# Régression linéaire généralisée PLS

Philippe Bastien(1), Vincenzo Esposito Vinzi(2), Michel Tenenhaus(3)
(1) L'Oréal Recherche (Aulnay), (2) Université Federico II (Naples), (3) Groupe HEC (Jouy-en-Josas)

## Introduction

La régression PLS univariée est un modèle non linéaire reliant une variable dépendante y à un ensemble de variables indépendantes quantitatives ou qualitatives $x_1$, …, $x_p$. Elle peut être obtenue par une suite de régressions simples et multiples. En exploitant les tests statistiques associés à la régression linéaire, il est possible de sélectionner les variables explicatives significatives à conserver dans la régression PLS et de choisir le nombre de composantes PLS à retenir. Le principe de l'algorithme présenté dans cette note peut également être utilisé pour obtenir une extension de la régression PLS à la régression linéaire généralisée PLS. Nous étudierons en détail la modification obtenue de la régression PLS usuelle, le cas de la régression logistique PLS et une utilisation de la régression linéaire généralisée PLS en données de survie. Les méthodes présentées seront illustrées par des exemples.

La régression PLS usuelle est issue d'une utilisation itérative des moindres carrés ordinaires et PLS est l'acronyme de *Partial Least Squares*. La régression linéaire généralisée PLS est construite sur une utilisation itérative du maximum de vraisemblance et PLS devient alors l'acronyme de *Projection onto Latent Structure*.

Brian Marx (1996) a proposé d'estimer les paramètres d'une régression linéaire généralisée en utilisant la régression PLS dans l'algorithme des moindres carrés pondérés itéré utilisé pour maximiser la vraisemblance. L'approche développée dans cette note est beaucoup plus simple, nécessite très peu de programmation nouvelle et est facilement généralisable à tous les modèles linéaires au niveau des variables explicatives.

## I. Présentation de la régression PLS

On suppose que les variables y, $x_1$,…,$x_p$ sont toutes centrées.

Le modèle de la régression PLS à m composantes s'écrit :

(1) $$y = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_{hj}^{*} x_j\right) + \text{résidu}$$

avec la contrainte que les composantes PLS $t_h = \sum_{j=1}^{p} w_{hj}^{*} x_j$ soient orthogonales. On peut considérer que les paramètres $c_h$ et $w_{hj}^{*}$ du modèle (1) sont des paramètres à estimer. D'où le côté non linéaire du modèle.

La régression PLS (Wold, Martens & Wold, 1983, Tenenhaus, 1998) est un algorithme d'estimation des paramètres du modèle (1) que nous allons re-décrire en reliant chaque étape de calcul à une régression linéaire simple ou multiple.

***Calcul de la première composante PLS $t_1$***

La première composante $t_1 = Xw_1^*$ est définie par

(2) $$t_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}(y,x_j)^2}} \sum_{j=1}^{p} \mathrm{cov}(y,x_j) x_j$$

La quantité $\mathrm{cov}(y, x_j)$ est aussi le coefficient de régression $a_{1j}$ dans la régression simple reliant la variable y à la variable $x_j/\mathrm{var}(x_j)$ :

(3) $$y = a_{1j}(x_j/\mathrm{var}(x_j)) + \text{résidu}$$

En effet :

$$a_{1j} = \frac{\mathrm{cov}(y, \frac{1}{\mathrm{var}(x_j)} x_j)}{\mathrm{var}(\frac{1}{\mathrm{var}(x_j)} x_j)} = \mathrm{cov}(y, x_j)$$

Un test sur le coefficient de régression $a_{1j}$ permet donc d'évaluer l'importance de la variable $x_j$ dans la construction de $t_1$. Ceci dit on pourrait tout aussi bien étudier la régression simple de y sur $x_j$ :

(4) $$y = a'_{1j} x_j + \text{résidu}$$

Les tests portant sur la nullité des $a_{1j}$ et $a'_{1j}$ sont naturellement équivalents.
Dans la formule (2) on pourrait remplacer par 0 les covariances non significatives.

***Calcul de la deuxième composante PLS $t_2$***

On construit les régressions simples de y et des $x_j$ sur $t_1$ :

(5) $$y = c_1 t_1 + y_1$$

et pour chaque j = 1 à p

(6) $$x_j = p_{1j} t_1 + x_{1j}$$

La deuxième composante $t_2$ est définie par

(7) $$t_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}(y_1, x_{1j})^2}} \sum_{j=1}^{p} \mathrm{cov}(y_1, x_{1j}) x_{1j}$$

La quantité $cov(y_1, x_{1j})$ est aussi le coefficient de régression $a_{2j}$ dans la régression multiple reliant la variable y aux variables $t_1$ et $x_{1j}/var(x_{1j})$ :

(8) $$y = c_{1j}t_1 + a_{2j}(x_{1j}/var(x_{1j})) + résidu$$

Ce résultat provient de l'orthogonalité entre le résidu $x_{1j}$ et la composante $t_1$.

La corrélation partielle entre y et $x_j$ conditionnellement à $t_1$ est définie comme la corrélation entre les résidus $y_1$ et $x_{1j}$. De même la covariance partielle entre y et $x_j$ conditionnellement à $t_1$ est définie comme la covariance entre les résidus $y_1$ et $x_{1j}$ :

(9) $$cov(y,x_j \mid t_1) = cov(y_1,x_{1j})$$

D'où une autre écriture de la deuxième composante PLS :

(10) $$t_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} cov(y,x_j|t_1)^2}} \sum_{j=1}^{p} cov(y,x_j|t_1)x_{1j}$$

Un test sur le coefficient de régression $a_{2j}$ permet d'évaluer l'importance de la variable $x_{1j}$ dans la construction de $t_2$. On pourrait aussi tester l'apport de la variable $x_j$ à la construction de la deuxième composante PLS en étudiant directement la régression de y sur $t_1$ et $x_j$ :

(11) $$y = c'_{1j}t_1 + a'_{2j}x_j + résidu$$

Les tests portant sur la nullité de $a_{2j}$ et $a'_{2j}$ sont équivalents car les vecteurs $(t_1, x_{1j})$ et $(t_1, x_j)$ engendrent le même espace.

Dans la formule (7) on pourrait remplacer par 0 les covariances non significatives.

La composante $t_2$ peut aussi s'exprimer en fonction des variables d'origine $x_j$ puisque les résidus $x_{1j} = x_j - p_{1j}t_1$ sont fonctions des $x_j$. La composante $t_2$ exprimée en fonction des $x_j$ s'écrit $t_2 = X w_2^*$.

### *Calcul des autres composantes et règle d'arrêt*

On procède de la même manière pour le calcul des autres composantes $t_h = X w_h^*$. On arrête en utilisant une procédure de validation croisée (cf. SIMCA-P) ou bien lorsque toutes les covariances partielles sont non significatives.

### *Formule de régression PLS*

Dans la formule (1) on estime les coefficients $c_h$ par régression multiple de y sur les composantes PLS $t_h$. L'équation de régression estimée peut ensuite s'exprimer en fonction des variables d'origine $x_j$ :

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

## II. Application

Nous allons étudier l'exemple des données de Cornell présenté dans Kettaneh-Wold (1992) et repris dans Tenenhaus (1998). Il s'agit de douze échantillons d'essence dont on connaît la composition donnée en proportion dans le tableau 1. On veut déterminer l'influence du mélange sur l'indice d'octane y. Les corrélations entre les variables sont données dans le tableau 2.

Description des variables :

| | | |
|---|---|---|
| Y | = | Indice d'octane |
| $x_1$ | = | Distillation directe (compris entre 0 et .21) |
| $x_2$ | = | Réformat (compris entre 0 et .62) |
| $x_3$ | = | Naphta de craquage thermique (compris entre 0 et .12) |
| $x_4$ | = | Naphta de craquage catalytique (compris entre 0 et .62) |
| $x_5$ | = | Polymère (compris entre 0 et .12) |
| $x_6$ | = | Alkylat (compris entre 0 et .74) |
| $x_7$ | = | Essence naturelle (compris entre 0 et .08) |

La somme des $x_j$ est égale à 1 pour chaque échantillon d'essence.

**Tableau 1 :** Données de Cornell

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | Y |
|---|---|---|---|---|---|---|---|
| .00 | .23 | .00 | .00 | .00 | .74 | .03 | 98.70 |
| .00 | .10 | .00 | .00 | .12 | .74 | .04 | 97.80 |
| .00 | .00 | .00 | .10 | .12 | .74 | .04 | 96.60 |
| .00 | .49 | .00 | .00 | .12 | .37 | .02 | 92.00 |
| .00 | .00 | .00 | .62 | .12 | .18 | .08 | 86.60 |
| .00 | .62 | .00 | .00 | .00 | .37 | .01 | 91.20 |
| .17 | .27 | .10 | .38 | .00 | .00 | .08 | 81.90 |
| .17 | .19 | .10 | .38 | .02 | .06 | .08 | 83.10 |
| .17 | .21 | .10 | .38 | .00 | .06 | .08 | 82.40 |
| .17 | .15 | .10 | .38 | .02 | .10 | .08 | 83.20 |
| .21 | .36 | .12 | .25 | .00 | .00 | .06 | 81.40 |
| .00 | .00 | .00 | .55 | .00 | .37 | .08 | 88.10 |

**Tableau 2 :** Le tableau des corrélations

Pearson Correlation Coefficients, N = 12
Prob > |r| under H0: Rho=0

| | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|---|---|---|---|---|---|---|---|---|
| **Y** | **1.00000** | **-0.83730**<br>**0.0007** | **-0.07082**<br>**0.8269** | **-0.83796**<br>**0.0007** | **-0.70671**<br>**0.0102** | **0.49380**<br>**0.1028** | **0.98507**<br>**<.0001** | **-0.74112**<br>**0.0058** |
| **X1** | -0.83730<br>0.0007 | 1.00000 | 0.10420<br>0.7473 | **0.99986**<br>**<.0001** | 0.37071<br>0.2355 | -0.54799<br>0.0651 | -0.80458<br>0.0016 | 0.60261<br>0.0381 |
| **X2** | -0.07082<br>0.8269 | 0.10420<br>0.7473 | 1.00000 | 0.10078<br>0.7553 | -0.53686<br>0.0719 | -0.29257<br>0.3561 | -0.19125<br>0.5516 | -0.59003<br>0.0434 |
| **X3** | -0.83796<br>0.0007 | 0.99986<br><.0001 | 0.10078<br>0.7553 | 1.00000 | 0.37400<br>0.2311 | -0.54820<br>0.0650 | -0.80520<br>0.0016 | 0.60708<br>0.0363 |
| **X4** | -0.70671<br>0.0102 | 0.37071<br>0.2355 | -0.53686<br>0.0719 | 0.37400<br>0.2311 | 1.00000 | -0.21133<br>0.5097 | -0.64566<br>0.0233 | **0.91588**<br>**<.0001** |
| **X5** | 0.49380<br>0.1028 | -0.54799<br>0.0651 | -0.29257<br>0.3561 | -0.54820<br>0.0650 | -0.21133<br>0.5097 | 1.00000 | 0.46292<br>0.1296 | -0.27436<br>0.3882 |
| **X6** | 0.98507<br><.0001 | -0.80458<br>0.0016 | -0.19125<br>0.5516 | -0.80520<br>0.0016 | -0.64566<br>0.0233 | 0.46292<br>0.1296 | 1.00000 | -0.65636<br>0.0204 |
| **X7** | -0.74112<br>0.0058 | 0.60261<br>0.0381 | -0.59003<br>0.0434 | 0.60708<br>0.0363 | 0.91588<br><.0001 | -0.27436<br>0.3882 | -0.65636<br>0.0204 | 1.00000 |

### *Utilisation de la régression multiple*

Ces données posent problème en régression. En effet les sorties de la Proc REG de SAS sur un modèle sans constante apparaissent dans le tableau 3 et donnent des résultats non conformes aux corrélations marginales pour les variables $x_1$, $x_2$, $x_3$, $x_5$ et $x_7$. Pour les variables $x_1$, $x_3$ et $x_7$ il s'agit d'un problème de multicolinéarité. Pour les variables $x_2$ et $x_5$ on trouvera une explication en utilisant la régression PLS.

**Tableau 3 :** Résultats de la régression multiple

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 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| X1 | 1 | 34.32023 | 209.12698 | 0.16 | 0.8761 |
| X2 | 1 | 85.92283 | 1.24820 | 68.84 | <.0001 |
| X3 | 1 | 141.25193 | 375.32840 | 0.38 | 0.7221 |
| X4 | 1 | 77.18010 | 9.21351 | 8.38 | 0.0004 |
| X5 | 1 | 87.75022 | 5.81572 | 15.09 | <.0001 |
| X6 | 1 | 100.30083 | 3.47382 | 28.87 | <.0001 |
| X7 | 1 | 116.92128 | 81.09563 | 1.44 | 0.2089 |

### *Utilisation de la régression PLS pas à pas ascendante*

Les variables d'origine y, $x_1$,…, $x_7$ centrées-réduites sont notées $y^*$, $x_1^*$,…, $x_7^*$.

### Construction de $t_1$

On décide de construire la composante $t_1$ en n'utilisant que les variables $x_j^*$ significatives au risque $\alpha = 0.05$.

En utilisant la formule (2) avec les seules corrélations significatives on obtient :

$$t_1 = \frac{-.8373x_1^* - .83796x_3^* - .70671x_4^* + .98507x_6^* - .74112x_7^*}{\sqrt{.8373^2 + .83796^2 + .70671^2 + .98507^2 + .74112^2}}$$

$$= -.4526x_1^* - .4530x_3^* - .3820x_4^* + .5325x_6^* - .4006x_7^*$$

**Construction de $t_2$**

On cherche dans un premier temps les variables $x_j^*$ contribuant de manière significative à la construction de $t_2$ On effectue donc sept régressions multiples de $y^*$ sur $t_1$ et chacun des $x_j^*$. On trouve dans le tableau 4 les variables significatives.

**Tableau 4 :** Résultats des régressions multiples de y sur $t_1$ et chaque $x_j^*$

| Variable | Coefficient de $x_j^*$ | Niveau de signification de $x_j^*$ |
|---|---|---|
| $x_1^*$ | .09699 | .6225 |
| $x_2^*$ | -.19728 | **.0101** |
| $x_3^*$ | .10358 | .6016 |
| $x_4^*$ | .01589 | .9055 |
| $x_5^*$ | .03634 | .7221 |
| $x_6^*$ | .65721 | **<.0001** |
| $x_7^*$ | .30464 | .0532 |

Seules les variables $x_2^*$ et $x_6^*$ sont significatives au risque $\alpha = 0.05$. On calcule les résidus $x_{12}$, $x_{16}$ des régressions de $x_2^*$, $x_6^*$ sur $t_1$. Les résidus $x_{12}$ et $x_{16}$ sont orthogonaux à $t_1$ et combinaisons linéaires des $x_j^*$. Par conséquent toute combinaison linéaire des résidus $x_{12}$ et $x_{16}$ est aussi une combinaison linéaire des $x_j^*$ orthogonale à $t_1$. Puis on effectue les deux régressions multiples de $y^*$ sur $t_1$ et $x_{1jn} = x_{1j}/\text{var}(x_{1j})$ , j = 2 et 6, dont les résultats sont donnés dans le tableau 5.

**Tableau 5 :** Résultats des régressions multiples de $y^*$ sur $t_1$ et $x_{1jn}$

| Variable | Coefficients de $x_{1jn}$ | Niveau de signification de $x_{1jn}$ |
|---|---|---|
| $x_{12n}$ | -.19403 | .0101 |
| $x_{16n}$ | .10098 | <.0001 |

<u>Conclusion</u> : On construit $t_2$ avec les deux résidus (non normalisés) $x_{12}$ et $x_{16}$ :

$$t_2 = \frac{-.19403x_{12} + .10098x_{16}}{\sqrt{.19403^2 + .10098^2}}$$

$$= .0731x_1^* - .8871x_2^* + .0732x_3^* + .0617x_4^* + .3756x_6^* + .0647x_7^*$$

On peut remarquer que la variable $x_2^*$ n'était pas du tout significative en terme de corrélation avec y. En revanche sa corrélation partielle avec y conditionnellement à $t_1$ est tout à fait significative. Et c'est la variable qui contribue le plus à la construction de la deuxième composante $t_2$.

**Construction de $t_3$**

On cherche tout d'abord les variables $x_j^*$ contribuant de manière significative à la construction de $t_3$ On effectue donc sept régressions multiples de y sur $t_1$, $t_2$, et sur chacun des $x_j^*$. On trouve les résultats de ces régressions dans le tableau 6.

Conclusion : Les variables $x_1^*$, $x_2^*$, $x_3^*$, $x_4^*$ et $x_6^*$ sont significatives au risque $\alpha = 0.05$; il faut donc rechercher une troisième composante PLS $t_3$.

**Tableau 6** : Résultats des régressions multiples de y sur $t_1$, $t_2$ et $x_j^*$.

| Variable | Niveau de signification des $x_j$ |
|---|---|
| $x_1^*$ | **.0034** |
| $x_2^*$ | **.0017** |
| $x_3^*$ | **.0028** |
| $x_4^*$ | **.0025** |
| $x_5^*$ | .4255 |
| $x_6^*$ | **.0017** |
| $x_7^*$ | .1085 |

On calcule ensuite les résidus $x_{2j}$ des régressions multiples des variables $x_j^*$ significatives sur les composantes $t_1$ et $t_2$. Ces résidus $x_{2j}$ sont tous orthogonaux à $t_1$ et $t_2$ et combinaisons linéaires des $x_j^*$. Par conséquent, toute combinaison linéaire des résidus $x_{2j}$ est aussi une combinaison linéaire des $x_j^*$ orthogonale à $t_1$ et $t_2$.

Puis, on effectue les régressions multiples de $y^*$ sur $t_1$, $t_2$ et les variables $x_{2jn} = x_{2j}/\mathrm{var}(x_{2j})$, significatives.

**Tableau 7** : Résultats des régressions multiples de $y^*$ sur $t_1$, $t_2$ et $x_{2jn}$.

| Variables | Coefficients de $x_{2jn}$ | Niveau de signification de $x_{2jn}$ |
|---|---|---|
| $x_{21n}$ | .05832 | .0034 |
| $x_{22n}$ | .01548 | .0017 |
| $x_{23n}$ | .05867 | .0028 |
| $x_{24n}$ | -.07732 | .0025 |
| $x_{26n}$ | .02974 | .0017 |

On construit $t_3$ avec les cinq résidus $x_{21}$, $x_{22}$, $x_{23}$, $x_{24}$, $x_{26}$ :

$$t_3 = \frac{.05832x_{21} + .01548x_{22} + .05867x_{23} - .07732x_{24} + .02974x_{26}}{\sqrt{.05832^2 + .01548^2 + .05867^2 + .07732^2 + .02974^2}}$$

$$= .493x_1^* - .286x_2^* + .496x_3^* - .655x_4^* + .469x_6^* + .001x_7^*$$

**Construction de $t_4$**

On recherche les variables $x_j^*$ contribuant de manière significative à la construction de $t_4$ en étudiant les régressions de y sur $t_1$, $t_2$, $t_3$ et $x_j^*$. On trouve les résultats de ces régressions dans le tableau 8.

**Tableau 8** : Résultats des régressions multiples de y sur $t_1$, $t_2$, $t_3$ et $x_j^*$

| Variable | Niveau de signification de $x_{3j}$ |
|---|---|
| $x_1^*$ | .2627 |
| $x_2^*$ | .3193 |
| $x_3^*$ | .4169 |
| $x_4^*$ | .9845 |
| $x_5^*$ | .4868 |
| $x_6^*$ | .3193 |
| $x_7^*$ | .1192 |

Conclusion : Aucune des variables $x_1^*$, …, $x_7^*$ n'est significative au risque $\alpha = 0.05$; il ne faut donc retenir que les trois premières composantes PLS.

***Construction de l'équation de régression PLS à trois composantes***

La régression de la variable d'origine y sur les trois composantes $t_1$, $t_2$, $t_3$ conduit à l'équation

$$\hat{y} = 88.58 + 3.25t_1 + 1.35t_2 + 1.15t_3$$

qui peut aussi s'écrire en fonction des variables $x_j$ d'origine :

$$\hat{y} = 93.317 - 8.755x_1 - 7.782x_2 - 14.969x_3 - 8.434x_4 + 9.488x_6 - 44.978x_7$$

En exploitant le fait que la somme des $x_j$ vaut 1, cette équation devient :

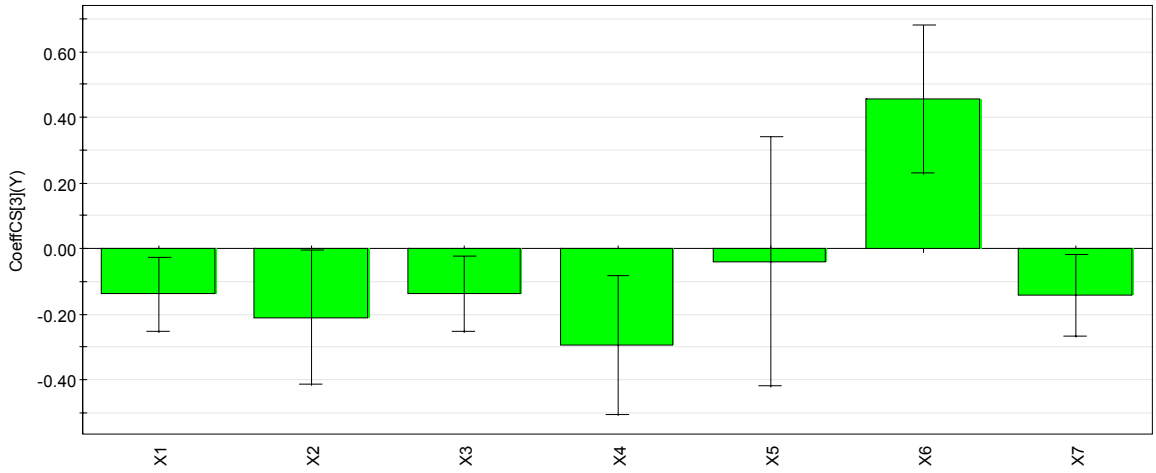
$$\begin{aligned}\hat{y} &= 93.317\left(\sum_{j=1}^{7} x_j\right) - 8.755x_1 - 7.782x_2 - 14.969x_3 - 8.434x_4 + 9.488x_6 - 44.978x_7 \\ &= 84.562x_1 + 85.535x_2 + 78.348x_3 + 84.883x_4 + 93.317x_5 + 102.805x_6 + 48.339x_7\end{aligned}$$

Si l'on veut fabriquer un mélange conduisant à un indice d'octane maximum il faut maximiser les composants par ordre décroissant des coefficients de régression. Ici on met un maximum de $x_6$ (74%); on complète par un maximum de $x_5$, soit 12%; on termine le mélange avec 14% de $x_2$. D'où un indice d'octane prédit en moyenne pour ce mélange égal à $\hat{y} = 85.535 \times 0.14 + 93.317 \times 0.12 + 102.805 \times 0.74 = 99.25$.

***Utilisation de SIMCA-P 9***

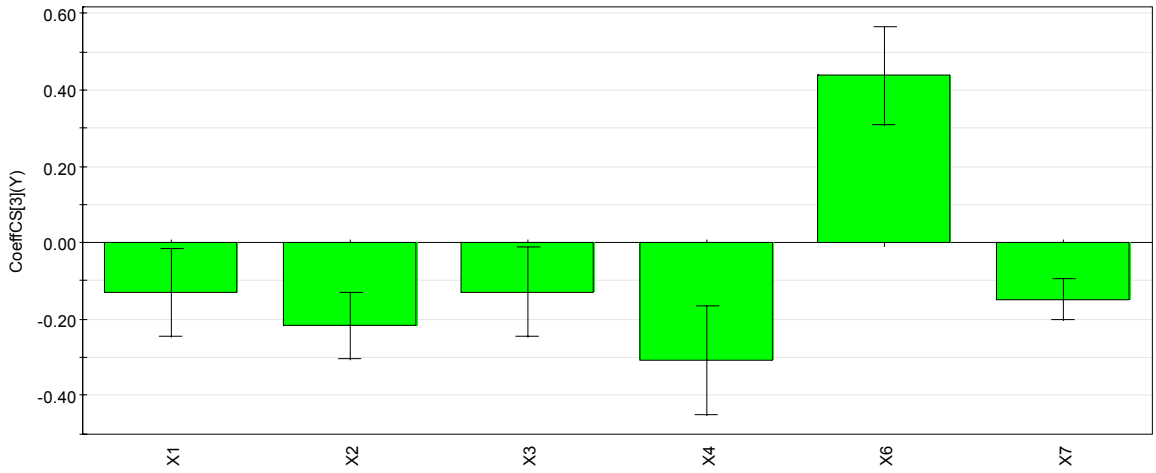
Nous avons utilisé sur ces données le logiciel SIMCA-P 10 (UMETRI, 2002). Cette version de SIMCA fournit les intervalles de confiance des coefficients de régression PLS à 95% calculés par validation croisée (Jack-knife). En réalisant la régression PLS de y sur les sept variables $x_j$ avec ce logiciel nous avons obtenu trois composantes PLS. On donne dans la figure 1 les intervalles de confiance des coefficients de régression PLS pour trois composantes.

**Figure 1 :** Régression PLS de y sur $x_1,\ldots, x_7$ :
Coefficients de régression et intervalles de confiance Jack-knife




Les résultats de la figure 1 confirment tout à fait la première analyse : la variable $x_5$ n'a pas d'influence directe sur l'indice d'octane y. La figure 2 montre les résultats obtenus en supprimant la variable $x_5$.

**Figure 2 :** Régression PLS de y sur $x_1,\ldots, x_4, x_6, x_7$ :
Coefficients de régression et intervalles de confiance Jack-knife



## III. Régression linéaire généralisée PLS

Nous avons vu que la régression PLS pouvait être obtenue en utilisant des régressions simples et multiples. En remplaçant ces régressions par des régressions linéaires généralisées on a accès a une nouvelle gamme de modèle : la régression linéaire généralisée PLS.

La régression linéaire généralisée PLS de y sur $x_1,\ldots, x_p$ à m composantes s'écrit :

$$(13) \qquad g(\theta) = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w^*_{hj} x_j\right)$$

où le paramètre $\theta$ peut représenter la moyenne $\mu$ d'une variable continue y, ou le vecteur des probabilités des valeurs prises par une variable y discrète, ou encore le rapport des risques $h(t)/h_0(t)$ en données de survie pour les modèles à risques proportionnels. La fonction de lien g est choisie par l'utilisateur en tenant compte de la loi de probabilité de y et de la qualité de l'ajustement du modèle aux données. On impose aux composantes PLS $T_h = \sum_{j=1}^{p} w^*_{hj} x_j$ d'être orthogonales.

Nous allons décrire cette nouvelle approche puis l'illustrer à l'aide de différents exemples.

### III.1 Algorithme de régression linéaire généralisée PLS

L'algorithme comprend quatre étapes :

1) La recherche des composantes PLS $T_h$
2) La régression linéaire généralisée de y sur les composantes PLS $T_h$
3) L'expression de la régression linéaire généralisée en fonction des variables d'origine
4) La validation Bootstrap des coefficients de l'équation de régression linéaire généralisée finale

Les trois dernières étapes étant évidentes, nous allons maintenant décrire en détail la première étape.

***Recherche des composantes PLS $T_h$***

On note X la matrice dont les colonnes sont formées des valeurs des p variables explicatives $x_j$. On recherche successivement m composantes PLS orthogonales $T_h$ combinaisons linéaires des $x_j$.

*Recherche de la première composante PLS $T_1$*

Etape 1 : Calculer le coefficient de régression $a_{1j}$ de $x_j$ dans la régression linéaire généralisée de y sur $x_j$ pour chaque variable $x_j$, j = 1 à p.
Etape 2 : Normer le vecteur colonne $a_1$ formé des $a_{1j}$ : $w_1 = a_1/\|a_1\|$
Etape 3 : Calculer la composante $T_1 = Xw_1/w_1'w_1$

*Recherche de la deuxième composante PLS $T_2$*

Etape 1 : Calculer le coefficient de régression $a_{2j}$ de $x_j$ dans la régression linéaire généralisée de y sur $T_1$ et $x_j$.
Etape 2 : Normer le vecteur colonne $a_2$ formé des $a_{2j}$ : $w_2 = a_2/||a_2||$
Etape 3 : Calculer le résidu $X_1$ de la régression linéaire de X sur $T_1$.
Etape 4 : Calculer la composante $T_2 = X_1w_2/w_2'w_2$.
Etape 5 : Exprimer la composante $T_2$ en fonction de X : $T_2 = Xw_2^*$

*Recherche de la h-ième composante PLS $T_h$*

On a obtenu aux étapes précédentes les composantes PLS $T_1,\ldots, T_{h-1}$. On obtient la composante $T_h$ en itérant la recherche de la deuxième composante.

Etape 1 : Calculer le coefficient de régression $a_{hj}$ de $x_j$ dans la régression linéaire généralisée de y sur $T_1, T_2, \ldots, T_{h-1}$ et $x_j$.
Etape 2 : Normer le vecteur colonne $a_h$ formé des $a_{hj}$ : $w_h = a_h/||a_h||$
Etape 3 : Calculer le résidu $X_{h-1}$ de la régression linéaire de X sur $T_1,\ldots,T_{h-1}$
Etape 4 : Calculer la composante $T_h = X_{h-1}w_h/w_h'w_h$.
Etape 5 : Exprimer la composante $T_h$ en fonction de X : $T_h = Xw_h^*$

***Commentaires***

1) On peut simplifier le calcul des composantes PLS $T_h$ en mettant à 0 les coefficients de régression $a_{hj}$ non significatifs. Seules les variables significatives contribuent à la construction de la composante PLS.
2) Le nombre m de composantes PLS à retenir peut être fixé par validation croisée sur les qualités prédictives du modèle ou bien en constatant que la composante $T_{m+1}$ n'est pas significative dans le sens où aucun $a_{m+1,j}$ ne l'est.
3) L'algorithme présenté peut fonctionner en présence de données manquantes. Notons $x_{h-1,i}$ le vecteur colonne formé de la transposée de la i-ième ligne de la matrice $X_{h-1}$. La valeur $T_{hi} = x_{h-1,i}'w_h/w_h'w_h$.de la composante $T_h$ pour l'individu i représente la pente de la droite des moindres carrés sans constante du nuage de points $(w_h, x_{h-1,i})$. Cette pente est aussi calculable lorsqu'il y a des données manquantes. Ainsi dans les étapes de calcul des composantes PLS, le calcul du dénominateur n'est réalisé que sur les données disponibles au numérateur.

## III.2 Application à la régression multiple

L'application de l'algorithme présenté en III.1 à un problème de régression multiple permet d'obtenir une variante de la régression PLS tout à fait naturelle et proposée en 1991 par Shenk et Westerhaus. Pour illustrer l'algorithme nous allons reprendre les données de Cornell. Nous décidons de travailler avec la variable y centrée-réduite et les variables $x_j$ centrées mais non réduites. Nous décidons également de ne conserver pour la construction des composantes PLS que les variables significatives.

***Calcul de la première composante PLS $T_1$***

La première composante $T_1 = Xw_1$ est définie par

(14) $$T_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{1j}^2}} \sum_{j=1}^{p} a_{1j} x_j$$

où $a_{1j}$ est le coefficient de régression de la variable centrée $x_j$ dans la régression de y = "Indice d'octane" sur chaque $x_j$. En remplaçant $a_{1j}$ par sa valeur $Cov(x_j,y)/Var(x_j)$ l'expression (14) devient

(15) $$T_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \left[ \frac{Cov(x_j, y)}{Var(x_j)} \right]^2}} \sum_{j=1}^{p} Cor(x_j, y) x_j^*$$

où $x_j^*$ représente la variable $x_j$ centrée-réduite. Il est certainement préférable de modifier la normalisation pour aboutir à :

(16) $$T_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} Cor(x_j, y)^2}} \sum_{j=1}^{p} Cor(x_j, y) x_j^*$$

Seules les variables $x_j$ à coefficient de corrélation avec y significatif contribuent à la construction de la première composante $T_1$. Il n'y a ici aucune différence entre cette nouvelle composante $T_1$ et l'ancienne.

***Calcul de la deuxième composante PLS $T_2$***

On recherche tout d'abord les variables $x_j$ contribuant de manière significative à l'explication de y en plus de $T_1$. On construit donc les régressions multiples de y sur $T_1$ et chaque $x_j$ :

(17) $$y = c_1 T_1 + a_{2j} x_j + \text{résidu}$$

Les résultats de ces régressions apparaissent dans le tableau 4.
Pour obtenir une deuxième composante $T_2$ orthogonale à $T_1$ il faut la construire à partir des résidus $x_{1j}$ des régressions des $x_j$ sur $T_1$ :

(18) $$x_j = p_{1j} T_1 + x_{1j}$$

La deuxième composante $T_2$ est alors définie par

(19) $$T_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{2j}^2}} \sum_{j=1}^{p} a_{2j} x_{1j}$$

où $a_{2j}$ est aussi le coefficient de régression de $x_{1j}$ dans la régression de y sur $T_1$ et $x_{1j}$ :

(20) $$\begin{aligned} y &= c_1 T_1 + a_{2j}(p_{1j} T_1 + x_{1j}) + \text{résidu} \\ &= (c_1 + p_{1j}) T_1 + a_{2j} x_{1j} + \text{résidu} \end{aligned}$$

L'orthogonalité entre $x_{1j}$ et $T_1$ implique qu'on a aussi, en modifiant la normalisation,

(21) $$T_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \text{Cor}(x_{1j}, y)^2}} \sum_{j=1}^{p} \text{Cor}(x_{1j}, y) x_{1j}^*$$

où $x_{1j}^*$ représente la variable $x_{1j}$ centrée-réduite.

Sur l'exemple on a obtenu :

(22) $$\begin{aligned} T_2 &= \frac{1}{\sqrt{.19565^2 + .25762^2}} (-.19565 x_{12}^* + .25762 x_{16}^*) \\ &= ..4211 x_1^* - .6098 x_2^* + .4214 x_3^* + .3554 x_4^* + 1.5363 x_6^* + .3727 x_7^* \end{aligned}$$

***Construction de la composante PLS $T_3$***

On cherche tout d'abord les variables $x_j$ contribuant de manière significative à la construction de $T_3$ On effectue donc sept régressions multiples de y sur $T_1$, $T_2$ , et sur chacun des $x_j$. On trouve les résultats de ces régressions dans le tableau 9.

**Tableau 9** : Résultats des régressions multiples de y sur $T_1$, $T_2$ et chaque $x_j$.

| Variables | Coefficients de $x_j$ | Niveau de signification de $x_j$ |
|---|---|---|
| $x_1$ | .1928 | **.0289** |
| $x_2$ | .2380 | **.0294** |
| $x_3$ | .1969 | **.0258** |
| $x_4$ | -.1478 | **.0177** |
| $x_5$ | -.0248 | .6356 |
| $x_6$ | .7929 | **.0294** |
| $x_7$ | -.2324 | .0922 |

Conclusion : Les variables $x_1$, $x_2$, $x_3$, $x_4$ et $x_6$ sont significatives au risque $\alpha = 0.05$; il faut donc rechercher une troisième composante PLS $T_3$.

On calcule ensuite les résidus $x_{2j}$ des régressions multiples des variables $x_j$ significatives sur les composantes $T_1$ et $T_2$.

On construit $T_3$ avec les cinq résidus réduits $x_{21}^*$, $x_{22}^*$, $x_{23}^*$, $x_{24}^*$, $x_{26}^*$ :

$$T_3 = \frac{.0859x_{21}^* + .0857x_{22}^* + .0872x_{23}^* - .0912x_{24}^* + .0857x_{26}^*}{\sqrt{.0859^2 + .0857^2 + .0872^2 + .0912^2 + .0857^2}}$$

$$= 1.600x_1^* + 1.248x_2^* + 1.622x_3^* - .243x_4^* + 3.260x_6^* + .541x_7^*$$

***Construction de la composante PLS $T_4$***

On recherche les variables $x_j$ contribuant de manière significative à la construction de $T_4$ en étudiant les régressions de y sur $T_1$, $T_2$, $T_3$ et chaque $x_j$. On trouve les résultats de ces régressions dans le tableau 10.

**Tableau 10** : Résultats des régressions multiples de y sur $T_1$, $T_2$, $T_3$ et chaque $x_j$.

| Variable | Niveau de signification de $x_j$ |
|---|---|
| $x_1$ | .7096 |
| $x_2$ | .9378 |
| $x_3$ | .8517 |
| $x_4$ | .5711 |
| $x_5$ | .6867 |
| $x_6$ | .9378 |
| $x_7$ | .3351 |

Conclusion : Aucune des variables $x_1$, …, $x_7$ n'est significative au risque $\alpha = 0.05$; il ne faut donc retenir que les trois premières composantes PLS.

***Construction de l'équation de régression PLS à trois composantes***

La régression de la variable d'origine y sur les trois composantes $T_1$, $T_2$, $T_3$ conduit à l'équation

$$\hat{y} = 88.58 + 3.25T_1 + 2.53T_2 + 1.20T_3$$

qui peut aussi s'écrire en fonction des variables $x_j$ d'origine :

$$\hat{y} = 87.682 - 5.920x_1 - 2.034x_2 - 10.060x_3 - 3.892x_4 + 15.133x_6 - 26.429x_7$$
$$= 87.682(\sum_{j=1}^{7} x_j) - 5.920x_1 - 2.034x_2 - 10.060x_3 - 3.892x_4 + 15.133x_6 - 26.429x_7$$
$$= 81.762x_1 + 85.648x_2 + 77.622x_3 + 83.790x_4 + 87.682x_5 + 102.815x_6 + 61.253x_7$$

L'indice d'octane maximum reste obtenu pour $x_2 = 0.14$, $x_5 = 0.12$, $x_6 = 0.74$ et atteint la valeur $\hat{y} = 85.648 \times 0.14 + 87.682 \times 0.12 + 102.815 \times 0.74 = 98.59$.

## III.3 Application à la régression logistique

Nous allons étudier l'application de l'algorithme général de régression linéaire généralisée à la régression logistique sur l'exemple des vins de Bordeaux.

Les variables suivantes ont été mesurées sur 34 années (1924 – 1957) :

TEMPÉRATURE : Somme des températures moyennes journalières (°C)
SOLEIL : Durée d'insolation (heures)
CHALEUR : Nombre de jours de grande chaleur
PLUIE : Hauteur des pluies (mm)
QUALITÉ du VIN : 1 = bonne, 2 = moyenne, 3 = médiocre

Les données figurent dans le tableau 11.

***La régression logistique ordinale***

La régression logistique ordinale de la qualité sur les quatre prédicteurs *centrées-réduits* correspond au modèle suivant :

$$\text{Prob}(Y \le i) = \frac{e^{\alpha_i+\beta_1 \text{Température}+\beta_2 \text{Soleil}+\beta_3 \text{Chaleur}+\beta_4 \text{Pluie}}}{1+e^{\alpha_i+\beta_1 \text{Température}+\beta_2 \text{Soleil}+\beta_3 \text{Chaleur}+\beta_4 \text{Pluie}}} \quad (23)$$

C'est un modèle à rapport des chances proportionnelles, accepté ici à l'aide d'un test du Score donné dans le tableau 12. Les résultats issus de la Proc Logistic de SAS sur les variables centrées-réduites apparaissent dans le tableau 12. Les niveaux de signification issus du test de Wald sur les quatre coefficients des variables prédictives de la qualité sont respectivement 0.0573, 0.1046, 0.4568, 0.0361. Seules les variables Température et Pluie sont significatives au risque de 10%. En utilisant le modèle (23) estimé on peut calculer la probabilité qu'une année soit bonne, moyenne ou médiocre. En affectant une année à la qualité la plus probable on obtient le tableau 13 croisant qualités observée et prévue. Il y a 7 années mal classées.

**Tableau 11 :** Les données vins de Bordeaux

| OBS | ANNÉE | TEMPÉRATURE | SOLEIL | CHALEUR | PLUIE | QUALITÉ |
|---|---|---|---|---|---|---|
| 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 |

**Tableau 12 :** Régression logistique de la qualité sur les variables météo

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 |
| SOLEIL | 1 | 1.7462 | 1.0760 | 2.6335 | 0.1046 |
| CHALEUR | 1 | -0.8891 | 1.1949 | 0.5536 | 0.4568 |
| PLUIE | 1 | -2.3668 | 1.1292 | 4.3931 | 0.0361 |

**Tableau 13:** Qualité de la prévision du modèle (11) en utilisant la régression logistique usuelle

| QUALITE OBSERVEE | PREVISION | | | |
|---|---|---|---|---|
| Effectif | 1 | 2 | 3 | Total |
| 1 | 8 | 3 | 0 | 11 |
| 2 | 2 | 8 | 1 | 11 |
| 3 | 0 | 1 | 11 | 12 |
| Total | 10 | 12 | 12 | 34 |

### *La régression logistique ordinale PLS*

Dans l'exemple des vins de Bordeaux, la multicolinéarité des prédicteurs conduit à deux difficultés : d'une part des variables influentes comme Soleil et Chaleur sont déclarées non significatives dans le modèle (23) alors que prises isolément elles le sont, et d'autre part la variable Chaleur apparaît dans l'équation du modèle avec un coefficient négatif, alors qu'elle a une influence positive sur la qualité. La régression logistique PLS permet en général d'obtenir un modèle cohérent au niveau des coefficients tout en conservant tous les prédicteurs. Elle fonctionne également lorsqu'il y a des données manquantes parmi les prédicteurs. Les régressions logistiques de la qualité sur chaque prédicteur centré-réduit pris séparément ont conduit aux coefficients $a_{1j}$ de Température, Soleil, Chaleur et Pluie égaux respectivement à 3.0117 (.0002), 3.3401 (.0002), 2.1445 (.0004) et -1.7906 (.0016), les niveaux de signification étant donnés entre parenthèses. Ces coefficients sont tous significatifs et ont des signes cohérents. Après normalisation de ces coefficients, on obtient la composante

$$T_1 = \frac{3.0117\ \text{Température} + 3.3401\ \text{Soleil} + 2.1445\ \text{Chaleur} - 1.7906\ \text{Pluie}}{\sqrt{(3.0117)^2 + (3.3401)^2 + (2.1445)^2 + (-1.7906)^2}}$$

$$= 0.5688\ \text{Température} + 0.6309\ \text{Soleil} + 0.4050\ \text{Chaleur} - 0.3382\ \text{Pluie}$$

Les résultats de la régression logistique de la qualité sur la composante $T_1$ sont donnés dans le tableau 14. Il est satisfaisant de constater qu'il y a une année mal classée de moins par rapport à la prévision réalisée avec la régression logistique usuelle.

Pour rechercher les variables contribuant de manière significative à la deuxième composante $T_2$ on construit les régression logistiques de la qualité sur $T_1$ et chaque prédicteur centré-réduit $x_j^*$ :

$$(24) \qquad (\text{Prob}(Y \le i) = \frac{e^{\alpha_i + \beta_1 T_1 + \beta_{2j} x_j^*}}{1 + e^{\alpha_i + \beta_1 T_1 + \beta_{2j} x_j^*}}$$

On obtient des niveaux de signification des prédicteurs égaux à .6765, .6027, .0983, .2544. On peut donc conclure que la deuxième composante PLS n'est pas significative. Par conséquent nous conservons le modèle à une composante.

**Tableau 14 :** Résultats de la régression logistique de la qualité sur la composante $T_1$

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| 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 |

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

| Effectif , | 1, | 2, | 3, | Total |
|---|---|---|---|---|
| 1 , | 9 , | 2 , | 0 , | 11 |
| 2 , | 2 , | 8 , | 1 , | 11 |
| 3 , | 0 , | 1 , | 11 , | 12 |
| Total | 11 | 11 | 12 | 34 |

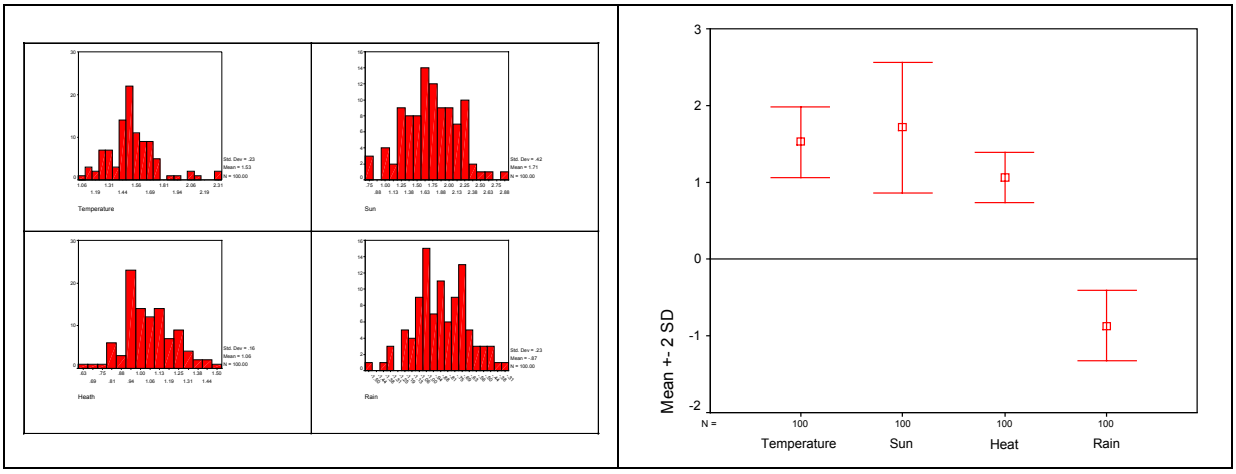
En exprimant la composante $T_1$ en fonction des variables Température, Soleil, Chaleur et Pluie centrées-réduites on obtient finalement des estimations plus cohérentes des paramètres du modèle (23) que celles obtenues en régression logistique usuelle :

$$\text{Prob}(Y = 1) = \frac{e^{-2.265 + 1.53 \times \text{Température} + 1.70 \times \text{Soleil} + 1.09 \times \text{Chaleur} - .91 \times \text{Pluie}}}{1 + e^{-2.5265 + 1.53 \times \text{Température} + 1.70 \times \text{Soleil} + 1.09 \times \text{Chaleur} - .91 \times \text{Pluie}}}$$

et

$$\text{Prob}(Y \le 2) = \frac{e^{2.2991 + 1.53 \times \text{Température} + 1.70 \times \text{Soleil} + 1.09 \times \text{Chaleur} - .91 \times \text{Pluie}}}{1 + e^{2.2991 + 1.53 \times \text{Température} + 1.70 \times \text{Soleil} + 1.09 \times \text{Chaleur} - .91 \times \text{Pluie}}}$$

Un bootstrap réalisé sur ce modèle de régression logistique PLS a donné les résultats donné dans le tableau suivant et valide donc les coefficients de régression :




## III.4 Exemple "Job Satisfaction"

Dans son livre *Models for Discrete Data* Daniel Zelterman propose un exemple de régression logistique sur variables qualitatives posant problème car l'auteur veut étudier les effets principaux et toutes les interactions d'ordre 2. Nous allons comparer sur cet exemple l'utilisation conjointe de la régression logistique et de la régression PLS sur variables qualitatives avec la régression logistique pas à pas de SAS version 8.2.

Voici la présentation de ces données par Zelterman :

*A large national corporation with more than 1 million employees sent out 100 000 surveys to determine the demographic factors influencing the satisfaction with their job. Approximately 75 000 surveys were returned , so there may have been a large bias due to those not responding. The data in* [Table 15] *tabulates the job satisfaction or dissatisfaction and demographic characteristics of 9949 employees in the 'craft' job classification within this company.*

**Table 15** : *Job satisfaction (Y/N) by sex (M/F), race, age, and region of residence for employees of a large U.S. corporation. source: Fowlkes et al. (J.A.S.A., 83, 611-622, 1988)*

| | White | | | | | | Nonwhite | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Under 35 | | 35-44 | | Over 44 | | Under 35 | | 35-44 | | Over 44 | |
| Region | M | F | M | F | M | F | M | F | M | F | M | F |
| Northeast | | | | | | | | | | | | |
| Y | 288 | 60 | 224 | 35 | 337 | 70 | 38 | 19 | 32 | 22 | 21 | 15 |
| N | 177 | 57 | 166 | 19 | 172 | 30 | 33 | 35 | 11 | 20 | 8 | 10 |
| Mid-Atlantic | | | | | | | | | | | | |
| Y | 90 | 19 | 96 | 12 | 124 | 17 | 18 | 13 | 7 | 0 | 9 | 1 |
| N | 45 | 12 | 42 | 5 | 39 | 2 | 6 | 7 | 2 | 3 | 2 | 1 |
| Southern | | | | | | | | | | | | |
| Y | 226 | 88 | 189 | 44 | 156 | 70 | 45 | 47 | 18 | 13 | 11 | 9 |
| N | 128 | 57 | 117 | 34 | 73 | 25 | 31 | 35 | 3 | 7 | 2 | 2 |
| Midwest | | | | | | | | | | | | |
| Y | 285 | 110 | 225 | 53 | 324 | 60 | 40 | 66 | 19 | 25 | 22 | 11 |
| N | 179 | 93 | 141 | 24 | 140 | 47 | 25 | 56 | 11 | 19 | 2 | 12 |
| Northwest | | | | | | | | | | | | |
| Y | 270 | 176 | 215 | 80 | 269 | 110 | 36 | 25 | 9 | 11 | 16 | 4 |
| N | 180 | 151 | 108 | 40 | 136 | 40 | 20 | 16 | 7 | 5 | 3 | 5 |
| Southwest | | | | | | | | | | | | |
| Y | 252 | 97 | 162 | 47 | 199 | 62 | 69 | 45 | 14 | 8 | 14 | 2 |
| N | 126 | 61 | 72 | 27 | 93 | 24 | 27 | 36 | 7 | 4 | 5 | 0 |
| Pacific | | | | | | | | | | | | |
| Y | 119 | 62 | 66 | 20 | 67 | 25 | 45 | 22 | 15 | 10 | 8 | 6 |
| N | 58 | 33 | 20 | 10 | 21 | 10 | 16 | 15 | 10 | 8 | 6 | 2 |

### III.4.1 Utilisation de la régression logistique PLS comme méthode de sélection de variables en régression logistique binaire sur données agrégées

La régression PLS peut être considérée comme une méthode de sélection de variables alternative à la régression pas à pas. La régression PLS est en quelque sorte une réponse au regret qu'éprouve l'utilisateur d'une régression pas à pas de devoir choisir une seule variable explicative par bloc de variables explicatives très corrélées entre elles. En régression PLS on remplace en quelque sorte ce bloc de variables par un résumé : la composante PLS. En régression pas à pas il y a exclusion de variables éventuellement importantes; en régression PLS toutes les variables importantes sont conservées et les variables sans importance sont soit exclues, soit participent au modèle, mais avec un poids faible.

Pour réaliser la régression logistique de *Job Satisfaction* sur les caractéristiques des employés Zelterman a construit toutes les variables indicatrices des modalités des facteurs et des interactions. Il s'est donc retrouvé avec une régression logistique binaire avec 44 variables explicatives indépendantes. Il a réalisé une régression logistique pas à pas ascendante et a abouti à des résultats peu satisfaisants car difficiles à interpréter. Il est vrai qu'il était limité par la version 6 de SAS. Dans la version 8 la proc Logistic contient une option "Class" et il est alors possible de réaliser une régression pas à pas ascendante au niveau des facteurs et des

interactions plutôt qu'au niveau des indicatrices. On obtient alors des résultats satisfaisants que nous présenterons plus loin.

Dans cette section nous souhaitons explorer l'utilisation de la régression logistique PLS comme outil de sélection de variables dans un problème de régression logistique binaire avec des prédicteurs qualitatifs.

***Construction de la première composante PLS $T_1$***

On sélectionne tout d'abord les facteurs et les interactions ayant une contribution significative dans la régression logistique de la variable *Job satisfaction* sur les variables décrivant la population étudiée. Pour les facteurs on les considère séparément. Pour chaque interaction on la considère en plus des deux effets principaux la formant. D'où le tableau 16. Pour construire la première composante PLS $T_1$ on va donc utiliser les variables Race, Age, Sexe, Région, Race*Sexe et Age*Sexe.

**Tableau 16** : Régression logistique de *Job Satisfaction* sur chaque facteur pris séparément et les interactions en plus des effets principaux les formant

| Variable | Wald | Niveau de signification |
|---|---|---|
| Race | 2.687 | .1012 |
| Age | 51.4856 | **<.0001** |
| Sexe | 20.8241 | **<.0001** |
| Région | 33.9109 | **<.0001** |
| Race*Age | 1.0578 | .5893 |
| Race*Sexe | 10.77 | **.001** |
| Race*Région | 3.4125 | .7556 |
| Age*Sexe | 7.9389 | **.0189** |
| Age*Région | 7.8771 | .7947 |
| Sexe*Région | 4.1857 | .6516 |

On souhaite donc construire une première composante PLS $T_1$ de la forme

$$T_1 = \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}\beta_1\\ -\beta_1\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}\beta_2\\ \beta_3\\ -\beta_2-\beta_3\end{bmatrix} + \begin{matrix}\text{Homme}\\ \text{Femme}\end{matrix}\begin{bmatrix}\beta_4\\ -\beta_4\end{bmatrix} + \begin{matrix}\text{Northeast}\\ \text{Mid}-\text{Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{matrix}\begin{bmatrix}\beta_5\\ \beta_6\\ \beta_7\\ \beta_8\\ \beta_9\\ \beta_{10}\\ -\beta_5-\ldots-\beta_{10}\end{bmatrix}$$

$$+ \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}\beta_{11} & -\beta_{11}\\ -\beta_{11} & \beta_{11}\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}\beta_{12} & -\beta_{12}\\ \beta_{13} & -\beta_{13}\\ -\beta_{12}-\beta_{13} & \beta_{12}+\beta_{13}\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

Les contraintes sur les coefficients sont plus naturelles que celles consistant à annuler les dernières modalités.

Pour obtenir la composante $T_1$ on réalise les régressions logistiques de la Satisfaction sur chacun des termes formant $T_1$ :

*Non Blanc - Blanc, $Age_{<35}$ - $Age_{>44}$, $Age_{35-44}$ - $Age_{>44}$, Homme - Femme,*
*Northeast - Pacific, ..., Southwest – Pacific, (Non Blanc- Blanc)* (Homme - Femme),*
*($Age_{<35}$ - $Age_{>44}$)*(Homme - Femme), ($Age_{35-44}$ - $Age_{>44}$)*(Homme – Femme).*

Il est préférable de ne pas réduire ces variables car on a alors accès aux odds-ratios en prenant l'exponentiel de deux fois les coefficients de régression. Les résultats de ces régressions logistiques sont donnés dans le tableau 17. La composante PLS $T_1$ est donnée dans le tableau 18. Ensuite on réalise la régression logistique de *Job Satisfaction* sur la composante PLS $T_1$ (Tableau 18), puis on l'exprime en fonction des variables d'origine. D'où le tableau 19.

**Tableau 17** : Les régressions logistiques de *Job Satisfaction* sur chaque variable

| Variable | Coefficient de Régression $a_{1j}$ | Coefficient normé $w_{1j}$ |
|---|---|---|
| Non blanc - Blanc | -.0486 | -.1424 |
| $Age_{<35}$ - $Age_{>44}$ | -.1775 | -.5203 |
| $Age_{35-44}$ - $Age_{>44}$ | -.1185 | -.3474 |
| Homme - Femme | .1050 | .3078 |
| Northeast - Pacific | -.1605 | -.4705 |
| Mid-Atlantic - Pacific | .0224 | .0656 |
| Southern - Pacific | -.0718 | -.2105 |
| Midwest - Pacific | -.1191 | -.3491 |
| Northwest - Pacific | -.0900 | -.2638 |
| Southwest -Pacific | .0136 | .0398 |
| (Non blanc - Blanc)*(Homme - Femme) | -.0124 | -.0363 |
| ($Age_{<35}$ - $Age_{>44}$)*(Homme - Femme) | .0095 | .0278 |
| ($Age_{35-44}$ - $Age_{>44}$)*(Homme - Femme) | -.0572 | -.1677 |

**Tableau 18** : Calcul de la première composante PLS $T_1$

$$T_1 = \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}-.14\\ +.14\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}-.52\\ -.35\\ +.87\end{bmatrix} + \begin{matrix}\text{Homme}\\ \text{Femme}\end{matrix}\begin{bmatrix}+.31\\ -.31\end{bmatrix} + \begin{matrix}\text{Northeast}\\ \text{Mid}-\text{Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{matrix}\begin{bmatrix}-.47\\ +.07\\ -.21\\ -.35\\ -.27\\ +.04\\ +1.19\end{bmatrix}$$

$$+ \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}+.04 & -.04\\ -.04 & +.04\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}+.03 & -.03\\ -.17 & +.17\\ +.15 & -.15\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

**Tableau 19** : Régression logistique de *Job Satisfaction* sur $T_1$

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | 0.5770 | 0.0210 | 755.5361 | <.0001 |
| T1 | 1 | 0.2328 | 0.0254 | 83.7304 | <.0001 |

**Tableau 20** : Régression logistique de *Job Satisfaction* sur $T_1$ exprimée en fonction des variables d'origine

$$\text{Logit(Prob(Satisfait))} =$$

$$0.58 + \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}-.03\\ +.03\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}-.12\\ -.08\\ +.20\end{bmatrix} + \begin{matrix}\text{Homme}\\ \text{Femme}\end{matrix}\begin{bmatrix}+.07\\ -.07\end{bmatrix} + \begin{matrix}\text{Northeast}\\ \text{Mid}-\text{Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{matrix}\begin{bmatrix}-.11\\ +.02\\ -.05\\ -.08\\ -.06\\ +.01\\ +.18\end{bmatrix}$$

$$+ \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}-.01 & +.01\\ +.01 & -.01\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}+.01 & -.01\\ -.04 & +.04\\ +.03 & -.03\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

### *Construction de la deuxième composante PLS $T_2$*

On commence par étudier la contribution des facteurs Race, Sexe, Age, Région et des interactions à la construction de $T_2$ en réalisant des tests sur ces facteurs en utilisant un modèle de régression logistique incluant $T_1$. D'où le tableau 21.

**Tableau 21** : Régression logistique de *Job Satisfaction* sur $T_1$ et chaque facteur pris séparément et les interactions en plus des effets principaux les formant

| Variables | Wald | Niveau de signification |
|---|---|---|
| Race | .56 | .45 |
| Age | 4.25 | .12 |
| Sexe | .32 | .57 |
| Région | 13.14 | **.04** |
| Race*Age | 1.21 | .54 |
| Race*Sexe | 11.6 | **.00** |
| Race*Région | 3.32 | .77 |
| Age*Sexe | 8.68 | **.01** |
| Age*Région | 7.73 | .80 |
| Sexe*Région | 3.49 | .75 |

Trois variables vont contribuer à la construction de la deuxième composante PLS $T_2$ : *Région, Race*Sexe* et *Age*Sexe*.

Pour obtenir une composante $T_2$ orthogonale à $T_1$, on construit $T_2$ à partir des résidus des régressions des variables

> *Northeast - Pacific, ..., Southwest - Pacific, (Non blanc - blanc)*(Homme - Femme), (Age$_{<35}$ - Age$_{>44}$)*(Homme - Femme), (Age$_{35\text{-}44}$ - Age$_{>44}$)*(Homme – Femme)*

sur $T_1$.

La composante $T_2$ est donnée dans le tableau 22 et dépend donc en fait des mêmes variables que $T_1$.

**Tableau 22** : Calcul de la deuxième composante PLS $T_2$

$$T_2 = \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}-.033\\ +.033\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}-.12\\ +.08\\ +.04\end{bmatrix} + \begin{matrix}\text{Homme}\\ \text{Femme}\end{matrix}\begin{bmatrix}+.07\\ -.07\end{bmatrix} + \begin{matrix}\text{Northeast}\\ \text{Mid}-\text{Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{matrix}\begin{bmatrix}-.40\\ +.67\\ +.15\\ -.17\\ -.04\\ +.38\\ -.59\end{bmatrix}$$

$$+ \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}+.28 & -.28\\ -.28 & +.28\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}+.33 & -.33\\ +.29 & -.29\\ -.62 & +.62\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

Ensuite on réalise la régression logistique de *Job Satisfaction* sur les composantes PLS $T_1$ et $T_2$ (Tableau 23), puis on l'exprime en fonction des variables d'origine. D'où le tableau 24.

**Tableau 23** : Régression logistique de *Job Satisfaction* sur $T_1$, $T_2$

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | 0.6046 | 0.0218 | 771.5523 | <.0001 |
| T1 | 1 | 0.2562 | 0.0257 | 99.6029 | <.0001 |
| T2 | 1 | 0.2039 | 0.0392 | 27.1008 | <.0001 |

**Tableau 24** : Régression logistique de *Job Satisfaction* sur $T_1$ et $T_2$ exprimée en fonction des variables d'origine

$$\text{Logit(Prob(Satisfait))} =$$

$$0.60 + \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}-.04\\ +.04\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}-.16\\ -.11\\ +.27\end{bmatrix} + \begin{matrix}\text{Homme}\\ \text{Femme}\end{matrix}\begin{bmatrix}+.09\\ -.09\end{bmatrix} + \begin{matrix}\text{Northeast}\\ \text{Mid}-\text{Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{matrix}\begin{bmatrix}-.20\\ +.15\\ -.02\\ -.12\\ -.07\\ +.09\\ +.17\end{bmatrix}$$

$$+ \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}+.05 & -.05\\ -.05 & +.05\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}+.07 & -.07\\ +.02 & -.02\\ -.09 & +.09\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

***Construction de la troisième composante PLS $T_3$***

On commence par étudier la contribution des facteurs Race, Sexe, Age, Région et des interactions à la construction de $T_3$ en réalisant des tests sur ces facteurs à l'aide d'un modèle de régression logistique incluant $T_1$ et $T_2$. D'où le tableau 25.

**Tableau 25** : Régression logistique de *Job Satisfaction* sur $T_1$, $T_2$ et chaque facteur pris séparément et les interactions en plus des effets principaux les formant

| Variables | Wald | Niveau de signification |
|---|---|---|
| Race | 1.39 | .24 |
| Age | 3.21 | .20 |
| Sexe | .08 | .77 |
| Région | 1.14 | .98 |
| Race*Age | .51 | .78 |
| Race*Sexe | 1.34 | .25 |
| Race*Région | 3.94 | .68 |
| Age*Sexe | 2.58 | .27 |
| Age*Région | 8.23 | .77 |
| Sexe*Région | 3.05 | .80 |

Il n'y a plus aucun effet significatif. On en conclut que deux composantes PLS sont suffisantes pour décrire la liaison entre *Job Satisfaction* et les caractéristiques des employés.

***Utilisation de la régression logistique usuelle sur les variables sélectionnées par PLS***

Les variables ayant été sélectionnées – celles qui apparaissent dans le tableau 24 - on peut maintenant revenir à la régression logistique usuelle pour étudier le modèle du tableau 24. La Proc Logistic de SAS Version 8.2 permet de prendre en compte les variables qualitatives en choisissant les contraintes imposées aux modalités des variables qualitatives. Nous choisissons comme contrainte d'imposer la nullité à la somme des coefficients d'un facteur ou

des lignes et des colonnes du tableau des coefficients décrivant une interaction d'ordre 2. C'est l'option par défaut de la Proc Logistic (Option Param = Effect de l'instruction Class).
On donne dans le tableau 26 les résultats de cette analyse et dans le tableau 27 le modèle obtenu par maximum de vraisemblance.

**Tableau 26** : Régression logistique de Job Satisfaction sur les variables sélectionnées par PLS. Tests de Wald

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| race | 1 | 0.1007 | 0.7510 |
| age | 2 | 50.7100 | <.0001 |
| sex | 1 | 14.0597 | 0.0002 |
| region | 6 | 37.7010 | <.0001 |
| race*sex | 1 | 7.5641 | 0.0060 |
| age*sex | 2 | 5.9577 | 0.0509 |

**Tableau 27** : Régression logistique de Job Satisfaction sur les variables sélectionnées par PLS : Le modèle

$$\text{Logit(Prob(Satisfait))} =$$

$$0.65 + \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}-.01\\ +.01\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}-.20\\ -.02\\ +.22\end{bmatrix} + \begin{matrix}\text{Homme}\\ \text{Femme}\end{matrix}\begin{bmatrix}+.12\\ -.12\end{bmatrix} + \begin{matrix}\text{Northeast}\\ \text{Mid}-\text{Atlantic}\\ \text{Southern}\\ \text{Midwest}\\ \text{Northwest}\\ \text{Southwest}\\ \text{Pacific}\end{matrix}\begin{bmatrix}-.22\\ +.22\\ -.04\\ -.13\\ -.09\\ +.07\\ +.19\end{bmatrix}$$

$$+ \begin{matrix}\text{Non}-\text{Blanc}\\ \text{Blanc}\end{matrix}\begin{bmatrix}+.09 & -.09\\ -.09 & +.09\\ \text{Homme} & \text{Femme}\end{bmatrix} + \begin{matrix}<35\\ 35-44\\ >44\end{matrix}\begin{bmatrix}+.08 & -.08\\ -.03 & +.03\\ -.05 & +.05\\ \text{Homme} & \text{Femme}\end{bmatrix}$$

**Interprétation** : La satisfaction augmente avec l'age. Les hommes sont plus satisfaits que les femmes. Les habitants des régions Mid-Atlantic, Southwest et Pacific sont plus satisfaits que ceux des autres régions. La différence de satisfaction entre les hommes et les femmes dépend de la race : elle est augmentée chez les Non-Blancs et diminuée chez les Blancs. De même la différence de satisfaction entre les hommes et les femmes dépend de l'age : elle est augmentée chez les jeunes et diminuée chez les plus vieux.

### III.4.2 Utilisation de la méthode de régression logistique pas à pas ascendante de la Proc Logistic de SAS version 8.2

Nous avons réalisé également une régression logistique pas à pas ascendante en utilisant la proc logistic de la version 8.2 de SAS. L'option *Hierarchy = none* permet d'introduire des interactions significatives même si les effets principaux ne le sont pas. C'est important sur cet exemple. Les résultats obtenus sont identiques à ceux issus de l'utilisation de la régression PLS décrite dans la section III.4.1.

Sur cet exemple la régression logistique PLS a conduit à sélectionner les mêmes variables que la régression logistique pas à pas ascendante. Elle n'offre donc pas ici d'intérêt particulier. Mais elle nous a permis d'exposer une méthodologie qui devrait être efficace lorsque le nombre de prédicteurs est élevé et le nombre d'observations faible. De plus en utilisant les principes de l'algorithme NIPALS on pourrait construire un algorithme permettant d'intégrer les données manquantes.

## III.5 Application aux données censurées : détermination des facteurs de risques associés à l'âge d'apparition des premiers cheveux blancs chez l'homme

### III.5.1 Présentation générale des données

Initiée en 1994 par le Professeur Serge Hercberg, l'étude épidémiologique SU.VI.MAX *(«SUpplémentation en VItamines et Minéraux Antioxydants»)* s'est donnée pour mission d'évaluer l'état nutritionnel de la population française et d'apprécier l'incidence d'un apport oral de vitamines et minéraux anti-oxydants sur différents indicateurs de santé : infarctus, maladies cardiovasculaires, cancers, dans lesquels les radicaux libres sont fortement impliqués (Hercberg 1997). Cette étude qui est prévue pour s'achever après 8 ans en 2002 comprend plus de 12.000 volontaires âgés entre 25 et 65 ans représentatifs de la population française, la moitié recevant un traitement antioxydant et l'autre moitié un placebo.

L'Oréal a conduit à partir de cette cohorte une étude sur l'état de santé des cheveux et des ongles chez plus de 10323 sujets (4057 hommes et 6266 femmes). Sur la base des réponses à un questionnaire comportant plus de 150 items on a cherché à mettre en évidence les facteurs de risques associés à une apparition prématurée de cheveux blancs chez l'homme.

### III.5.2 Description des variables

La variable réponse est l'âge d'apparition des premiers cheveux blancs chez l'homme. C'est une variable discrète à 6 modalités : 30 ans et moins, 31 à 35 ans, 36 à 40 ans, 41 à 45 ans, 45 à 50 ans, et plus de 50 ans.

Les idées de base pour analyser ce type de données sont décrites dans Allison (1995). Chaque individu i est dupliqué $m_i$ fois où $m_i$ est le numéro de sa classe d'âge lors de l'apparition de ses premiers cheveux blancs ou le numéro de sa classe d'âge actuelle s'ils ne sont pas encore apparus. On définit ensuite une variable de réponse $y_{it}$, qui vaut 0 si l'individu i n'a pas de cheveux blancs à l'âge t et 1 sinon. Pour le dernier enregistrement ($t = m_i$) $y_{it}$ vaut 1 si l'événement est apparu, et 0 sinon.

Par exemple on donne dans le tableau 28 le cas 1 d'un homme de 45 ans dont les premiers cheveux blancs sont apparus à 38 ans (il a 3 enregistrements), puis le cas 2 d'un homme de 45 ans sans cheveux blancs (il a 4 enregistrements).

**Tableau 28** : Préparation des données et construction de la fonction de réponse

| i | t | $y_{it}$ |
|---|---|---|
| Cas 1 | ≤30 | 0 |
| Cas 1 | 31-35 | 0 |
| Cas 1 | 36-40 | 1 |
| Cas 2 | ≤30 | 0 |
| Cas 2 | 31-35 | 0 |
| Cas 2 | 36-40 | 0 |
| Cas 2 | 41-45 | 0 |

Les facteurs de risques sélectionnés à partir de la base de donnée SUVIMAX sont les suivants :

*Antécédents familiaux (réponse binaire 1 = oui, 0 = non) :*

- La mère a eu des cheveux blancs avant 30 ans ($X_1$)
- La sœur a eu des cheveux blancs avant 30 ans ($X_2$)
- La mère n'a pas eu de cheveux blancs après 60 ans ($X_3$)
- Le père a eu des cheveux blancs avant 30 ans ($X_4$)
- Le père n'a pas eu de cheveux blancs après 60 ans ($X_5$)

*Caractéristiques des cheveux*

- Epaisseur des cheveux (très fin, fin, moyen, épais)
- Couleur naturelle des cheveux (roux, blond, châtain clair, châtain foncé, brun, noir)

Seuls des facteurs de risques significatifs ont été retenus pour cette application. De plus, afin de permettre une comparaison entre le modèle linéaire généralisé usuel et le modèle linéaire généralisé PLS, seuls les individus ayant des valeurs renseignées sur les facteurs de risques sélectionnés ont été pris en compte.

### III.5.3 Le modèle

Soit $P_{it}$ la probabilité que les premiers cheveux blancs de l'individu i soient apparus à l'instant t sachant qu'ils ne sont pas apparus aux instants précédents 1, 2, …, t-1, :

$$P_{it} = \text{Prob}(y_{it} = 1 \,/\, y_{i1} = 0, \ldots, y_{i,t-1} = 0)$$

C'est la probabilité conditionnelle d'observer « $y_{it} = 1$ » sachant que « $y_{i1} = 0, \ldots, y_{i,t-1} = 0$ ».

Supposons que le modèle continu sous-jacent soit un modèle de Cox à risques proportionnels

(25) $$\begin{aligned}\text{Log}[h_i(t)/h_0(t)] &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \\ &\quad + \beta_6 \text{Roux} + \beta_7 \text{Blond} + \beta_8 \text{Châtain Clair} + \beta_9 \text{Châtain Foncé} \\ &\quad + \beta_{10} \text{Très Fins} + \beta_{11} \text{Fins} + \beta_{12} \text{Moyens} \\ &= \sum_j \beta_j V_j\end{aligned}$$

Pour modéliser la variable réponse en fonctions des prédicteurs on peut alors utiliser un modèle linéaire généralisé avec une fonction de lien de type « Complementary Log Log » (Prentice and Gloeckler, 1978).

(26) $$\text{Log}[-\text{Log}(1-P_{it})] = \alpha_t + \sum_j \beta_j V_j$$

L'interprétation des coefficients de régression du modèle (26) en terme de risques relatifs est identique à celle du modèle (25) sous-jacent. Par exemple l'apparition des premiers cheveux blancs avant 30 ans chez la mère augmente le risque, toutes chose égales par ailleurs, de $100(\exp(\beta_1) - 1)$ %.

La forme particulière du vecteur réponse associé à un même individu (toutes les coordonnées sont nulles si l'individu est censuré, toutes les coordonnées sont nulles à l'exception de la dernière qui vaut un si l'individu n'est pas censuré) permet de factoriser la vraisemblance des données à l'aide des probabilités conditionnelles $P_{it}$ :

$$\begin{aligned}L &= \prod_{i=1}^{N} \text{prob}(y_{i1} = 0, y_{i2} = 0, \ldots, y_{it_i - 1} = 0, y_{it_i} = 1) \\ &= \prod_{i=1}^{N} \text{prob}(y_{it_i} = 1 / y_{i1} = 0, \ldots, y_{it_i - 1} = 0) \times \text{prob}(y_{i1} = 0, \ldots, y_{it_i - 1} = 0) \\ &= \prod_{i=1}^{N} P_{it_i}(1 - P_{it_i - 1}) \times \ldots \times (1 - P_{i1}) \\ &= \prod_{i=1}^{N} \prod_{l \le t_i} P_{il}^{y_{il}} (1 - P_{il})^{1 - y_{il}}\end{aligned}$$

Tout ce passe comme si les réponses $y_{il}$ suivaient indépendamment une loi de Bernoulli de paramètre $P_{il}$.

Le modèle (26) est un modèle linéaire généralisé bien défini que nous allons étudier successivement à l'aide de la proc GENMOD et à l'aide du modèle linéaire généralisé PLS.

### III.5.4 Utilisation du modèle linéaire généralisé

Les résultats de la procédure GENMOD de SAS version 8.2 appliquée au modèle (26) sont donnés dans le tableau suivant :

**Tableau 29** : Résultats de la procédure GENMOD

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 > ChiSq |
|---|---|---|---|
| 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 |
| épaisseur | 3 | 16.75 | 0.0008 |
| couleur | 5 | 31.12 | <.0001 |

Analysis Of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| 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 |
| Très fins | 1 | -0.4598 | 0.1157 | -0.6865 | -0.2330 | 15.80 | <.0001 |
| Fins | 1 | -0.3518 | 0.0982 | -0.5442 | -0.1594 | 12.84 | 0.0003 |
| Moyens | 1 | -0.2791 | 0.0975 | -0.4701 | -0.0881 | 8.20 | 0.0042 |
| Epais | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |
| roux | 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 |
| châtain clair | 1 | -0.3959 | 0.1055 | -0.6027 | -0.1891 | 14.07 | 0.0002 |
| châtain foncé | 1 | -0.3141 | 0.1032 | -0.5165 | -0.1118 | 9.26 | 0.0023 |
| brun | 1 | -0.1310 | 0.1120 | -0.3506 | 0.0886 | 1.37 | 0.2422 |
| noir | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | . | . |

## *Commentaires*

1) Le rapport de la déviance sur le nombre de degrés de libertés est inférieur à un et tend à valider l'adéquation du modèle aux données.
2) La probabilité conditionnelle d'apparition des premiers cheveux blancs augmente avec l'âge. C'est la traduction statistique d'un phénomène quasiment inéluctable dans la population étudiée.
3) Les antécédents familiaux vont tous dans le sens attendu.
4) Le risque d'apparition des cheveux blancs augmente avec l'épaisseur des cheveux

5) Le risque d'apparition des cheveux blancs est d'autant plus important que la couleur naturelle des cheveux est foncée.
6) La couleur rousse se positionne de façon singulière entre le châtain foncé et le brun. Ce résultat confirme la spécificité de cette couleur d'un point de vue bio-chimique et montre la nécessité de traiter ce facteur qualitativement.

### III.5.5 Utilisation du modèle linéaire généralisé PLS

#### III.5.5.1 Construction de la première composante PLS

On construit un modèle linéaire généralisé contenant l'âge et chaque prédicteur pris séparément. Les variables qualitatives sont prises en compte dans le modèle dans leur globalité.

Toutes les variables sont significatives (Tableau 30). Ce n'est pas un résultat surprenant dans la mesure ou seuls les facteurs de risque significatifs ont été retenus pour cette application. Elles vont donc toutes contribuer à la construction de la première composante T1.

**Tableau 30** : Régressions linéaires généralisés de y sur l'âge et chacun des prédicteurs

| Variables | Wald | | |
|---|---|---|---|
| | ddl | Khi-deux | NS |
| X1 | 1 | 42.53 | < 0.0001 |
| X2 | 1 | 59.76 | < 0.0001 |
| X3 | 1 | 35.16 | < 0.0001 |
| X4 | 1 | 74.05 | < 0.0001 |
| X5 | 1 | 31.93 | < 0.0001 |
| Epaisseur | 3 | 27.87 | < 0.0001 |
| Couleur | 5 | 46.93 | < 0.0001 |

Le tableau 31 présente les coefficients de régression $a_{1j}$ de $V_j$ dans la régression linéaire généralisée de y sur l'âge et chaque $V_j$ pris séparément.

La première composante PLS s'écrit en fonction des variables et des coefficients du tableau 31 :

$$T_1 = \sum_j a_{1j}(V_j - \overline{V}_j) / \sqrt{\sum a_{1j}^2}$$

**Tableau 31** : Régressions linéaires généralisées de y sur l'âge et chaque prédicteur $V_j$

| Variables : $V_j$ | Coefficients de régression $a_{1j}$ |
|---|---|
| X1 | 0.73 |
| X2 | 0.92 |
| X3 | -0.50 |
| X4 | 0.93 |
| X5 | -0.51 |
| Très fins | -0.56 |
| Fins | -0.38 |
| Moyens | -0.28 |
| Epais | 0.00 |
| Roux | -0.41 |
| Blond | -0.80 |
| Châtain clair | -0.49 |
| Châtain foncé | -0.40 |
| Brun | -0.19 |
| Noir | 0.00 |

### III.5.5.2 Construction de la seconde composante PLS

On construit un modèle linéaire généralisé contenant l'âge, T1, et chaque prédicteur pris séparément. Les résultats sont donnés dans le tableau 32. Aucune des variables n'étant significative, seule la composante PLS T1 est retenue.

**Tableau 32** : Régressions linéaires généralisées de y sur l'âge, T1 et chacun des prédicteurs

| Variables | Wald | | |
|---|---|---|---|
| | ddl | Khi-deux | NS |
| X1 | 1 | 1.23 | 0.27 |
| X2 | 1 | 1.63 | 0.20 |
| X3 | 1 | 0.41 | 0.52 |
| X4 | 1 | 0.48 | 0.49 |
| X5 | 1 | 0.26 | 0.61 |
| Epaisseur | 3 | 0.41 | 0.94 |
| Couleur | 5 | 0.26 | 0.99 |

### III.5.5.3 Utilisation du modèle linéaire généralisé sur T1

Les résultats de la régression linéaire généralisée de y sur l'âge et $T_1$ sont donnés dans le tableau 33.

***Remarque :***

On peut constater, comme le montre le tableau 34 ci-dessous, que les critères de qualité d'ajustement des modèles cloglog et cloglog PLS sont très proches et conduisent à accepter ces modèles avec des valeurs de la déviance et du Khi-deux de Pearson proches de leurs degrés de liberté.

**Tableau 33** : Régression linéaire généralisée de y sur l'âge et T1

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 > ChiSq |
|---|---|---|---|---|---|---|---|
| <=30 | 1 | -2.9125 | 0.1041 | -3.1165 | -2.7085 | 783.28 | <.0001 |
| 31-35 | 1 | -2.4855 | 0.0897 | -2.6614 | -2.3096 | 767.06 | <.0001 |
| 36-40 | 1 | -1.4883 | 0.0610 | -1.6078 | -1.3688 | 595.66 | <.0001 |
| 41-45 | 1 | -0.6544 | 0.0498 | -0.7521 | -0.5567 | 172.38 | <.0001 |
| 46-50 | 1 | -0.2500 | 0.0580 | -0.3637 | -0.1363 | 18.56 | <.0001 |
| >50 | 1 | 0.7060 | 0.0829 | 0.5436 | 0.8684 | 72.59 | <.0001 |
| T1 | 1 | 1.6864 | 0.1036 | 1.4833 | 1.8894 | 264.98 | <.0001 |

**Tableau 34** : Adéquation des modèles cloglog et cloglog PLS aux données

| | Modèle cloglog | | | Modèle cloglog PLS | | |
|---|---|---|---|---|---|---|
| | ddl | valeur | Valeur/ddl | ddl | valeur | Valeur/ddl |
| Déviance | 6147 | 5162 | 0.84 | 6159 | 5165 | 0.84 |
| Khi-deux de Pearson | 6147 | 5761 | 0.94 | 9159 | 5780 | 0.94 |

### III.5.5.4 Expression du modèle PLS en fonction des variables d'origine

En utilisant l'expression de T1 en fonction des variables d'origine on peut exprimer le modèle linéaire généralisé PLS en fonction des variables d'origine. Les estimations des coefficients des modèles cloglog et cloglog PLS ainsi que leurs intervalles de confiance à 95% (Wald pour le modèle classique et bootstrap pour le modèle PLS) sont très proches. Les résultats des deux modèles sont présentés dans le tableau 35.

***Commentaires :***

1) La méthode de ré-échantillonnage de type "Balanced bootstrap" introduite par Davison en 1986 a été utilisée afin d'obtenir des intervalles de confiances des paramètres. Cette méthode offre de meilleures performances qu'un simple « Uniform bootstrap ».
2) L'estimation des intervalles de confiance bootstrap a été faite en sélectionnant les percentiles; l'utilisation de l'estimation bootstrap de la variance des coefficients donne des résultats très comparables.
3) Avec un modèle PLS à deux composantes on retrouve à la deuxième décimale près les estimations du modèle classique.

**Tableau 35** : Coefficients des modèles cloglog et cloglog PLS à 1 composante en fonction des variables d'origine.

| Variables | cloglog | | | | cloglog PLS | | | |
|---|---|---|---|---|---|---|---|---|
| | Parameter estimates | Standard error | Hazard ratio | Wald 95% CI | Parameter estimates | Standard error | Hazard ratio | Bootstrap 95% CI |
| ≤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 | . | . |
| X1 | 0.50 | 0.12 | 1.65 | 1.28-2.10 | 0.59 | 0.12 | 1.80 | 1.45-2.22 |
| X2 | 0.63 | 0.13 | 1.88 | 1.45-2.44 | 0.74 | 0.12 | 2.09 | 1.72-2.59 |
| X3 | -0.45 | 0.09 | 0.64 | 0.54-0.76 | -0.39 | 0.06 | 0.67 | 0.58-0.75 |
| X4 | 0.82 | 0.11 | 2.27 | 1.82-2.80 | 0.74 | 0.12 | 2.11 | 1.79-2.97 |
| X5 | -0.43 | 0.09 | 0.65 | 0.54-0.78 | -0.40 | 0.07 | 0.67 | 0.58-0.77 |
| Très fins | -0.46 | 0.12 | 0.63 | 0.50-0.79 | -0.44 | 0.09 | 0.64 | 0.54-0.75 |
| Fins | -0.35 | 0.10 | 0.70 | 0.58-0.85 | -0.30 | 0.08 | 0.74 | 0.63-0.87 |
| Moyens | -0.28 | 0.10 | 0.76 | 0.63-0.91 | -0.22 | 0.08 | 0.80 | 0.70-0.93 |
| Epais | 0.00 | . | . | . | 0.00 | 0.00 | . | . |
| Roux | -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 |
| Châtain clair | -0.40 | 0.11 | 0.67 | 0.55-0.83 | -0.39 | 0.09 | 0.68 | 0.56-0.79 |
| Châtain foncé | -0.31 | 0.10 | 0.73 | 0.59-0.90 | -0.32 | 0.10 | 0.73 | 0.62-0.84 |
| Brun | -0.13 | 0.11 | 0.88 | 0.70-1.09 | -0.15 | 0.11 | 0.86 | 0.72-1.03 |
| Noir | 0.00 | . | . | . | 0.00 | . | . | . |

### III.5.6 Modèle de Cox PLS

Le modèle de Cox (25) peut aussi être utilisé avec des données discrètes (ici la variable durée de vie est l'âge d'apparition des premiers cheveux blancs et prend les valeurs t = 1 à 6). Nous avons utilisé l'approximation de Efron pour la prise en compte des ex-æquo. Comme pour le modèle (26) sur le complementary log log une seule composante PLS a été retenue. Les résultats des modèles de Cox classique et Cox PLS sont présentés dans le tableau 36.

**Tableau 36** : Coefficients des modèles de Cox classique et Cox PLS à une composante en fonction des variables d'origine

| | Cox | | | | Cox PLS | | | |
|---|---|---|---|---|---|---|---|---|
| Variables | Parameter estimates | Standard error | Hazard ratio | Wald 95% CI | Parameter estimates | Standard error | Hazard ratio | Bootstrap 95% CI |
| X1 | 0.47 | 0.12 | 1.60 | 1.27-2.02 | 0.55 | 0.11 | 1.72 | 1.44-2.16 |
| X2 | 0.58 | 0.13 | 1.78 | 1.39-2.29 | 0.69 | 0.11 | 1.99 | 1.62-2.36 |
| X3 | -0.43 | 0.08 | 0.65 | 0.55-0.77 | -0.37 | 0.06 | 0.69 | 0.61-0.77 |
| X4 | 0.76 | 0.11 | 2.15 | 1.75-2.64 | 0.69 | 0.10 | 2.01 | 1.64-2.46 |
| X5 | -0.41 | 0.09 | 0.67 | 0.56-0.79 | -0.39 | 0.07 | 0.68 | 0.61-0.77 |
| Très fins | -0.43 | 0.11 | 0.65 | 0.52-0.81 | -0.42 | 0.08 | 0.66 | 0.57-0.79 |
| Fins | -0.33 | 0.10 | 0.72 | 0.60-0.87 | -0.29 | 0.07 | 0.75 | 0.66-0.85 |
| Moyens | -0.26 | 0.09 | 0.77 | 0.64-0.93 | -0.21 | 0.07 | 0.81 | 0.70-0.95 |
| Epais | 0.00 | . | . | . | 0.00 | . | . | . |
| Roux | -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 |
| Châtain clair | -0.38 | 0.10 | 0.69 | 0.56-0.84 | -0.37 | 0.09 | 0.69 | 0.56-0.81 |
| Châtain foncé | -0.30 | 0.10 | 0.74 | 0.61-0.90 | -0.30 | 0.09 | 0.74 | 0.62-0.89 |
| Brun | -0.13 | 0.11 | 0.88 | 0.71-1.09 | -0.15 | 0.10 | 0.86 | 0.71-1.06 |
| Noir | 0.00 | . | . | . | 0.00 | . | . | . |

***Commentaires :***

1) Comme on pouvait s'y attendre les résultats du modèle de Cox et du cloglog sont très proches.
2) Les remarques faites précédemment dans la comparaison des résultats des modèles cloglog et cloglog PLS s'appliquent ici.
3) Les coefficients sont estimés par maximisation de la vraisemblance partielle (Cox, 1972).

## IV Conclusion

La re-formulation de l'algorithme de régression PLS que nous avons présentée dans cette note présente plusieurs avantages.

1) Elle permet de relier la régression PLS aux procédures habituelles de régressions simple et multiple et d'utiliser en régression PLS les tests habituels à ces méthodes. Ces tests statistiques permettent d'identifier les variables explicatives ne contribuant pas à la construction des composantes PLS et par conséquent ayant peu d'influence sur la variable à expliquer. Une composante PLS est considérée comme non significative lorsque aucune variable explicative n'a un poids significatif dans sa construction. Sur l'exemple présenté dans la section II notre approche a conduit à la même sélection de variables que l'utilisation de la régression PLS usuelle pas à pas descendante. Il faudrait bien sûr valider l'approche présentée sur une plus grande variété d'exemples. Il faudrait aussi comparer notre méthode de sélection des variables explicatives à des approches présentées dans Forina, Casolino & Pizarro Millan (1999), Gauchi et Chagnon (2001), Höskuldsson (2001), Lingren, Geladi, Rannar & Wold (1994), Sarabia, Ortiz, Sánchez & Herrero (2001).

2) Dans la pratique, dans les cas de forte multi-colinéarité, il est habituel d'utiliser la régression multiple pas à pas. Cette méthode a l'inconvénient de conduire à l'abandon de variables explicatives très corrélées à la variable à expliquer et donc importantes pour le chercheur. La régression PLS permet au contraire de conserver dans le modèle toutes les variables à fort pouvoir explicatif.

3) En cas de données manquantes les composantes PLS sont calculées en utilisant le principe de l'algorithme NIPALS. S'il y a des données manquantes les composantes PLS sont cependant corrélées. L'algorithme de régression PLS d'origine ne prend pas en compte cette situation. Notre formulation au contraire tient compte de la corrélation entre les composantes PLS en utilisant la régression multiple.

4) Notre présentation de la régression PLS se généralise de manière immédiate à la régression linéaire généralisée. Des premiers résultats ont déjà été obtenus en régression logistique PLS (Esposito Vinzi & Tenenhaus, 2001) et en données de survie avec le modèle de Cox PLS (Bastien & Tenenhaus, 2001).

## Références


1. Allison, Paul D. (1995) : *Survival Analysis Using the SAS System : A practical guide*, SAS Inc, Cary, NC.
2. Bastien P., Tenenhaus M. (2001) : PLS generalized linear regression. Application to the analysis of life time data. In *PLS and Related Methods, Proceedings of the PLS'01 International Symposium*, Esposito Vinzi V., Lauro C., Morineau A. & Tenenhaus M. (Eds). CISIA-CERESTA Editeur, Paris, p. 131-140.
3. Cox,D.R. (1972). *Regression models and life-tables (with discussion)*. Journal of the Royal Statistical Society, Series B, 34,187-220.
4. Davison, A. C., Hinkley, D. V. and Schechtman, E. (1986). *Efficient bootstrap simulations*, Biometrika, 73, 555-566.
5. Efron B., Tibshirani R.J. (1993) - *An introduction to the Bootstrap*. Chapman and Hall, New York.
6. Esposito Vinzi V., Tenenhaus M. (2001) : PLS Logistic Regression. In *PLS and Related Methods, Proceedings of the PLS'01 International Symposium*, Esposito Vinzi V., Lauro C., Morineau A. & Tenenhaus M. (Eds). CISIA-CERESTA Editeur, Paris, p. 117-130.
7. Forina M., Casolino C. Pizarro Millan C. (1999) : Iterative Predictor Weighting PLS (IPW): A technique for the elimation of useless predictors in regression problems. *Journal of Chemometrics,* **13**, 165-184.
8. Gauchi J.-P., Chagnon P. (2001) : Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data, in *Special issue of Chemometrics & Intelligent Laboratory Systems on PLS* (in press).
9. Hercberg, S. et al. (1997) : *A primary prevention trial of nutritional doses of antioxydant vitamins and minerals on cardiovascular diseases and cancers in general populations : The SUVIMAX Study. Design, methods and participants characteristics*. *Control Clin. Trials.*
10. Höskuldsson A. (2001) : Variable and subset selection in PLS regression. *Chemometrics & Intelligent Laboratory Systems, **55**, 23-38.*
11. Kettaneh-Wold N. (1992) : Analysis of mixture data with partial least squares. *Chemometrics & Intelligent Laboratory Systems, **14**, 57-69.*
12. Lingren F., Geladi P., Rannar S., Wold S. (1994) : Interactive Variable Selection (IVS) for PLS. Part 1:Theory and Algorithms. *Journal of Chemometrics,* **8**, 349-363.
13. Marx B.D. (1996) : Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression. *Technometrics,* vol. 38, n°4, pp. 374-381.
14. Prentice R.L. and Gloeckler L.A. (1978). *Regression Analysis of grouped survival data with application to breast cancer data, Biometrics 34,57-67.*
15. 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 *PLS and Related Methods, Proceedings of the PLS'01 International Symposium*, Esposito Vinzi V., Lauro C., Morineau A. & Tenenhaus M. (Eds). CISIA-CERESTA Editeur, Paris, p. 327-339.
16. Shenk J.S., and Westerhaus M. O (1991) : Population structuring of near infrared spectra and modified partial least squares regression. Crop Sci. 31:1548-1555.
17. Tenenhaus M. (1998) : *La régression PLS*. Technip, Paris
18. Umetri AB (2002) : *SIMCA-P 10, User Guide and Tutorial* Umetri AB, Box 7960, S-90719 Umeå, Sweden.
19. Wold S., Martens & Wold H. (1983) : The multivariate calibration problem in chemistry solved by the PLS method. In *Proc. Conf. Matrix Pencils*, Ruhe A. & Kåstrøm B. (Eds), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, p. 286-293.
20. Zelterman D. (1999) : *Models for discrete data*. Oxford Science Publications, New York.