La régression logistique PLS

Michel Tenenhaus
Groupe HEC, 78351 Jouy-en-Josas

1 Introduction

La régression PLS permet de relier une ou plusieurs variables de réponse $\mathbf{y}$ à un ensemble de variables prédictives $\mathbf{x}_1, ..., \mathbf{x}_k$ dans des conditions où la régression multiple fonctionne mal ou plus du tout (forte multicolinéarité, plus de variables que d'observations, données manquantes). On peut rencontrer des problèmes analogues en régression logistique et plus généralement dans l'utilisation du modèle linéaire généralisé. Il est tout à fait possible de transposer les principes de la régression PLS à la régression logistique et au modèle linéaire généralisé. Nous présentons dans ce chapitre la régression logistique PLS. On en déduit de manière immédiate le modèle linéaire généralisé PLS qui ne sera donc pas présenté. A côté de l'algorithme de régression logistique PLS de base, nous avons étudié la régression logistique sur composantes PLS, plus pragmatique, mais qui devrait être de qualité comparable. Nous présentons ces deux méthodes autour d'un exemple posant problème en régression logistique usuelle.

2 L'exemple des vins de Bordeaux.

Les données

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 1. Toutes les analyses de ce chapitre ont été réalisées sur les variables météo centrées-réduites.

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

$$\Pr\text{ob}(\mathbf{y} \leq i) = \frac{e^{\alpha_i + \beta_1 \text{Température} + \beta_2 \text{Soleil} + \beta_3 \text{Chaleur} + \beta_4 \text{Pluie}}}{1 + e^{\alpha_i + \beta_1 \text{Température} + \beta_2 \text{Soleil} + \beta_3 \text{Chaleur} + \beta_4 \text{Pluie}}} \qquad (1)$$

C'est un modèle à rapport des chances proportionnelles. Ce modèle est accepté ici à l'aide du test du Score donné dans le tableau 2 des résultats issus de la Proc Logistic de SAS. Les niveaux de signification issus du test de Wald des deux constantes et des quatre coefficients des variables prédictives de la qualité sont respectivement 0.0143, 0.0177, 0.0573, 0.1046, 0.4568, 0.0361. Seules les variables Température et Pluie sont significatives au risque de 10%. En utilisant le modèle (1) 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 3 croisant qualités observée et prévue. Il y a 7 années mal classées.

Observation	Température	Soleil	Chaleur	Pluie	Qualité
1	3064	1201	10	361	2
2	3000	1053	11	338	3
3	3155	1133	19	393	2
4	3085	970	4	467	3
5	3245	1258	36	294	1
6	3267	1386	35	225	1
7	3080	966	13	417	3
8	2974	1189	12	488	3
9	3038	1103	14	677	3
10	3318	1310	29	427	2
11	3317	1362	25	326	1
12	3182	1171	28	326	3
13	2998	1102	9	349	3
14	3221	1424	21	382	1
15	3019	1230	16	275	2
16	3022	1285	9	303	2
17	3094	1329	11	339	2
18	3009	1210	15	536	3
19	3227	1331	21	414	2
20	3308	1366	24	282	1
21	3212	1289	17	302	2
22	3361	1444	25	253	1
23	3061	1175	12	261	2
24	3478	1317	42	259	1
25	3126	1248	11	315	2
26	3458	1508	43	286	1
27	3252	1361	26	346	2
28	3052	1186	14	443	3
29	3270	1399	24	306	1
30	3198	1259	20	367	1
31	2904	1164	6	311	3
32	3247	1277	19	375	1
33	3083	1195	5	441	3
34	3043	1208	14	371	3

Tableau 1 : Les données des vins de Bordeaux

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 2 : Régression logistique de la qualité sur les variables météo

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

Tableau 3 : Qualité de la prévision du modèle (1)

3 La régression logistique 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 (1) 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. La régression logistique PLS consiste à adapter l'algorithme de régression PLS1 (Wold, Ruhe, Wold, Dunn, III (1984) ou Tenenhaus (1998)) au cas d'une variable de réponse binaire ou ordinale. On note $\mathbf{y}$ la variable de réponse et $\mathbf{X}$ la matrice dont les colonnes sont formées des valeurs des variables explicatives $\mathbf{x}_j$, $j = 1, ..., k$. Toutes les variables $\mathbf{x}_j$ sont centrées-réduites.

Nous allons maintenant décrire la construction des composantes PLS.

3.1 Recherche des composantes PLS

On recherche successivement des composantes PLS orthogonales $\mathbf{t}_h$ combinaisons linéaires des $\mathbf{X}$. Dans cet algorithme, si la variable de réponse est ordinale, la régression logistique est construite en supposant un modèle à rapport des chances proportionnel.

Recherche de la première composante PLS $\mathbf{t}_1$

Etape 1 : Calculer le coefficient de régression w_{1j} de $\mathbf{x}_j$ dans la régression logistique simple de $\mathbf{y}$ sur $\mathbf{x}_j$, pour chaque $j = 1, ..., k$.

Etape 2 : Normer le vecteur colonne $\mathbf{w}_1$ formé des w_{1j}.

Etape 3 : Calculer la composante $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1/\mathbf{w}_1^{'}\mathbf{w}_1$**.**

Recherche de la deuxième composante PLS $\mathbf{t}_2$

Etape 1 : Calculer le résidu $\mathbf{X}_1$ de la régression de $\mathbf{X}$ sur la composante $\mathbf{t}_1$. On note $\mathbf{x}_{1j}$ la j-ième colonne de la matrice $\mathbf{X}_1$.

Etape 2 : Calculer le coefficient de régression w_{2j} de $\mathbf{x}_{1j}$ dans la régression logistique multiple de $\mathbf{y}$ sur les variables $\mathbf{t}_1, \mathbf{x}_{1j}$, pour chaque $j = 1, ..., k$.

Etape 3 : Normer le vecteur colonne $\mathbf{w}_2$ formé des w_{2j}.

Etape 4 : Calculer la composante $\mathbf{t}_2 = \mathbf{X}_1\mathbf{w}_2/\mathbf{w}_2^{'}\mathbf{w}_2$.

Etape 5 : Exprimer la composante $\mathbf{t}_2$ en fonction de $\mathbf{X}$: $\mathbf{t}_2 = \mathbf{X}\mathbf{w}_2^*$.

Recherche de la h-ième composante PLS $\mathbf{t}_h$

On a obtenu aux étapes précédentes les composantes PLS $\mathbf{t}_1, ..., \mathbf{t}_{h-1}$. On obtient la composante $\mathbf{t}_h$ en prolongeant la recherche de la deuxième composante.

Etape 1 : Calculer le résidu $\mathbf{X}_{h-1}$ de la régression de $\mathbf{X}$ sur les composantes $\mathbf{t}_1, ..., \mathbf{t}_{h-1}$. On note $\mathbf{x}_{h-1,j}$ la j-ième colonne de la matrice $\mathbf{X}_{h-1}$.

Etape 2 : Calculer le coefficient de régression w_{hj} de $\mathbf{x}_{h-1,j}$ dans la régression logistique multiple de $\mathbf{y}$ sur les variables $\mathbf{t}_1, ..., \mathbf{t}_{h-1}$, $\mathbf{x}_{h-1,j}$, pour chaque $j = 1, ..., k$.

Etape 3 : Normer le vecteur colonne $\mathbf{w}_h$ formé des w_{hj}.

Etape 4 : Calculer la composante $\mathbf{t}_h = \mathbf{X}_{h-1}\mathbf{w}_h/\mathbf{w}_h^{'}\mathbf{w}_h$.

Etape 5 : Exprimer la composante $\mathbf{t}_h$ en fonction de $\mathbf{X}$: $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h^*$.

3.2 Commentaire

Notons $\mathbf{x}_{h-1(i)}$ le vecteur-colonne transposé de la i-ième ligne de la matrice $\mathbf{X}_{h-1}$. La valeur t_{hi} de la composante $\mathbf{t}_h$ pour l'individu i représente la coordonnée de la projection du vecteur $\mathbf{x}_{h-1(i)}$ sur l'axe engendré par le vecteur $\mathbf{w}_h$. C'est aussi la pente de la droite des moindres carrés sans constante du nuage de points $(\mathbf{w}_h, \mathbf{x}_{h-1(i)})$. 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.

3.3 Construction de l'équation de régression logistique PLS

A chaque étape h, on construit la régression logistique de $\mathbf{y}$ sur les composantes $\mathbf{t}_1, ..., \mathbf{t}_h$. L'équation de régression logistique PLS est obtenue en exprimant cette équation en fonction des variables d'origine. Ainsi pour une réponse $\mathbf{y}$ binaire (0/1), notant π la probabilité de l'évènement ($\mathbf{y} = 1$), on obtient :

$$\begin{aligned} \widehat{\log(\frac{\pi}{1-\pi})} &= c_1\mathbf{t}_1 + ... + c_h\mathbf{t}_h \\ &= c_1\mathbf{X}\mathbf{w}_1^* + ... + c_h\mathbf{X}\mathbf{w}_h^* \\ &= \mathbf{X}\mathbf{b} \end{aligned}$$

où

$$\mathbf{b} = c_1\mathbf{w}_1^* + ... + c_h\mathbf{w}_h^*$$

Comme en régression PLS, on peut encore construire une carte des variables en utilisant les coordonnées $(\mathbf{w}_1^*, c_1), (\mathbf{w}_2^*, c_2)$. Les produits scalaires entre les représentations (w_{1j}^*, w_{2j}^*) des variables $\mathbf{x}_j$ et (c_1, c_2) de $\mathbf{y}$ donnent alors des valeurs approchées des coefficients de régression logistique PLS.

3.4 Choix du nombre de composantes PLS

Le nombre de composantes PLS $\mathbf{t}_h$ est déterminé en régression PLS par validation croisée.Une composante $\mathbf{t}_h$ est ajoutée si le PRESS (PRedicted Error Sum of Squares) de l'étape h est nettement plus petit que le RESS (Residual Sum of Squares) de l'étape $h-1$. Wold propose dans le logiciel SIMCA (Umetri, 1999) d'introduire la h-ième composante si l'indice de Stone-Geisser

$$Q^2 = 1 - \frac{PRESS_h}{RESS_{h-1}}$$

est au moins égal à 0.0975. On peut utiliser le même type d'approche en régression logistique PLS.

Voici une manière de procéder lorsque la variable $\mathbf{y}$ est binaire (0/1). On utilise le χ^2 de Pearson défini par

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \pi_i)^2}{\pi_i(1-\pi_i)}$$

où y_i est la valeur de la variable $\mathbf{y}$ pour l'individu i et π_i la probabilité de l'événement ($\mathbf{y} = 1$) pour un individu ayant les caractéristiques de l'individu

i. Le χ^2 de l'étape h peut être calculé par substitution en remplaçant π_i par son estimation à l'aide de la régression logistique sur les composantes $\mathbf{t}_1, ..., \mathbf{t}_h$. Il peut aussi être calculé par validation croisée en estimant π_i sans utiliser l'observation i. On considère que la composante $\mathbf{t}_h$ est significative si le χ^2 calculé à l'étape h par validation croisée est nettement inférieur au χ^2 calculé à l'étape $h-1$ par substitution. En reprenant l'approche de Wold, on décide que la composante th est significative si l'indice

$$Q^2 = 1 - \frac{\chi^2_{\text{validation croisée, étape } h}}{\chi^2_{\text{substitution, étape } h-1}} \tag{2}$$

est au moins égal à 0.0975.

On peut généraliser cette approche sans difficulté pour une variable ordinale $\mathbf{y}$ à plus de deux modalités. Voici la manière de procéder décrite sur l'exemple des vins de Bordeaux. On note $\mathbf{y}_i = (\mathbf{y}_{1i}, \mathbf{y}_{2i}, \mathbf{y}_{3i})$ le vecteur indiquant la qualité de l'année i : la coordonnée $\mathbf{y}_{ji}$ vaut 1 si l'année i est de qualité j, 0 sinon. L'espérance de $\mathbf{y}_i$ est égal au vecteur $(\pi_{i1}, \pi_{i2}, \pi_{i3})$ formé des probabilités que l'année i soit bonne, moyenne ou médiocre. On obtient aussi la variance du vecteur $\mathbf{y}_i$:

$$Var(\mathbf{y}_i) = \begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} & -\pi_{i1}\pi_{i3} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) & -\pi_{i2}\pi_{i3} \\ -\pi_{i3}\pi_{i1} & -\pi_{i3}\pi_{i2} & \pi_{i3}(1-\pi_{i3}) \end{bmatrix}$$

En utilisant les deux premières composantes $(\mathbf{y}_{1i}, \mathbf{y}_{2i})$ de $\mathbf{y}_i$ on peut construire une mesure de l'écart entre les données et le modèle analogue au khi-deux de Pearson utilisé en régression logistique binaire :

$$\chi^2 = \sum_{i=1}^{n} (\mathbf{y}_{1i} - \pi_{i1}, \mathbf{y}_{2i} - \pi_{i2}) \begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_{1i} - \pi_{i1} \\ \mathbf{y}_{2i} - \pi_{i2} \end{bmatrix}$$

soit

$$\chi^2 = \sum_{i=1}^{n} (\mathbf{y}_{1i} - \pi_{i1}, \mathbf{y}_{2i} - \pi_{i2}) \left\{ \frac{1}{\pi_{i1}\pi_{i2}\pi_{i3}} \begin{bmatrix} \pi_{i2}(1-\pi_{i2}) & \pi_{i1}\pi_{i2} \\ \pi_{i2}\pi_{i1} & \pi_{i1}(1-\pi_{i1}) \end{bmatrix} \right\} \begin{bmatrix} \mathbf{y}_{1i} - \pi_{i1} \\ \mathbf{y}_{2i} - \pi_{i2} \end{bmatrix}$$

et en développant

$$\chi^2 = \sum_{i=1}^{n} \frac{1-\pi_{i2}}{\pi_{i1}\pi_{i3}} (\mathbf{y}_{1i} - \pi_{i1})^2 + \sum_{i=1}^{n} \frac{1-\pi_{i1}}{\pi_{i2}\pi_{i3}} (\mathbf{y}_{2i} - \pi_{i2})^2 + \sum_{i=1}^{n} \frac{2}{\pi_{i3}} ((\mathbf{y}_{1i} - \pi_{i1})\mathbf{y}_{12} - \pi_{i2})$$

On décide que la composante $\mathbf{t}_h$ est significative si l'indice Q^2 toujours défini comme en (2) est au moins égal à 0.0975.

On peut aussi chercher le nombre de composantes minimisant le pourcentage de mal classés.

3.5 Application aux vins de Bordeaux

Les régressions logistiques de la qualité sur chaque prédicteurs centrés-réduits (notés en italique) ont conduit aux coefficients w_{1j} de *Température, Soleil, Chaleur et Pluie* égaux respectivement à 3.0117, 3.3402, 2.1446 et -1.7906. Les signes des coefficients sont maintenant tous cohérents. Après normalisation de ces coefficients, on obtient la composante

$$\mathbf{t}_1 = 0.5688 \times Température + 0.6309 \times Soleil + 0.4051 \times Chaleur - 0.3382 \times Pluie$$

Les résultats de la régression logistique de la qualité sur la composante $\mathbf{t}_1$ sont donnés dans le tableau 4. 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.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept1	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	1	8	2	11
3	0	1	11	12
Total	10	11	13	34

Tableau 4 : Résultats de la régression logistique de la qualité sur la composante $\mathbf{t}_1$

Pour rechercher la deuxième composante $\mathbf{t}_2$ on calcule tout d'abord les résidus notés Température$_1$, Soleil$_1$, Chaleur$_1$, Pluie$_1$ de Température, Soleil, Chaleur, Pluie sur $\mathbf{t}_1$. Puis on réalise les quatre régressions logistiques de la qualité sur les variables $\mathbf{t}_1$ et chacune des variables Température$_1$, Soleil$_1$, Chaleur$_1$, Pluie$_1$. On obtient des coefficients w_{2j} égaux respectivement à – 0.6308, 0.6461, -1.9407, -0.9798. D'où après normalisation de ces coefficients la

composante

$$\mathbf{t}_2 = -0.2680 \times \text{Température}_1 + 0.2745 \times \text{Soleil}_1 - 0.8244 \times \text{Chaleur}_1 - 0.4162 \times \text{Pluie}_1$$

La composante $\mathbf{t}_2$ peut aussi s'écrire en fonction des variables d'origine centrées-réduites et on obtient :

$$\mathbf{t}_2 = -0.1081 \times Température + 0.4518 \times Soleil - 0.7105 \times Chaleur - 0.5113 \times Pluie$$

Les résultats de la régression logistique de la qualité sur les composantes $\mathbf{t}_1, \mathbf{t}_2$ sont donnés dans le tableau 5.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	-2.5362	0.8718	8.4638	0.0036
INTERCP2	1	2.1444	0.8950	5.7414	0.0166
T1	1	3.0254	0.8200	13.6141	0.0002
T2	1	1.4076	0.8802	2.5576	0.1098

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

Effectif	1	2	3	Total
1	9	2	0	11
2	1	8	2	11
3	0	1	11	12
Total	10	11	13	34

Tableau 5 : Résultats de la régression logistique de la qualité sur les composantes $\mathbf{t}_1, \mathbf{t}_2$

Le modèle à deux composantes classe toujours mal six années. Mais avec une différence au niveau de la qualité 2 : le modèle à une composante classait deux années moyennes en bonnes et une moyenne en médiocre, alors que le modèle à deux composantes classe une année moyenne en bonne et deux en médiocre. Cette dernière situation est évidemment préférable pour le consommateur.

Nous avons ensuite essayé le modèle à trois composantes. Il conduit à sept années mal classées et un coefficient négatif pour la variable Chaleur. Nous préférons donc conserver le modèle à deux composantes. En exprimant les composantes $\mathbf{t}_1$ et $\mathbf{t}_2$ en fonction des variables Température, Soleil, Chaleur et

Pluie on obtient finalement des estimations plus cohérentes des paramètres du modèle (1) que celles obtenues en régression logistique

$$\text{Prob}(\mathbf{y}=1) = \frac{e^{-2.54+1.57\times\text{Température}\ +2.54\times\text{Soleil}\ +0.23\times\text{Chaleur}\ -1.74\times\text{Pluie}}}{1+e^{-2.54+1.57\times\text{Température}\ +2.54\times\text{Soleil}\ +0.23\times\text{Chaleur}\ -1.74\times\text{Pluie}}}$$

et

$$\text{Prob}(\mathbf{y}\leq 2) = \frac{e^{2.14+1.57\times\text{Température}\ +2.54\times\text{Soleil}\ +0.23\times\text{Chaleur}\ -1.74\times\text{Pluie}}}{1+e^{2.14+1.57\times\text{Température}\ +2.54\times\text{Soleil}\ +0.23\times\text{Chaleur}\ -1.74\times\text{Pluie}}}$$

Nous allons maintenant illustrer la démarche de la régression logistique PLS en présence de données manquantes. Supposons que la valeur de la variable Température ne soit pas disponible pour l'année 1924. On trouve dans le tableau 6 les vecteurs $\mathbf{w}_1$ et $\mathbf{w}_2$ calculés sur les données complètes et incomplètes

	Données complètes		Données incomplètes	
	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_1$	$\mathbf{w}_2$
Température	0.5688	-0.2680	0.5743	-0.3168
Soleil	0.6309	0.2745	0.6280	0.2533
Chaleur	0.4050	-0.8244	0.4132	-0.8168
Pluie	-0.3382	-0.4162	-0.3366	-0.4102

Tableau 6 : $\mathbf{w}_1$ et $\mathbf{w}_2$ sur les données complètes et incomplètes

Détaillons le calcul de la composante $\mathbf{t}_1$ pour l'année 1924 :

$$\begin{aligned}
\mathbf{t}_{11} &= \frac{0.6280\times Soleil_1 + 0.4132\times Chaleur_1 - 0.3366\times Pluie_1}{0.6260^2+0.4132^2+0.3366^2} \\
&= \frac{0.6280\times(-0.36584)+0.4132\times(-0.88089)-0.3366\times(0.00611)}{0.6260^2+0.4132^2+0.3366^2} \\
&= -0.88146
\end{aligned}$$

On avait trouvé $\mathbf{t}_{11} = -0.968$ en utilisant les données complètes. On trouve de même

$$\begin{aligned}
\mathbf{t}_{21} &= \frac{0.2533\times \text{Soleil}_{11} - 0.8168\times \text{Chaleur}_{11} - 0.4102\times \text{Pluie}_{11}}{0.2533^2+0.8168^2+0.4102^2} \\
&= \frac{0.2533\times 0.10667 - 0.8168\times(-0.43483)-0.4102\times(-0.33856)}{0.2533^2+0.8168^2+0.4102^2} \\
&= 0.57923
\end{aligned}$$

On avait trouvé 0.529 en utilisant les données complètes. Les composantes PLS calculées sur les données complètes et incomplètes sont évidemment très proches. La corrélation entre $\mathbf{t}_1$ calculée sur les données complètes et $\mathbf{t}_1$ calculée

sur les données incomplètes est égale à 0.99994. Au niveau des secondes composantes PLS on trouve une corrélation de 0.99972. La corrélation entre les deux composantes PLS calculées sur les données incomplètes est égale à -0.01294 et elle n'est donc plus strictement nulle. On trouve dans le tableau 7 les résultats de la régression logistique de la qualité sur les composantes $\mathbf{t}_1, \mathbf{t}_2$ calculées sur les données incomplètes.

Les modèles calculés sur les données complètes et incomplètes sont évidemment très proches, mais curieusement le modèle calculé sur les données incomplètes conduit à une année mal classée de moins. Le modèle construit sur les données complètes classait comme médiocre l'année 1924 en réalité moyenne : les probabilités de bon, moyen, médiocre calculées étaient respectivement de 0.00884, 0.48147, 0.50969. Le modèle construit en supprimant la température de l'année 1924 conduit aux probabilités 0.01181, 0.55215, 0.43604 et donc à un bon classement. Ainsi la qualité de l'année 1924 est moyenne *malgré* une température un peu faible.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	-2.5490	0.8768	8.4507	0.0036
INTERCP2	1	2.1349	0.8955	5.6837	0.0171
T1	1	3.0797	0.8350	13.6032	0.0002
T2	1	1.4148	0.8849	2.5563	0.1099

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

Effectif	1	2	3	Total
1	9	2	0	11
2	1	9	1	11
3	0	1	11	12
Total	10	12	12	34

Tableau 7 : Résultats de la régression logistique de la qualité sur les composantes $\mathbf{t}_1; \mathbf{t}_2$

4 La régression logistique sur composantes PLS

On peut réaliser sur les données Vins de Bordeaux une régression PLS des variables Bon, Moyen, Médiocre, indicatrices des niveaux de qualité, sur les variables

de météo. Nous avons continué à utiliser le jeu de données avec la température de l'année 1924 manquante. Par validation croisée une seule composante PLS a été retenue. Cette composante PLS $\mathbf{t}_1$ est égale à :

$$\mathbf{t}_1 = 0.5531 \times Température + 0.54895 \times Soleil + 0.48061 \times Chaleur - 0.40218 \times Pluie$$

Pour l'année 1924, elle vaut

$$\begin{aligned} \mathbf{t}_{11} &= \frac{0.54895 \times Soleil1 + 0.48061 \times Chaleur1 - 0.40218 \times Pluie1}{0.54895^2 + 0.48061^2 + 0.40218^2} \\ &= -0.90285 \end{aligned}$$

On construit ensuite la régression logistique de la qualité sur cette composante PLS. Les résultats apparaissent dans le tableau 8. La régression logistique obtenue exprimée en fonction des variables d'origine sera cohérente, vu les signes des coefficients des variables dans la composante $\mathbf{t}_1$. Il y a maintenant 6 mal classés. Sur les données complètes, la même procédure a également conduit à 6 mal classés.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	-2.1492	0.8279	6.7391	0.0094
INTERCP2	1	2.2845	0.8351	7.4841	0.0062
T1	1	2.6592	0.7028	14.3182	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

Tableau 8 : Régression logistique sur la composante PLS (Température 1924 manquente)

Ainsi, sur notre exemple, la régression logistique PLS et la régression logistique sur composantes PLS ont conduit à des résultats tout à fait comparables. Cette dernière méthode a cependant l'avantage de ne nécessiter aucune programmation et de profiter des performances conjointes des logiciels SIMCA et SAS.

5 Analyse discriminante PLS

Nous avons également étudié l'utilisation de l'analyse discriminante PLS. On calcule pour chaque année les valeurs prédites des variables indicatrices Bon, Moyen, Médiocre par régression PLS de ces variables sur les variables météo et on affecte l'année à la qualité définie par la variable ayant la valeur prédite la plus forte. On aboutit à 12 mal classés lorsqu'on retient une composante PLS (la qualité Moyenne n'est jamais affectée) et à 7 mal classés lorsqu'on retient deux composantes PLS (voir tableau 9). Ainsi, sur cet exemple, nous avons obtenu de meilleurs résultats avec la régression logistique PLS qu'avec l'analyse discriminante PLS.

1 composante

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

Effectif	1	3	Total
1	11	0	11
2	4	7	11
3	1	11	12
Total	16	18	34

2 composantes

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

Effectif	1	2	3	Total
1	11	0	0	11
2	3	6	2	11
3	1	1	10	12
Total	15	7	12	34

Tableau 9 : Prévision de la qualité par régression PLS (Données complètes)

6 Conclusion

Nous avons montré dans cette note que les principes de la régression PLS pouvaient s'étendre sans difficulté à la régression logistique et au modèle linéaire généralisé. Plus généralement la méthodologie présentée permet d'étendre la régression PLS à toute méthode consistant à modéliser linéairement une transformation $g(\pi)$ de la loi de probabilité de la variable de réponse $\mathbf{y}$ en fonction de prédicteurs $\mathbf{X}$ (cf. les procédures LOGISTIC et CATMOD de SAS) ou bien une transformation $g(\mu)$ de la moyenne de $\mathbf{y}$ en fonction des $\mathbf{X}$ (cf. la procédure GENMOD de SAS). Nous avons comparé sur un exemple la régression logistique PLS et la régression logistique sur composantes PLS. Nous n'avons pu conclure à la supériorité d'une approche par rapport à l'autre. D'un point de vue pratique, il faut cependant noter que la seconde approche est immédiate au niveau informatique et réalisable directement avec SAS, du moins lorsqu'il n'y a pas de données manquantes (cf. la procédure PLS de SAS, Tobias, 1996). S'il y a des données manquantes, on peut utiliser conjointement SAS et SIMCA. Marx (1996) a proposé une autre généralisation PLS du modèle linéaire généralisé. Son approche consiste à remplacer par une suite de régressions PLS l'étape " moindres carrés pondérés itérés " de l'algorithme permettant d'obtenir les estimations du maximum de vraisemblance des coefficients de régression de la régression logistique. Notre approche est beaucoup plus simple, s'appuie sur des logiciels existants et est une généralisation immédiate de la méthodologie PLS. Il nous resterait à comparer les deux approches au niveau des résultats pratiques.

References

[1] Marx, B.D. (1996) : Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression. *Technometrics*, vol. 38, n° 4, pp. 374-381.

[2] Tenenhaus, M. (1998) : *La régression PLS.* Paris : Technip.

[3] Tobias, R.D. (1996) : An introduction to Partial Least Squares Regression. SAS Institute Inc., Cary, NC.

[4] Umetri AB (1999) : SIMCA 8.0, Graphical Software for Multivariate Modeling. Umetri AB, Box 7960, S-90719 Umeå, Sweden.

[5] Wold S., Ruhe A., Wold H. & Dunn III, W. J. (1984) : The collinearity problem in linear regression. The Partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, vol. 5, n° 3, pp. 735-743.