with a stronger explanatory power.
(3) In case of missing data, PLS components are computed in accordance with the NIPALS algorithm. However, PLS components are correlated in this case. The original PLS regression algorithm does not consider this feature. On the contrary, the new formulation let the correlation between PLS components play a role as multiple regression is used.
(4) There is an immediate extension to generalised linear regression. Some preliminary results were already obtained in PLS logistic regression Esposito Vinzi and Tenenhaus (2001) and in survival data with the Cox PLS model Bastien and Tenenhaus (2001).

Acknowledgements

We are very grateful to Silvano Amato for his very valid help in setting up the bootstrap-based validation procedures presented in the paper.

References

Allison, P.D., 1995. Survival Analysis Using the SAS System: A Practical Guide. SAS Inc, Cary, NC, USA.

Bastien, P., Tenenhaus, M., 2001. PLS generalised linear regression. Application to the analysis of life time data. In: Esposito Vinzi, V., Lauro, C., Morineau, A., Tenenhaus, M. (Eds.), PLS and Related Methods, Proceedings of the PLS'01 International Symposium, CISIA-CERESTA, Paris, pp. 131–140.

Camo, ASA., 2003. The Unscrambler® 8.0, CAMO ASA, Nedre Vollgate 8, 0158 Oslo, Norway.

Cox, D.R., 1972. Regression models and life-tables (with discussion). J. Roy. Statist. Soc., Series B 34, 187–220.

Davison, A.C., Hinkley, D.V., Schechtman, E., 1986. Efficient bootstrap simulations. Biometrika 73, 555–566.

Efron, B., 1983. Estimating the error rate of a prediction rule: some improvements on cross-validation. J. Amer. Statist. Assoc. 78, 316–331.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Esposito Vinzi, V., Tenenhaus, M., 2001. PLS Logistic Regression. In: Esposito Vinzi, V., Lauro, C., Morineau, A., Tenenhaus, M. (Eds.), PLS and Related Methods, Proceedings of the PLS'01 International Symposium, CISIA-CERESTA, Paris, pp. 117–130.

Forina, M., Casolino, C., Pizarro Millan, C., 1999. Iterative predictor weighting PLS (IPW): a technique for the elimination of useless predictors in regression problems. J. Chemometrics 13, 165–184.

Garthwaite, P.H., 1994. An interpretation of partial least squares. J. Amer. Statist. Assoc. 89 (425), 122–127.

Gauchi, J.-P., Chagnon, P., 2001. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. Chemometrics & Intelligent Laboratory Systems 58 (2), 171–193.

Hercberg, S., Preziosi, P., Briancon, S., Galan, P., Triol, I., Malvy, D., Roussel, A.M., Favier, A., 1998. A primary prevention trial of nutritional doses of antioxydant vitamins and minerals on cardiovascular diseases and cancers in general populations: the SU.VI.MAX Study—design, methods and participant characteristics. Supplementation en Vitamines et Mineraux AntioXydants. Control Clin. Trials 19 (4), 336–351.

Höskuldsson, A., 2001. Variable and subset selection in PLS regression. Chemometrics & Intelligent Laboratory Systems 55, 23–38.

Kettaneh-Wold, N., 1992. Analysis of mixture data with partial least squares. Chemometrics & Intelligent Laboratory Systems 14, 57–69.

Lazraq, A., Cléroux, R., Gauchi, J.P., 2003. Selecting both latent and explanatory variables in the PLS1 regression model. Chemometrics and Intelligent Laboratory Systems 66, 117–126.

Lingren, F., Geladi, P., Rannar, S., Wold, S., 1994. Interactive variable selection (IVS) for PLS. Part 1: theory and algorithms. J. Chemometrics 8, 349–363.

Marx, B.D., 1996. Iteratively reweighted partial least squares estimation for generalized linear regression. Technometrics 38 (4), 374–381.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd Edition. Chapman & Hall, London.

Moulton, L., Zager, S., 1991. Bootstrapping generalized linear models. Computat. Statist. Data Anal. 11, 53–63.

Nguyen, D.V., Rocke, D.M., 2002. Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics 18 (12), 1625–1632.

Prentice, R.L., Gloeckler, L.A., 1978. Regression analysis of grouped survival data with application to breast cancer data. Biometrics 34, 57–67.

Sarabia, L.A., Ortiz, M.C., Sánchez, M.S., Herrero, A., 2001. Dimension-wise selection in partial least squares regression with a bootstrap estimated signal-noise relation to weight the loadings. In: Esposito Vinzi, V., Lauro, C., Morineau, A., Tenenhaus, M. (Eds.), PLS and Related Methods, Proceedings of the PLS'01 International Symposium, CISIA-CERESTA, Paris, pp. 327–339.

Shao, J., Tu, D., 1995. The Jackknife and the Bootstrap. Springer, New York.

Shenk, J.S., Westerhaus, M.O., 1991. Population structuring of near infrared spectra and modified partial least squares regression. Crop Science 31, 1548–1555.

Tenenhaus, M., 1998. La régression. PLS Technip, Paris.

Tenenhaus, A., 2002. Régression Logistique PLS validée par bootstrap. Mémoire de DEA de Statistique, Université Pierre et Marie Curie, Paris and L'Oréal, Aulnay-sous-Bois.

Umetri, AB., 2004. SIMCA-P 10.5, User Guide and Tutorial. Umetri AB, Box 7960, S-90719 Umeå, Sweden.

Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration problem in chemistry solved by the PLS method. In: Ruhe, A., Kåstrøm, B. (Eds.), Proceedings of the Conference Matrix Pencils, March 1982, Lecture Notes in Mathematics, Springer, Heidelberg, pp. 286–293.