# Calcul des coefficients de régression et du PRESS en régression PLS1

Marie Chavent, Brigitte Patouille

Mathématiques Appliquées de Bordeaux (UMR 5466)
Universtité Bordeaux 1,
351 cours de la libération,
33405 Talence cedex

Email : chavent@math.u-bordeaux1.fr, be@sm.u-bordeaux2.fr

## 1 Introduction

Lors de l'implémentation sous Splus de l'algorithme de régression PLS1 tel qu'il est présenté dans [Tenenhaus, 1998], nous avons utilisé une formule de récurrence simple pour le calcul des coefficients de régression et précisé certains choix nécessaires pour le calcul du PRESS. Nous allons rappeler dans un premier temps l'algorithme PLS1 afin d'introduire les notations, puis donner dans un second temps la formule de récurrence permettant d'implémenter le calcul des coefficients de régression. Enfin, nous montrerons dans une dernière partie les différentes alternatives possibles pour le calcul du PRESS, et la solution adoptée. Cette solution permet en effet de retrouver certains résultats numériques présentés dans [Tenenhaus, 1998] et obtenus avec le logiciel SIMCA-P. Les fonctions Splus suivantes sont disponibles auprès des auteurs : $PLS1(X, y, H)$ pour le modèle à H composantes et $PLS1cv(X, y)$ pour le modèle avec choix du nombre de composantes par validation croisée.

# 2 L'algorithme PLS1

La section 2.1 donne le principe de l'algorithme PLS1 sans données manquantes et permet d'introduire les écritures matricielles utilisées pour implémenter l'algorithme de la section 2.2.

## 2.1 Le principe de l'algorithme et passage à l'écriture matricielle

On considère que les vecteurs $x_j$ des variables explicatives et que le vecteur $y$ de la variable à expliquer sont centrés. Évidemment, les variables peuvent être centrées-réduites, mais seul le centrage des variables est utilisé lors du passage à l'écriture matricielle. On note $X$ la matrice individus $\times$ variables de dimensions $n \times p$. On va donc chercher le vecteur $a_h = (a_{h1}, ..., a_{hp})$ des coefficients de régressions du modèle à h composantes.

**Étape 1 :** On construit la première composante $t_1$ comme une combinaison linéaire des $p$ variables explicatives $x_j$. Les coefficients $w_1^{'} = (w_{11}, ..., w_{1j}, ..., w_{1p})$ de cette combinaison linéaire cherchent à "résumer" au mieux les variables explicatives $x_j$ et à "expliquer" au mieux la variable $y$ :

$$\begin{array}{rcl} t_1 & = & w_{11}x_1 + .... + w_{1p}x_p \\ w_{1j} & = & \frac{cov(x_j,y)}{\sqrt{\sum_{j=1}^{p} cov^2(x_j,y)}} \end{array}$$

On effectue ensuite une régression simple de $y$ sur $t_1$ :

$$y = c_1 t_1 + y_1$$

où $y_1$ est le vecteur des résidus et $c_1$ est le coefficient de régression :

$$c_1 = \frac{cov(y, t_1)}{\sigma_{t_1}^2}$$

On en déduit une première équation de régression :

$$y = \overbrace{c_1 w_{11}}^{a_{11}} x_1 + ... + \overbrace{c_1 w_{1p}}^{a_{1p}} x_p + y_1$$

Passage à l'écriture matricielle sachant que les vecteurs $x_j$ et $y$ sont centrés :

$$\begin{array}{lcccl} cov(x_j, y) & = & \sum_{i=1}^{n} x_{ij} y_i & = & x_j^{'} y \\ cov(y, t_1) & = & \sum_{i=1}^{n} y_i t_{1i} & = & y^{'} t_1 \\ \sigma_{t_1}^2 & = & \sum_{i=1}^{n} t_{1i} t_{1i} & = & t_1^{'} t_1 \end{array}$$

d'où :

$$\begin{aligned} w_1 &= \frac{X'y}{\|X'y\|} \\ t_1 &= Xw_1 \\ c_1 &= \frac{y't_1}{t_1't_1} \\ a_1 &= c_1w_1 \end{aligned}$$

**Étape 2 :** On construit une deuxième composante $t_2$, non corrélée à $t_1$ et expliquant bien le résidu $y_1$. Cette composante $t_2$ est combinaison linéaire des résidus $x_{1j}$ des régressions simples des variables $x_j$ sur $t_1$ :

$$\begin{aligned} t_2 &= w_{21}x_{11} + ... + w_{2j}x_{1j} + w_{2p}x_{1p} \\ w_{2j} &= \frac{cov(x_{1j},y_1)}{\sqrt{\sum_{j=1}^{p} cov^2(x_{1j},y_1)}} \end{aligned}$$

Pour calculer les résidus $x_{1j}$, on réalise une régression linéaire de toutes les variables $x_j$ sur $t_1$ :

$$x_j = p_{1j}t_1 + x_{1j}$$

où $x_{1j}$ est le vecteur de résidus et $p_{1j}$ est le coefficient de régression :

$$p_{1j} = \frac{cov(x_j, t_1)}{\sigma^2_{t_1}}$$

d'où

$$x_{1j} = x_j - p_{1j}t_1$$

On effectue ensuite une régression de $y$ sur $t_1$ et $t_2$:

$$y = c_1t_1 + c_2t_2 + y_2$$

où $c_1$ est le coefficient de régression de la première étape, $c_2$ est le coefficient de la régression simple de $y_1$ sur $t_2$ et $y_2$ le vecteur des résidus de cette régression :

$$y_1 = c_2t_2 + y_2$$

d'où

$$c_2 = \frac{cov(y_1, t_2)}{\sigma^2_{t_2}}$$

Nous verrons section 2.3 comment calculer le vecteur $a_2$ des coefficients de l'équation de régression :

$$y = a_{21}x_1 + ... + a_{2p}x_p + y_2$$

Passage à l'écriture matricielle sachant que les $x_{1j}$ et $y_1$ sont centrés et en notant $X_1 = (x_{11}, ..., x_{1p})$ la matrice des résidus $x_{1j}$ :

$$\begin{aligned} p_1 &= \frac{X^{'} t_1}{t_1^{'} t_1} \\ X_1 &= X - t_1 p_1^{'} \\ y_1 &= y - c_1 t_1 \\ w_2 &= \frac{X_1^{'} y_1}{\|X_1^{'} y_1\|} \\ t_2 &= X_1 w_2 \\ c_2 &= \frac{y_1^{'} t_2}{t_2^{'} t_2} \end{aligned}$$

**Étapes suivantes :** Cette procédure itérative peut se poursuivre en utilisant de la même manière les résidus $y_2$ et $x_{21}$,..., $x_{2p}$. Le nombre de composantes $t_1$,...,$t_H$ à retenir est habituellement déterminé par validation croisée. Cette procédure sera présentée section 3.

## 2.2 L'algorithme

Cette version de l'algorithme PLS1 ne traite pas les données manquantes. Il s'agit donc d'une version simplifiée de celui p. 99 dans [Tenenhaus, 1998].

Étape 1 : $X_0 = X$ et $y_0 = y$

Étape 2 : Pour $h = 1, ..., H$ :

Étape 2.1 : Calcul du vecteur $w_h = (w_{h1}, ..., w_{hj}, ..., w_{hp})$

$$w_h = \frac{X_{h-1}^{'} y_{h-1}}{\|X_{h-1}^{'} y_{h-1}\|} \tag{1}$$

Étape 2.2 : Calcul de la composante $t_h$

$$t_h = X_{h-1} w_h \tag{2}$$

Étape 2.3 : Calcul du coefficient de régression $c_h$ de $y_{h-1}$ sur $t_h$

$$c_h = \frac{y_{h-1}^{'} t_h}{t_h^{'} t_h} \tag{3}$$

Étape 2.4 : Calcul du vecteur $y_h$ des résidus de la régression de $y_{h-1}$ sur $t_h$

$$y_h = y_{h-1} - c_h t_h$$

Étape 2.5 : Calcul du vecteurs $p_h$ des coefficients des régressions de $x_{hj}$ sur $t_h$

$$p_h = \frac{X_{h-1}^{'} t_h}{t_h^{'} t_h}$$

Étape 2.6 : Calcul de la matrice $X_h$ des vecteurs des résidus des régressions de $x_{hj}$ sur $t_h$

$$X_h = X_{h-1} - t_h p_h^{'}$$

## 2.3 Calcul des coefficients de régression

L'algorithme tel qu'il est présenté section 2.2 ne calcule pas de manière explicite les coefficients de régression $a_h$ du modèle à h composantes :

$$y = a_{h1}x_1 + ... + a_{hp}x_p + y_h = Xa_h + y_h$$

On montre que :

$$a_h = c_1 w_1^* + ... + c_h w_h^* \tag{4}$$

où $C_h^{'} = (c_1, ..., c_h)$ est le vecteur des coefficients des régressions linéaires sur les $h$ composantes, et $W_h^* = (w_1^*, ..., w_h^*)$ est la matrice des $h$ vecteurs $w_h^*$ vérifiant :

$$t_h = w_{h1}^* x_1 + ... + w_{hp}^* x_p = X w_h^* \tag{5}$$

On montre que le vecteur $w_h^*$ est défini par la formule de récurrence suivante :

$$w_h^* = w_h - \sum_{k=1}^{h-1} w_k^* (p_k^{'} w_h) \tag{6}$$

Il suffit alors de rajouter à l'étape 2 de l'algorithme donné section 2.2 les étapes 2.7 et 2.8 suivantes :

Étape 2.7 : $w_h^* = w_h$
Pour $k = 1, ..., h$
$w_h^* = w_h^* - w_k^* (p_k^{'} w_h)$

Étape 2.8 : $a_h = W_h^* C_h$

Pour retrouver la formule (4) à partir de (5) :
On sait que

$$\begin{aligned} y &= c_1 t_1 + ... + c_h t_h + y_h \\ t_h &= X w_h^* \end{aligned}$$

d'où

$$\begin{aligned} y &= c_1 X w_1^* + ... + c_h X w_h^* + y_h \\ &= X(\underbrace{c_1 w_1^* + ... + c_h w_h^*}_{a_h}) + y_h \end{aligned}$$

Pour retrouver la formule de récurrence (6) :

- Le calcul de $w_1^*$ est immédiat car $t_1 = X\overbrace{w_1}^{w_1^*}$
- Calcul de $w_2^*$ : On cherche $w_2^*$ tel que $t_2 = Xw_2^*$. On a :

$$\begin{aligned} t_2 &= X_1 w_2 \\ &= (X - t_1 p_1^{'}) w_2 \\ &= (X - X w_1 p_1^{'}) w_2 \\ &= \underbrace{X(w_2 - w_1 p_1^{'} w_2)}_{w_2^*} \end{aligned}$$

- Calcul de $w_3^*$ : de la même manière on a :

$$\begin{aligned} t_3 &= X_2 w_3 \\ &= (X_1 - t_2 p_2^{'}) w_3 \\ &= (X - t_1 p_1^{'} - t_2 p_2^{'}) w_3 \\ &= X \underbrace{(w_3 - w_1(p_1^{'} w_3) - (w_2^*)(p_2^{'} w_3))}_{w_3^*} \end{aligned}$$

- D'où en généralisant on retrouve (6) :

$$\begin{aligned} w_h^* &= w_h - \overbrace{w_1}^{w_1^*} p_1^{'} w_h - w_2^* p_2^{'} w_h - ... - w_{h-1}^* p_{h-1}^{'} w_h \\ &= w_h - \textstyle\sum_{k=1}^{h-1} w_k^* p_k^{'} w_h \end{aligned}$$

# 3 Choix du nombre de composantes par validation croisée

Dans cette section, nous allons présenter en détails comment calculer les critères $PRESS$, $RSS$ et $Q_h^2$ afin de retrouver les résultats numériques obtenus avec le logiciel SIMCA-P sur l'exemple des données de Cornell [Tenenhaus, 1998].

## 3.1 Le principe

La procédure de validation croisée pour le choix du nombre de composantes [pp.77 et 83, Tenenhaus, 1998] est la suivante. A chaque étape $h$ et donc pour chaque nouvelle composante $t_h$ on calcule le critère :

$$Q_h^2 = 1 - \frac{PRESS_h}{RSS_{h-1}}$$

Pour $h = 1$, on a $RSS_0 = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = n - 1$ lorsque la variable est centrée réduite en utilisant la division par $n-1$ pour le calcul de la variance.

Une nouvelle composante $t_h$ est significative et donc conservée si $Q_h^2 \geq 0.0975$.

Il faut donc définir les critères $RSS_h$ et $PRESS_h$. Classiquement et c'est la présentation adoptée dans [Tenenhaus, 1998], on utilise le modèle de régression de $y$ sur les $h$ composantes suivant

$$y = \overbrace{c_1 t_1 + ... + c_h t_h}^{\hat{y_h}} + y_h \tag{7}$$

pour calculer la prédiction $\hat{y_h} = y - y_h$. En fait, pour chaque observation $i$ :

- on calcule la prédictions $\hat{y}_{hi}$ de $y_i$ à l'aide du modèle (7) obtenu en utilisant toutes les observations.
- on calcule la prédiction $\hat{y}_{h(-i)}$ de $y_i$ à l'aide du modèle (7) obtenu sans utiliser l'observation $i$.

Les critères $RSS_h$ (Residual Sum of Square) et $PRESS_h$ (PRediction Error Sum Of Squares) sont alors définis par :

$$RSS_h = \sum_{i=1}^{n}(y_i - \hat{y}_{hi})^2 \tag{8}$$

et

$$PRESS_h = \sum_{i=1}^{n}(y_i - \hat{y}_{h(-i)})^2 \tag{9}$$

L'implémentation sous S-plus de ces calculs ne nous permettant pas de retrouver les valeurs $Q_h^2$ fournies par SIMCA-P pour les données de Cornell [pp.83, Tenenhaus, 1998]), nous avons essayé de nombreuses autres méthodes de calcul. Finalement, l'idée qui nous a permis de retouver ces résultats numériques est la suivante. A chaque étape $h$, c'est-à-dire pour chaque nouvelle composante $t_h$, les critères $RSS_h$ et $PRESS_h$ sont calculés à partir du vecteur des résidus $y_{(h-1)}$ de l'étape précédente et non plus à partir de $y$ :

- Etape 1 : calcul du $PRESS$ et du $RSS$ sur $y$.
- Etape 2 : calcul du $PRESS$ et du $RSS$ sur $y_1$ vecteur des résidus de la régression de $y$ sur $t_1$.
- ...

A partir de là, à chaque étape $h$, trois modèles de régression peuvent être utilisés pour calculer la prédiction $\tilde{y_h} = y_{(h-1)} - y_h$ selon que l'on considère la régression du vecteur des résidus $y_{h-1}$ sur $t_h$, sur la matrice des résidus $X_{(h-1)}$ ou sur la matrice initiale $X$. En effet, sachant que

$$t_h = X_{(h-1)} w_h$$

et que

$$t_h = X w_h^*$$

nous avons les trois modèles suivant :

$$y_{(h-1)} = \overbrace{c_h t_h}^{\tilde{y_h}} + y_h \tag{10}$$

$$y_{(h-1)} = \overbrace{c_h X_{(h-1)} w_h}^{\tilde{y_h}} + y_h \tag{11}$$

$$y_{(h-1)} = \overbrace{c_h X w_h^*}^{\tilde{y_h}} + y_h \tag{12}$$

Pour une observation $i$, la prédiction $\tilde{y}_{hi}$ de $y_{(h-1)i}$ est la même avec les trois modèles obtenus en utilisant toutes les observations. En revanche, la prédiction $\tilde{y}_{h(-i)}$ de $y_{(h-1)i}$ est différente avec les trois modèles obtenus en retirant l'observation $i$. En conséquence, le choix du modèle n'intervient pas dans le calcul du $RSS$ mais uniquement dans le calcul du $PRESS$. Ces deux critères sont maintenant définis par :

$$RSS_h = \sum_{i=1}^{n} (y_{(h-1)i} - \tilde{y}_{hi})^2 \tag{13}$$

et

$$PRESS_h = \sum_{i=1}^{n} (y_{(h-1)i} - \tilde{y}_{h(-i)})^2 \tag{14}$$

## 3.2 Calcul du RSS

Nous venons de voir qu'il est équivalent de calculer le $RSS_h$ à partir de la formule (13) et de l'un des trois modèles (10), (11) ou (12). De plus, il est équivalent de calculer $RSS_h$ :

- sur $y$ avec avec la formule (8) et le modèle (7) :

$$y = \overbrace{c_1 t_1 + ... + c_h t_h}^{\hat{y_h}} + y_h$$

d'où

$$RSS_h = \sum_{i=1}^{n} (y_i - \hat{y}_{hi})^2 = \| y - \hat{y}_h \|^2 = \| y_h \|^2$$

- sur le vecteur des résidus $y_{(h-1)}$ avec la formule (13) et le modèle (10) :

$$y_{h-1} = \underbrace{c_h t_h}_{\tilde{y}_h} + y_h$$

d'où :

$$RSS_h = \| y_{h-1} - \tilde{y}_h \|^2 = \| y_{h-1} - c_h t_h \|^2 = \| y_h \|^2$$

Finalement, on calcule :

$$\begin{cases} RSS_1 &= \| y_1 \|^2 \\ \quad \vdots & \\ RSS_h &= \| y_h \|^2 \end{cases}$$

## 3.3 Calcul du $PRESS$

Nous venons de voir qu'il y a trois manières différentes de calculer la prédiction $\tilde{y}_h$ selon que l'on considère la régression de $y_{h-1}$ sur $t_h$, sur la matrice des résidus $X_{(h-1)}$ ou sur la matrice initiale $X$ c'est à dire selon que l'on considère le modèle (10), (11) ou (12). On a donc à chaque étape $h$ et pour chaque observation $i$ trois manières de calculer la prédiction $\tilde{y}_{hi}$ :

$$\begin{aligned} \tilde{y}_{hi} &= c_h t_{hi} \\ &= c_h X_{(h-1)i} w_h \\ &= c_h X_i w_h^* \end{aligned} \qquad (15)$$

où

- $X_{(h-1)i}$ est la i-ème ligne de la matrice des résidus $X_{(h-1)}$ calculée à l'étape 2.6 de l'algorithme (à l'étape $h-1$), $X_i$ est la i-ème ligne de la matrice initiale $X$ et $t_{hi}$ est la i-ème valeur de la composante $t_h$ calculée à l'étape 2.2 de l'algorithme.

- le coefficient $c_h$, les vecteurs $w_h$ et $w_h^*$ sont calculés respectivement aux étapes 2.3, 2.1 et 2.7 de l'algorithme.

Il y a donc à chaque étape $h$ et pour chaque observation $i$, trois manières non équivalentes cette fois, de calculer la prédiction $\tilde{y}_{h(-i)}$. En fait, pour retrouver les résultats numériques de SIMCA-P sur les données de Cornell nous avons utilisé l'égalité (15) c'est à dire la régression du vecteur des résidus $y_{h-1}$ sur la matrice des résidus $X_h$. Cette égalité s'écrit alors

$$\tilde{y}_{h(-i)} = c_{h(-i)} X_{(h-1)i} w_{h(-i)} \tag{16}$$

où

- $X_{(h-1)i}$ est la i-ème ligne de la matrice des résidus $X_{(h-1)}$

- $c_{h(-i)}$ et $w_{h(-i)}$ sont calculés sur $n-1$ individus, l'individu $i$ étant retiré. Plus précisément, ils sont calculés à partir des formules (3) et (1) utilisées aux étapes 2.3 et 2.1 de l'algorithme, mais ces formules sont appliquées à la matrice des résidus $X_{(h-1)}$ privée de la ligne $i$ et au vecteur des résidus $y_{(h-1)}$ privé de sa i-ème valeur.

Finalement, on note $X_{(h-1)(-i)}$ la matrice des résidus $X_{(h-1)}$ privée de sa i-ème ligne et $y_{(h-1)(-i)}$ le vecteur $y_{h-1}$ privé de sa i-ème valeur. Ces notations sont un peu lourdes, mais le calcul du $PRESS$ est simple et peut s'inclure à l'algorithme à la suite de l'étape 2.3 :

Étape (2.3)' : Calcul de $PRESS_h$

- Pour $i = 1, ..., n$ en utilisant les formules (1), (2) et (3) on calcule :

$$\begin{aligned}
w_{h(-i)} &= \frac{X'_{(h-1)(-i)} y_{(h-1)(-i)}}{\| X'_{(h-1)(-i)} y_{(h-1)(-i)} \|} \\
t_{h(-i)} &= X_{(h-1)(-i)} w_{h(-i)} \\
c_{h(-i)} &= \frac{y'_{(h-1)(-i)} t_{h(-i)}}{t'_{h(-i)} t_{h(-i)}} \\
\tilde{y}_{h(-i)} &= c_{h(-i)} X_{(h-1)i} w_{h(-i)}
\end{aligned}$$

- Puis on calcule :

$$PRESS_h = \sum_{i=1}^{n} (y_{(h-1)i} - \tilde{y}_{h(-i)})^2$$

où $y_{(h-1)i}$ est la i-ème valeur du vecteur des résidus $y_{(h-1)}$

# 4 Référence bibliographique

TENENHAUS, M. (1998). *la régression PLS, Théorie et Pratique*. Editions Technip.