Journal de la Société Française de Statistique
Vol. 154 No. 3 (2013)

Régression Bêta PLS

Title: PLS Beta Regression

Frédéric Bertrand[1], Nicolas Meyer[2], Michèle Beau-Faller[3], Karim El Bayed[4], Izzie-Jacques Namer[5] et Myriam Maumy-Bertrand[1]

Résumé : De nombreuses variables d'intérêt, comme par exemple des résultats expérimentaux, des rendements ou des indicateurs économiques, s'expriment naturellement sous la forme de taux, de proportions ou d'indices dont les valeurs sont nécessairement comprises entre zéro et un ou plus généralement deux valeurs fixes connues à l'avance. La régression Bêta permet de modéliser ces données avec beaucoup de souplesse puisque les fonctions de densité des lois Bêta peuvent prendre des formes très variées. Toutefois, comme tous les modèles de régression usuels, elle ne peut s'appliquer directement lorsque les prédicteurs présentent des problèmes de multicolinéarité ou pire lorsqu'ils sont plus nombreux que les observations. Ces situations se rencontrent fréquemment de la chimie à la médecine en passant par l'économie ou le marketing. Pour circonvenir cette difficulté, nous formulons une extension de la régression PLS pour les modèles de régression Bêta. Celle-ci, ainsi que plusieurs outils comme la validation croisée et des techniques bootstrap, est disponible pour le langage R dans la bibliothèque `plsRbeta`.

Abstract: Many responses, for instance experimental results, yields or economic indices, can be naturally expressed as rates or proportions whose values must lie between zero and one or between any two given values. The Beta regression often allows to model these data accurately since the shapes of the densities of Beta laws are very versatile. Yet, as any of the usual regression model, it cannot be applied safely in case of multicollinearity and not at all when the model matrix is rectangular. These situations are frequently found from chemistry to medicine through economics or marketing. To circumvent this difficulty, we derived an extension of PLS regression to Beta regression models. It, as well as several other tools, such as cross validation or bootstrap techniques, is available for the R language in the `plsRbeta` package.

Mots-clés : Régression Bêta, Régression PLS, Régression Bêta PLS, Validation croisée, Techniques bootstrap, Langage R
Keywords: Beta Regression, PLS Regression, PLS Beta Regression, Cross validation, Bootstrap techniques, R language
Classification AMS 2000 : 62F40, 62J07, 62J12, 62P10, 62P20, 62P30

[1] Institut de Recherche Mathématique Avancée, UMR 7501, Université de Strasbourg et CNRS.
E-mail : frederic.bertrand@math.unistra.fr et E-mail : myriam.maumy@math.unistra.fr
[2] Hôpitaux Universitaires de Strasbourg et Faculté de Médecine.
E-mail : nmeyer@unistra.fr
[3] Hôpitaux Universitaires de Strasbourg et INSERM U682.
E-mail : Michele.Faller@chru-strasbourg.fr
[4] RMN et Biophysique des Membranes - ISIS.
E-mail : elbayed@unistra.fr
[5] Service de Biophysique et de Médecine Nucléaire. Hôpitaux Universitaires de Strasbourg.
E-mail : izzie.jacques.namer@chru-strasbourg.fr

 ISSN: 2102-6238

1. Introduction

La régression PLS, fruit de l'algorithme NIPALS initialement développée par Wold (1966) et exposée en détails par Tenenhaus (1998), a déjà été étendue avec succès aux modèles linéaires généralisés par Bastien et al. (2005) et aux modèles de Cox par Bastien (2008).

Nous proposons ici une extension de la régression PLS aux modèles de régression Bêta. En effet, l'intérêt pratique de la loi Bêta a été plusieurs fois affirmé par exemple par Johnson et al. (1995) : "Beta distributions are very versatile and a variety of uncertanties can be usefully modelled by them. This flexibility encourages its empirical use in a wide range of applications."

Plusieurs articles récents se sont intéressés à l'étude de la régression Bêta et de ses propriétés. Mentionnons en particulier, l'article de Ferrari et al. (2004) pour une introduction à ces modèles et ceux de Kosmidis and Firth (2010), Simas et al. (2010) et Grün et al. (2012) pour des extensions ou des amélioration des techniques d'estimation de ces modèles.

Nous supposerons dans la suite de l'article que la réponse étudiée est à valeurs dans l'intervalle $[0;1]$. Le modèle que nous proposons peut bien sûr s'utiliser dès que la réponse Y est à valeurs dans un intervalle borné $[a;b]$, avec $a<b$ fixes et connus, en étudiant $(Y-a)/(b-a)$ à la place de Y.

2. Régression Bêta PLS

2.1. *La régression Bêta*

Lorsqu'elle est non nulle, la fonction de densité de la loi Beta(p,q) est donnée par :

$$\pi(y;p;q)=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}y^{p-1}(1-y)^{q-1},\quad 0<y<1 \tag{1}$$

avec $p>0$, $q>0$ et $\Gamma(\cdot)$ la fonction gamma d'Euler. Si Y suit une loi Beta(p,q), son espérance et sa variance sont égales à :

$$\mathbb{E}(Y)=\frac{p}{p+q}\quad\text{et}\quad \mathrm{Var}(Y)=\frac{pq}{(p+q)^2(p+q+1)}. \tag{2}$$

Afin de pouvoir appliquer des techniques similaires à celles utilisées pour les modèles linéaires généralisés par McCullagh and Nelder (1995), Ferrari et al. (2004) proposent de reparamétriser la loi Bêta de la manière suivante. En posant $\mu=p/(p+q)$ et $\phi=p+q$, c'est-à-dire $p=\mu\phi$ et $q=(1-\mu)\phi$, l'Équation (2) devient :

$$\mathbb{E}(Y)=\mu\quad\text{et}\quad \mathrm{Var}(Y)=\frac{V(\mu)}{1+\phi}$$

où $V(\mu)=\mu(1-\mu)$. Ainsi μ est la valeur moyenne de la réponse et ϕ peut être interprété comme un paramètre de précision puisque, pour un μ fixé, plus la valeur de ϕ est élevée, plus la variance de la réponse est petite. Avec ces nouveaux paramètres, la fonction de densité donnée à l'Équation (1) est égale à :

$$\pi(y;\mu;\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\quad 0<y<1. \tag{3}$$

 ISSN: 2102-6238

Il n'est pas possible d'utiliser directement la théorie du modèle linéaire généralisé, telle qu'introduite par l'article fondateur de Nelder and Wedderburn (1972) puis reprise dans le livre éponyme de McCullagh and Nelder (1995), car celle-ci repose sur l'utilisation d'une famille de lois mise sous une forme exponentielle naturelle (NEF). Or, comme l'a indiqué Morris (1982), la famille des lois bêtas, avec la paramétrisation reprise par Ferrari et al. (2004) et introduite ci-dessus, est une certes une famille exponentielle univariée mais pas une famille exponentielle naturelle. En effet, ces dernières sont caractérisées par leur fonction de variance. Or la fonction de variance de cette paramétrisation de la famille des lois bêtas est déjà celle de la famille exponentielle naturelle formée par les lois binomiales, d'où le résultat. Pour la même raison, il n'est pas possible de caractériser, au sein des familles exponentielles, la famille des lois bêtas à l'aide de leur fonction de variance.

Soient $Y_1, \ldots, Y_n$ des variables aléatoires indépendantes et distribuées suivant la fonction de densité donnée à l'Équation (3) de moyenne μ_t et de précision inconnue ϕ.

Nous obtenons le modèle de régression Bêta en supposant que la moyenne de Y_t, $1 \leqslant t \leqslant n$, peut s'écrire :

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_k \tag{4}$$

où $\beta = (\beta_1, \ldots, \beta_k)'$ est un vecteur de paramètres de régression inconnus ($\beta \in \mathbb{R}^k$), $'$ désignant la transposition, et $x_{t1,}, \ldots, x_{tk}$ sont les observations des k prédicteurs avec $k < n$ qui sont supposées connues et fixes. Enfin, $g(\cdot)$ est une fonction de lien strictement monotone, deux fois dérivable, surjective et définie sur l'intervalle $]0;1[$ et à valeurs dans $\mathbb{R}$. La variance de Y_t est une fonction de μ_t et de ce fait dépend de la valeur des covariables. Par conséquent, le modèle prend automatiquement en compte les éventuels défauts d'homoscédasticité.

Il existe plusieurs choix usuels pour la fonction de lien $g(\cdot)$. Par exemple, le lien logit $g(\mu) = \log \mu/(1-\mu)$, le lien probit $g(\mu) = \Phi^{-1}(\mu)$, avec Φ la fonction de répartition de la loi normale centrée et réduite, le lien complémentaire log-log $g(\mu) = \log(-\log(1-\mu))$, le lien log-log $g(\mu) = -\log(-\log(\mu))$. Une étude détaillée de ces liens a été réalisée par McCullagh and Nelder (1995) et Atkinson (1985) en a proposé d'autres encore. Ici aussi l'utilisation du lien logit permet d'interpréter l'exponentielle des coefficients des covariables en termes d'odds ratio.

2.2. *La régression PLS*

Considérons les variables centrées $Y, x_1, \ldots, x_j, \ldots, x_p$. Soit X la matrice des prédicteurs $x_1, \ldots, x_j, \ldots, x_p$. La régression PLS est bien connue et décrite de manière exhaustive notamment par Höskuldsson (1988), Wold et al. (2001) et Tenenhaus (1998). La présentation classique de la régression PLS est sous forme algorithmique. Nous n'en rappellerons que les éléments utiles pour la suite. La régression PLS est un modèle non-linéaire qui permet de construire des composantes orthogonales t_h obtenues en maximisant les quantités $cov(Y, t_h)$. Soit T la matrice formée de ces composantes, nous avons :

$$Y = Tc' + \varepsilon, \tag{5}$$

où ε est le vecteur des résidus et c' le vecteur des coefficients des composantes.
En posant $T = XW^*$, où W^* est la matrice des coefficients des variables x_j dans chaque composante

 ISSN: 2102-6238

t_h, nous avons l'expression directe de la réponse Y à l'aide des prédicteurs x_j :

$$Y = XW^*c' + \varepsilon. \tag{6}$$

En développant le membre de droite de l'Équation (6), nous obtenons pour chaque composante Y_i de Y :

$$Y_i = \sum_{h=1}^{H} \left(c_h w^*_{1h} x_{i1} + \cdots + c_h w^*_{ph} x_{ip} \right) + \varepsilon_i, \tag{7}$$

H étant le nombre de composantes retenues dans le modèle final avec $H \leqslant \mathrm{rg}(X)$, H étant en général très inférieur au rang de X et p étant égal au nombre de variables contenues dans la matrice X. Les coefficients $c_h w^*_{jh}$, où $1 \leqslant j \leqslant p$, suivant la notation avec $*$ de Wold et al. (2001), traduisent la relation entre le vecteur Y et les variables x_j à travers les composantes t_h.

2.3. *La régression Bêta PLS*

La régression Bêta PLS de la réponse Y sur les variables $x_1, \ldots, x_j, \ldots, x_p$ avec H composantes $t_h = w^*_{1h} x_{i1} + \cdots + w^*_{ph} x_{ip}$ s'écrit :

$$g(\mu) = \sum_{h=1}^{H} \left(c_h \sum_{j=1}^{p} w^*_{jh} x_{ij} \right), \tag{8}$$

où μ est l'espérance de Y. Le lien $g(\cdot)$ est à choisir parmi les liens logit, probit, complémentaire log-log, log-log, Cauchit et log en fonction du type de données et de la qualité de l'ajustement du modèle au données. L'algorithme permettant de déterminer les composantes PLS t_h d'un modèle de régression Bêta PLS est le suivant :

- Calcul de la première composante PLS t_1 :
  1. Calculer le coefficient a_{1j} de x_j dans la régression Bêta de Y sur x_j pour chaque prédicteur x_j, $1 \leqslant j \leqslant p$.
  2. Normer le vecteur colonne a_1 : $w_1 = a_1 / \|a_1\|$.
  3. Calculer la composante $t_1 = 1/(w_1' w_1) X w_1$.
- Calcul de la deuxième composante PLS t_2 :
  1. Calculer le coefficient a_{2j} de x_j dans la régression Bêta de Y sur t_1 et x_j pour chaque prédicteur x_j, $1 \leqslant j \leqslant p$.
  2. Normer le vecteur colonne a_2 : $w_2 = a_2 / \|a_2\|$.
  3. Calculer la matrice résiduelle X_1 de la régression linéaire de X sur t_1.
  4. Calculer la composante $t_2 = 1/(w_2' w_2) X_1 w_2$.
  5. Exprimer la composante t_2 en termes de prédicteurs X : $t_2 = X w^*_2$.
- Nous supposons construites les $h-1$ composantes $t_1, \ldots, t_{h-1}$.
  Calcul de la $h-$ème composante PLS t_h :

 ISSN: 2102-6238

1. Calculer le coefficient a_{hj} de x_j dans la régression Bêta de Y sur $t_1, t_2, \ldots, t_{h-1}$ et x_j pour chaque prédicteur x_j, $1 \leqslant j \leqslant p$.
2. Normer le vecteur colonne a_h : $w_h = a_h/\|a_h\|$.
3. Calculer la matrice résiduelle X_{h-1} de la régression linéaire de X sur $t_1, t_2, \ldots, t_{h-1}$.
4. Calculer la composante $t_h = 1/(w_h' w_h) X_{h-1} w_h$.
5. Exprimer la composante t_h en termes de prédicteurs X : $t_h = X w_h^*$.

Proposition 1. *Les composantes PLS* $(t_h)_{1\leqslant h\leqslant H}$ *sont orthogonales.*

Il est facilement possible de modifier l'algorithme précédent pour pouvoir traiter les jeux de données incomplets (Tenenhaus, 1998).

3. Bootstrap, validations croisées et implémentation logicielle

3.1. Bootstrap

Nous supposons avoir retenu le nombre m adéquat de composantes d'un modèle de régression Bêta PLS de Y sur $x_1, \ldots, x_j, \ldots, x_p$. Nous proposons l'algorithme suivant pour construire des intervalles de confiance et des tests de significativité pour les prédicteurs x_j, $1 \leqslant j \leqslant p$, à l'aide de techniques bootstrap.
Soit $\widehat{F}_{(T|Y)}$ la fonction de répartition empirique étant données la matrice T formée des m composantes PLS et la réponse Y.

Étape 1. Tirer B échantillons de $\widehat{F}_{(T|Y)}$.

Étape 2. Pour tout $b = 1, \ldots, B$, calculer :

$$c^{(b)} = (T^{(b)\prime} T^{(b)})^{-1} T^{(b)\prime} Y^{(b)} \quad \text{et} \quad b^{(b)} = W^* c'^{(b)},$$

où $[T^{(b)}, Y^{(b)}]$ est le $b-$ème échantillon bootstrap, $c'^{(b)}$ est le vecteur des coefficients des composantes et $b^{(b)}$ est le vecteur des coefficients des p prédicteurs d'origine pour cet échantillon et enfin W^* est la matrice fixe des poids des prédicteurs dans le modèle d'origine comportant m composantes.

Étape 3. Pour chaque j, notons Φ_{b_j} l'approximation de Monte-Carlo de la fonction de répartition de la statistique bootstrap de b_j.
Pour chaque b_j, des boîtes à moustaches et des intervalles de confiance peuvent être construits à l'aide des percentiles de Φ_{b_j}. Un intervalle de confiance peut être défini par $I_j(\alpha) =]\Phi_{b_j}^{-1}(\alpha), \Phi_{b_j}^{-1}(1-\alpha)[$ où $\Phi_{b_j}^{-1}(\alpha)$ et $\Phi_{b_j}^{-1}(1-\alpha)$ sont les valeurs obtenues à partir de la fonction de répartition de la statistique bootstrap de telle sorte qu'un niveau nominal de confiance de niveau $100(1-2\alpha)\%$ soit atteint. Afin d'améliorer la qualité de l'intervalle de confiance en termes de taux de couverture, c'est-à-dire la capacité de $I_j(\alpha)$ à fournir les taux de couverture attendus, il est possible d'utiliser plusieurs techniques de construction : normale, percentile ou BC_a (Efron et Tibshirani 1993 ou Davison et Hinkley 1997). Les intervalles ainsi obtenus ne sont pas conçus pour servir à réaliser des comparaisons multiples ou deux à deux et doivent être interprétés séparément.

 ISSN: 2102-6238

3.2. Points forts de l'implémentation logicielle

La bibliothèque de fonctions `plsRbeta` pour le langage R implémente les modèles de régression Bêta PLS. Elle utilise la régression Bêta implémentée dans la bibliothèque de fonctions `betareg` (Cribari-Neto and Zeileis, 2010) pour le langage R afin de réaliser l'étape **1.**.

- Modèles de régression Bêta PLS avec des données complètes ou incomplètes.
- Choix du nombre de composantes grâce à différents critères AIC, BIC, R^2 modifié, arrêt de significativité de la composante t_{m+1} lorsqu'aucun des coefficients a_{m+1} n'est plus significatif dans le modèle à un niveau α' donné ou en utilisant un critère Q^2 estimé par validation croisée.
- Validation croisée « repeated k-fold cross-validation » avec des données complètes ou incomplètes. Les Tableaux 3 et 5 sont des exemples d'utilisation de la validation croisée pour déterminer le nombre adéquat de composantes.
- Bootstrap des coefficients des prédicteurs pour des modèles de régression Bêta PLS avec des données complètes ou incomplètes. Différentes constructions d'intervalles, détaillées dans Efron and Tibshirani (1993) ou Davison and Hinkley (1997), sont disponibles et reposent sur la bibliothèque de fonctions `boot` Canty and Ripley pour le langage R. La Figure 4 est un exemple de boîtes à moustaches construites à partir de la distribution bootstrap des coefficients des prédicteurs d'un modèle de régression PLS Bêta. Les Figures 1, 2 et 5 sont des exemples d'utilisation des techniques bootstrap pour établir la significativité des prédicteurs d'un modèle de régression PLS Bêta.

4. Sélection du nombre de composantes

Un problème crucial pour une utilisation correcte de la régression PLS est la détermination du nombre de composantes. Si, dans le cas de la régression PLS originale, le critère du Q^2 est extrêmement efficace (Tenenhaus, 1998), ses bonnes propriétés disparaissent malheureusement pour les modèles de régression linéaire généralisée PLS. Une étude par simulation s'impose donc afin de déterminer un critère fonctionnel pour choisir le nombre de composantes. Nous proposons de comparer les critères suivants *AIC* et *BIC* (Cribari-Neto and Zeileis, 2010), χ^2 de Pearson, R^2 de Pearson et pseudo-R^2 (Ferrari et al., 2004) ou critères $Q^2\chi^2$ et $Q^2\chi^2$ cumulé estimés par validation croisée en 5 groupes (5-CV) ou en 10 groupes (10-CV) (Bastien et al., 2005).

Les paramètres suivants ont été utilisés pour le plan de simulations.

– Nombre d'individus : 25, 50 et 100.
– Nombre de variables : 10, 25, 50 et 100.
– Nombre de composantes : 2, 4 et 6.
– Dispersion ϕ : 2,5, 5, 10 et 15.

L'algorithme utilisé pour créer les données simulées est une adaptation directe de l'algorithme de Li et al. (2002) qui est lui-même une généralisation multivariée de celui de Naes and Martens (1985). Ce type de généralisation a déjà été utilisé avec succès dans le cas des modèles de régression logistique PLS (Meyer et al., 2010).

Le Tableau 1 est un exemple de résultat pour 25 individus, 10 variables, 2 composantes et un paramètre de dispersion ϕ égal à 2,5. Pour 100 jeux de données simulés, un nombre maximum de 6 composantes devait être calculé. Le nombre moyen de composantes qui a pu être effectivement

 ISSN: 2102-6238

TABLEAU 1. *Choix du nombre de composantes par simulation. Valeur cible 2.*

Critère	Moyenne	Éc-type	Médiane	MAD	Critère	Moyenne	Éc-type	Médiane	MAD
ML.AIC	3.53	0.94	3.5	0.74	$K5.ML.\chi^2$	2.86	2.05	2	1.48
ML.BIC	3.1	0.8	3	0	$K10.ML.Q^2\chi^2_{\text{cum}}$	4.71	1.97	6	0
$ML.\chi^2$	2.84	2.05	2	1.48	$K10.ML.Q^2\chi^2$	0	0	0	0
ML.RSS	5.15	1.31	6	0	$K10.ML.\text{pre}\chi^2$	1.37	0.49	1	0
ML.pseudo-R^2	3.96	1.59	4	2.97	$K10.ML.\chi^2$	2.84	2.05	2	1.48
$ML.R^2$	5.15	1.31	6	0	$K5.BC.Q^2\chi^2_{\text{cum}}$	4.64	1.98	6	0
BC.AIC	3.42	0.87	3	1.48	$K5.BC.Q^2\chi^2$	0.06	0.24	0	0
BC.BIC	3.02	0.79	3	0	$K5.BC.\text{pre}\chi^2$	1.48	0.5	1	0
$BC.\chi^2$	5.3	1.57	6	0	$K5.BC.\chi^2$	2.89	2.04	2	1.48
BC.RSS	5.14	1.33	6	0	$K10.BC.Q^2\chi^2_{\text{cum}}$	4.84	1.87	6	0
BC.pseudo-R^2	3.98	1.58	4	2.97	$K10.BC.Q^2\chi^2$	0.07	0.26	0	0
$BC.R^2$	5.14	1.33	6	0	$K10.BC.\text{pre}\chi^2$	1.54	0.5	2	0
BR.AIC	3.43	0.81	3	1.48	$K10.BC.\chi^2$	2.87	2.04	2	1.48
BR.BIC	3.04	0.74	3	0	$K5.BR.Q^2\chi^2_{\text{cum}}$	4.42	2.02	6	0
$BR.\chi^2$	5.16	1.64	6	0	$K5.BR.Q^2\chi^2$	0.07	0.26	0	0
BR.RSS	5.03	1.28	6	0	$K5.BR.\text{pre}\chi^2$	1.47	0.5	1	0
BR.pseudo-R^2	3.95	1.53	4	1.48	$K5.BR.\chi^2$	2.89	2.06	2	1.48
$BR.R^2$	5.03	1.28	6	0	$K10.BR.Q^2\chi^2_{\text{cum}}$	4.53	1.94	6	0
$K5.ML.Q^2\chi^2_{\text{cum}}$	4.69	1.89	6	0	$K10.BR.Q^2\chi^2$	0.08	0.27	0	0
$K5.ML.Q^2\chi^2$	0	0	0	0	$K10.BR.\text{pre}\chi^2$	1.52	0.5	2	0
$K5.ML.\text{pre}\chi^2$	1.43	0.5	1	0	$K10.BR.\chi^2$	2.87	2.04	2	1.48

calculé, ainsi que le nombre d'échec complet de l'ajustement du modèle de régression bêta sont indiqués dans le Tableau 2. Les abréviations *ML*, *BR*, et *BC* indiquent que les régression Bêta utilisées pour ajuster la régression Bêta PLS sont basées sur le maximum de vraisemblance, la réduction de biais ou la correction de biais (Kosmidis and Firth, 2010). Les préfixes $K5$ et $K10$ signifient que la valeur a été obtenue après une validation croisée en $K = 5$ ou $K = 10$ groupes.

Précisons que le MAD est la médiane des écarts absolus à la médiane. Il s'agit d'un indicateur robuste de dispersion naturellement associé à la médiane. Une valeur de 0 pour le MAD signifie donc que, dans plus de la moitié des simulations, la valeur retenue pour le nombre de composantes est égale à la valeur médiane de toutes les simulations.

De manière générale, les résultats de l'étude par simulations montrent que le $Q^2\chi^2$ (5-CV et 10-CV), déjà connu pour son comportement surprenant en régression logistique PLS (Bastien et al., 2005; Meyer et al., 2010), ne se comporte guère mieux pour les modèles de régression Bêta PLS. La maximisation des critères du R^2 ou du pseudo-R^2, s'avère également inefficace. Les critères *AIC* et *BIC* retiennent systématiquement quelques composantes de trop. Cette tendance est également connue dans le cas de la Régression PLS traditionnelle (Kraemer and Sugiyama, 2011) comme dans celui de la Régression Logistique PLS (Meyer et al., 2010).

Nous constatons que l'algorithme de Régression Bêta PLS proposé parvient à extraire très régulièrement le nombre maximal de composantes demandées et que l'ajustement n'est impossible que dans l'une des 100 simulations uniquement pour la technique BR. La validation croisée, quant à elle, se termine dans 98,5 % des simulations.

 ISSN: 2102-6238

TABLEAU 2. *Nombre maximal de composantes calculées et d'échecs d'ajustement*

	Moyenne	Éc-type	Médiane	MAD	Échecs
ML	5.55	1.22	6	0	0
BC	5.52	1.24	6	0	0
BR	5.43	1.22	6	0	1
ML (5-CV)	5.59	1.17	6	0	1
ML (10-CV)	5.55	1.22	6	0	0
BC (5-CV)	5.63	1.08	6	0	2
BC (10-CV)	5.6	1.13	6	0	1
BR (5-CV)	5.62	1.09	6	0	4
BR (10-CV)	5.6	1.13	6	0	1

5. Exemple d'application en médecine

Les tumeurs cancéreuses représentent l'une des trois principales causes de mortalité dans le monde occidental. La compréhension des mécanismes des pathologies cancéreuses repose actuellement sur l'étude des relations mutuelles des anomalies génétiques acquises, apparaissant dans les tissus au cours du processus de la cancérisation. Ces anomalies sont fréquemment analysées par allélotypages, permettant de déterminer pour un nombre plus ou moins important de sites géniques, la présence ou non d'une modification du nombre de copies de chaque gène. La description multivariée de ces anomalies est informative sur le processus de cancérogénèse. Par ailleurs, l'ensemble de ces sites géniques porteur ou non d'anomalie peut être utilisé pour tenter de prédire certaines caractéristiques cliniques ou biologiques de la tumeur telles que le taux de cellules tumorales sur la biopsie d'une lésion. La modélisation dans un modèle statistique de taux, variable dont l'espace de variation est contenu dans l'intervalle fermé $[0;1]$ comme variable prédite suggère l'utilisation d'une régression Bêta. Par ailleurs, les données d'allélotypage sont caractérisées par une fréquente colinéarité et par une proportion importante de données manquantes. De plus la matrice des données a souvent des dimensions $(i;j)$ telles que $j > i$, ce qui rend la matrice non-inversible, posant des difficultés dans l'ajustement d'un modèle de régression. La régression Bêta de type PLS que nous avons développée est donc particulièrement adaptée pour traiter les données d'allélotypage dans le contexte particulier de la prédiction d'une variable de type taux.

L'exemple est celui de données d'allélotypage obtenues sur une série de 93 patients atteints de différents types de cancer du poumon et comportant 23,2% de valeurs manquantes. La variable prédite est le taux de cellularité tumorale du prélèvement peropératoire de la tumeur. Les variables explicatives sont composées de 56 variables binaires indicatrices de la présence d'une anomalie sur chacun des 56 microsatellites et de trois variables cliniques.

La sélection de variables est dans cette exemple très importante car elle permet de définir un sous-ensemble de prédicteurs, c'est-à-dire de sites géniques, capable de prédire le taux de cellules tumorales. En effet, les pathologies cancéreuses sont des pathologies génétiques acquises et certaines de ces anomalies sont la cause et d'autre la conséquence de la pathologie tumorale. Par ailleurs, l'information contenue dans les différents sites géniques microsatellites est potentiellement redondante. La sélection de variable, séparant les variables jouant probablement un rôle moteur dans le développement tumoral des variables ne faisant que traduire un bruit de fond aléatoire induit par des anomalies causées par ce développement tumoral, est alors une aide

 ISSN: 2102-6238

TABLEAU 3. *Choix du nombre de composantes, méthode BR et lien logit*

Nb Composantes	0	1	2	3	4	5	6	7	8	9
AIC	−23.9	−48.2	−63.4	−76.2	−90.2	−101.5	−114.2	−116.9	−120.5	−121.6
BIC	−18.5	−40.2	−52.7	−62.9	−74.2	−82.8	−92.9	−92.9	−93.8	−92.3
Pred Sig		15	7	1	2	0	2	0	0	0
$Q^2\chi^2$ (5-CV)		−0.3	−1.0	−1.8	−2.8	−3.8	−5.9	−8.9	−10.4	−8.0
χ^2 Pearson	98.3	97.2	96.0	100.0	96.8	91.5	89.1	88.6	87.4	86.4
pseudo-R^2		0.21	0.31	0.40	0.47	0.53	0.58	0.59	0.61	0.62
R^2 Pearson		0.22	0.35	0.41	0.50	0.57	0.64	0.65	0.67	0.68

indispensable à la compréhension des mécanismes sous-jacents de la tumorogenèse.

Le Tableau 3 résume les valeurs de différents critères servant à déterminer le nombre de composantes à utiliser pour modéliser convenablement les données d'allélotypages avec la méthode BR et un lien logit.

Le critère d'arrêt dès l'absence de prédicteur significatif nous inciterait à choisir 6 composantes. Ce choix est confirmé par les critères du pseudo-R^2 ou du R^2 pour lesquels un coude apparaît à partir de 6 composantes.

Le critère BIC nous invite à retenir 8 composantes et le critère AIC un nombre encore plus élevé. Lors de l'étude par simulation, nous avons constaté que ces critères sont libéraux et retiennent généralement quelques composantes de trop.

En fonction des critères, 6 ou 8 composantes sont à retenir pour un lien logit. Compte-tenu de l'étude par simulation précédente, nous décidons de retenir 6 composantes. La même étude a été réalisée avec la méthode BC ou un lien log-log et amène aux mêmes conclusions.

Des intervalles de confiance sont alors obtenus pour chacun des prédicteurs grâce à des échantillons bootstrap de taille 1000 et les techniques normal, basic, percentile ou BC_a.

L'utilisation de techniques bootstrap pour établir la significativité des prédicteurs d'un modèle de régression PLS Bêta est illustré par la Figure 1 pour le cas du modèle à 6 composantes avec la technique de réduction de biais BR.

Au final, nous utilisons la technique BC_a, connue pour ses bonnes propriétés (DiCiccio and Efron, 1996), pour sélectionner les variables significatives au seuil de 5 % en retenant celles pour lesquelles l'intervalle de confiance BC_a ne contient pas 0 comme illustré par la Figure 2 pour le cas du modèle à 6 composantes avec la technique de réduction de biais BR.

Nous retenons ainsi les variables du Tableau 4 comme ayant un effet significatif sur le taux de cellularité tumorale du prélèvement peropératoire de la tumeur. Nous constatons que les deux techniques permettant de traiter le biais de la régression Bêta ont des résultats qui ne diffèrent que par la seule variable EGF3, significative au seuil de 5 % pour la technique BC et non pour la technique BR.

La Figure 3 permet d'évaluer la stabilité des variables significatives sélectionnées par la technique bootstrap BC_a pour un nombre de composantes variant de 1 à 6 et les deux techniques BC et BR d'ajustement de modèle de régression Bêta.

 ISSN: 2102-6238

TABLEAU 4. *Variables significatives*

6 composantes	P4	C3M	RB	FL7A	W2	W4
	MT4	HLA	HLD	HLB	EA3	EA2
Bias Correction	EGF2	EGF3	FL7B	VSFGFR3	VSTOP1	VSTOP2A
	VSEGFR	AFRAEGFR	SRXRA	SMT	SHL	SEB
6 composantes	P4	C3M	RB	FL7A	W2	W4
	MT4	HLA	HLD	HLB	EA3	EA2
Bias Reduction	EGF2	FL7B	VSFGFR3	VSTOP1	VSTOP2A	VSEGFR
	AFRAEGFR	SRXRA	SMT	SHL	SEB	

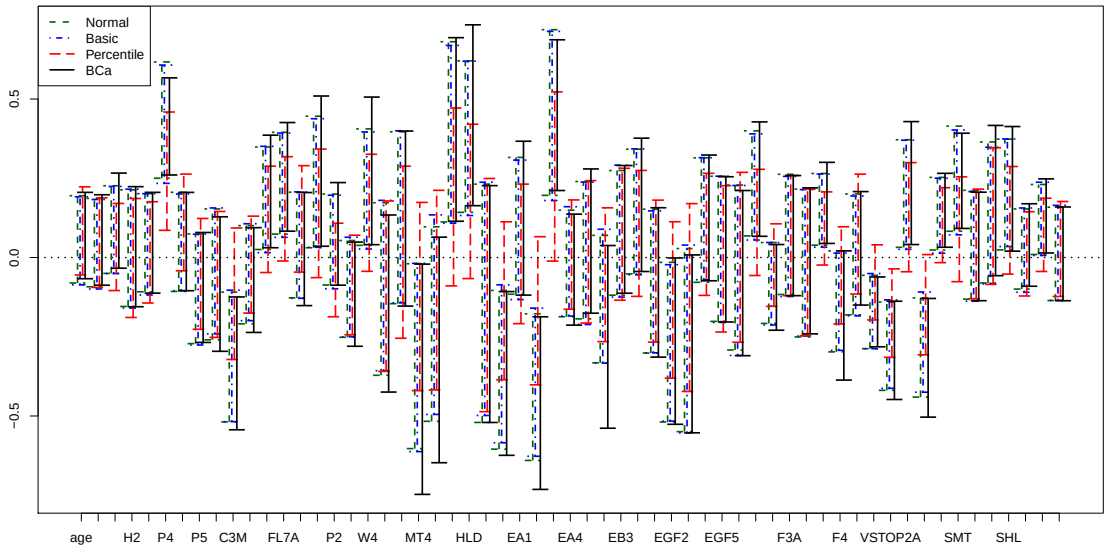

FIGURE 1. *Intervalles de confiance bootstrap à 95%, 6 composantes BR*

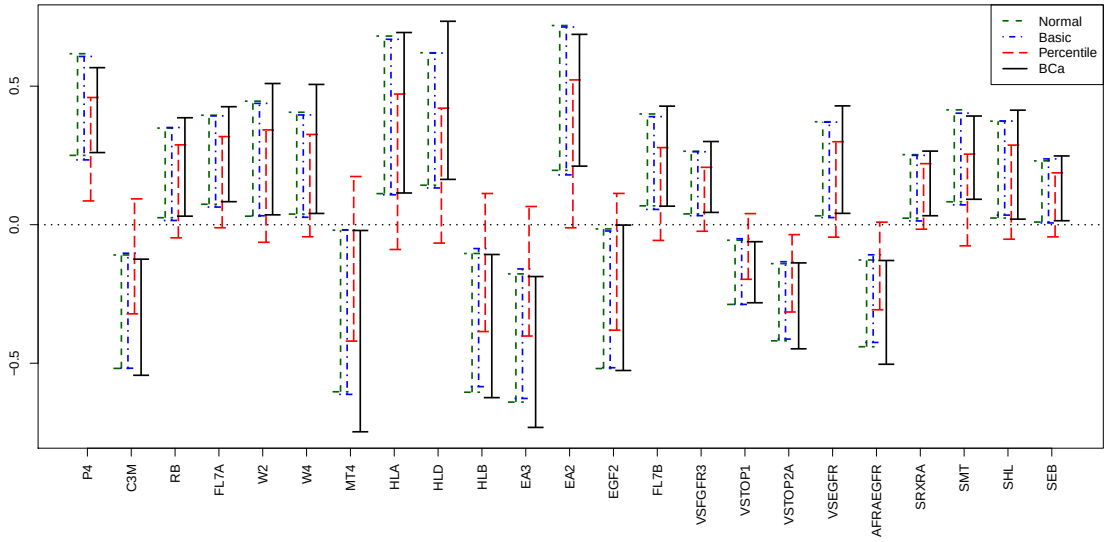

FIGURE 2. *Intervalles de confiance bootstrap à 95%, variables significatives BC_a, 6 composantes BR*

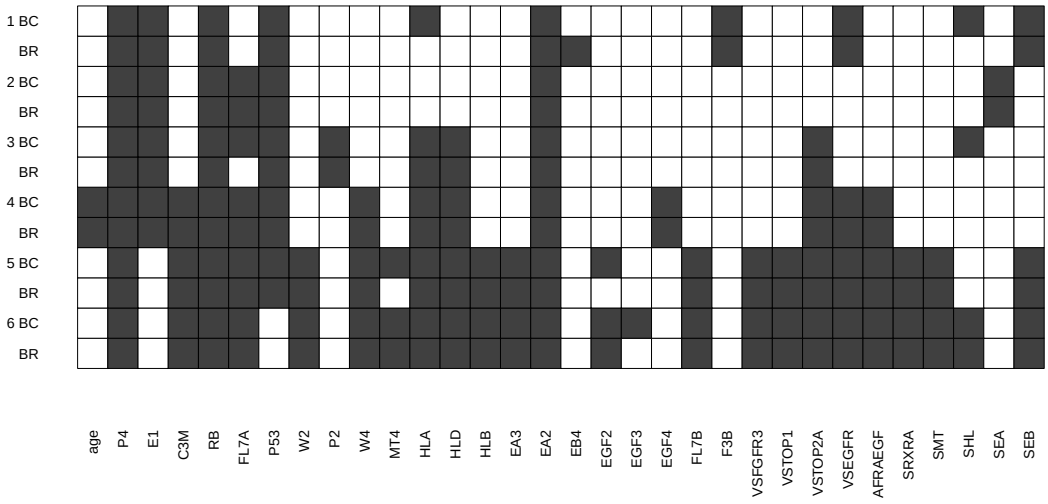

FIGURE 3. *Variables sélectionnées par la technique bootstrap BC_a*

TABLEAU 5. *Choix du nombre de composantes*

Nb Compos	0	1	2	3	4	5	6	7	8	9	10
AIC	−32,35	−110,26	−128,86	−141,93	−152,50	−165,93	−183,71	−193,49	−206,75	−217,21	−231,30
BIC	−27,58	−103,11	−119,33	−130,02	−138,21	−149,25	−164,66	−172,05	−182,93	−191,01	−202,71
$Q^2\chi^2$ (10-CV)		−1,05	−2,68	−1,00	−1,51	−2,34	−3,71	−8,04	−211,19	−5,16*E*3	−4,16*E*4
χ^2 Pearson	76,78	76,44	80,58	76,20	81,95	81,08	74,06	76,39	72,62	73,21	71,82
pseudo-R^2		0,70	0,78	0,82	0,85	0,87	0,89	0,90	0,92	0,93	0,94
R^2 Pearson		0,57	0,64	0,73	0,75	0,81	0,86	0,88	0,90	0,92	0,94

6. Exemple d'application en chimiométrie

L'objectif est de trouver des composés permettant de prédire le taux d'infiltration de patients en cellules cancéreuses à partir de données de spectrométrie. Un patient sain a 0 % de cellules cancéreuses dans une biopsie tandis qu'un patient malade aura dans un biopsie un % d'autant plus élevé que l'échantillon contient des cellules cancereuses. L'intérêt de cette expérience est d'essayer de réduire considérablement le temps d'analyse des biopsies en évitant d'avoir recours à un comptage par un médecin spécialiste en anatomie et cytologie pathologique. Une difficulté statistique supplémentaire apparaît ici : il y a plus de variables (180) que d'individus (80). Plus de détails sur le protocole expérimental suivi sont disponibles dans l'article dans lequel ce jeu de donnés a déjà été initialement publié (Piotto et al., 2012).

Nous commençons par déterminer le nombre de composantes à utiliser pour ajuster un modèle avec toutes les zones du spectre comme variables explicatives. Le tableau 5 contient les éléments de sélection du nombre de composantes. Nous constatons à nouveau la tendance des critères AIC et BIC à retenir un nombre très élevé de composantes, à savoir plus de 10. Par contre, dans cet exemple le critère du $Q^2\chi^2$, estimé à l'aide d'une validation croisée en 10 groupes (10-CV), semble pertinent et nous invite à conserver 3 composantes. Il en va de même du χ^2 Pearson. Le pseudo-R^2 va également dans ce sens puisque nous constatons une décroissance rapide de l'apport explicatif suite à l'ajout de composantes supplémentaires au-delà de la troisième.

Nous construisons alors des échantillons bootstrap de taille 250 des coefficients des prédicteurs pour un modèle à 3 composantes avec la méthode de réduction de biais (BR). La Figure 4 donne les boîtes à moustaches de ces distributions. Des intervalles de confiance sont alors obtenus pour chacun des prédicteurs avec les techniques normal, basic, percentile ou BC_a. Au final, nous utilisons la technique BC_a, connue pour ses bonnes propriétés (DiCiccio and Efron, 1996), pour sélectionner les variables significatives au seuil de 5 % en retenant celles pour lesquelles l'intervalle de confiance BC_a ne contient pas 0 comme illustré par la Figure 5 pour le cas du modèle à 3 composantes avec la technique de réduction de biais BR.

Enfin, nous ajustons un modèle de régression Bêta PLS construit à partir des seuls 79 prédicteurs significatifs pour la technique bootstrap BC_a. Les éléments de sélection du nombre de composantes, non reproduits ici pour des raisons de manque de place, nous amènent à en retenir 2. Nous constatons sur la Figure 6 que ce modèle est validé par le recours à une enveloppe simulée (Atkinson, 1981). Celle-ci est construite, comme recommandé par Atkinson, à partir de 19 statistiques d'ordre.

Comme nous le voyons sur la Figure 7, les deux composantes du modèle formé des seuls

 ISSN: 2102-6238

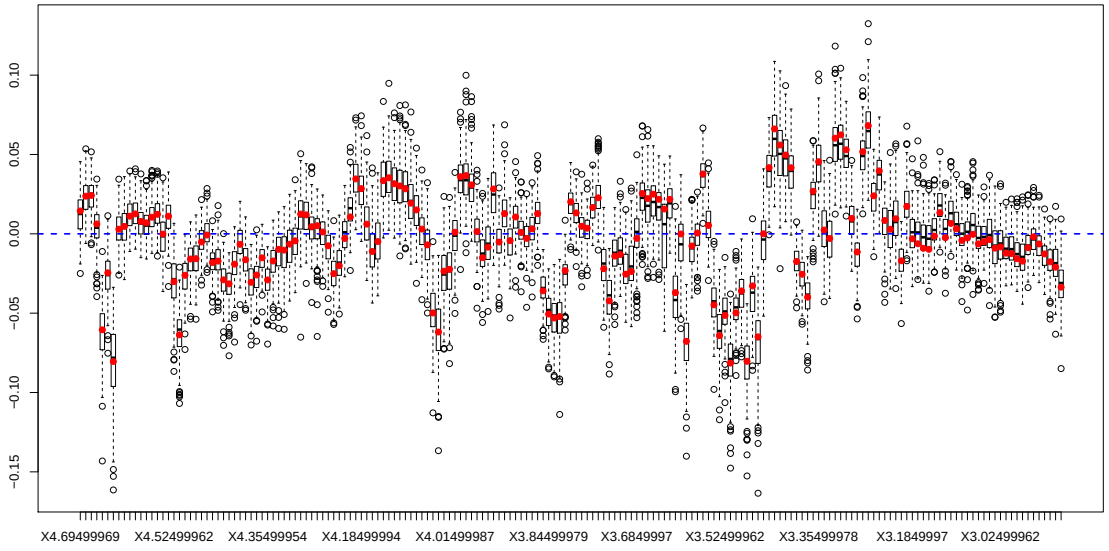

FIGURE 4. *Boîtes à moustaches des distributions bootstrap des prédicteurs, 3 composantes BR*

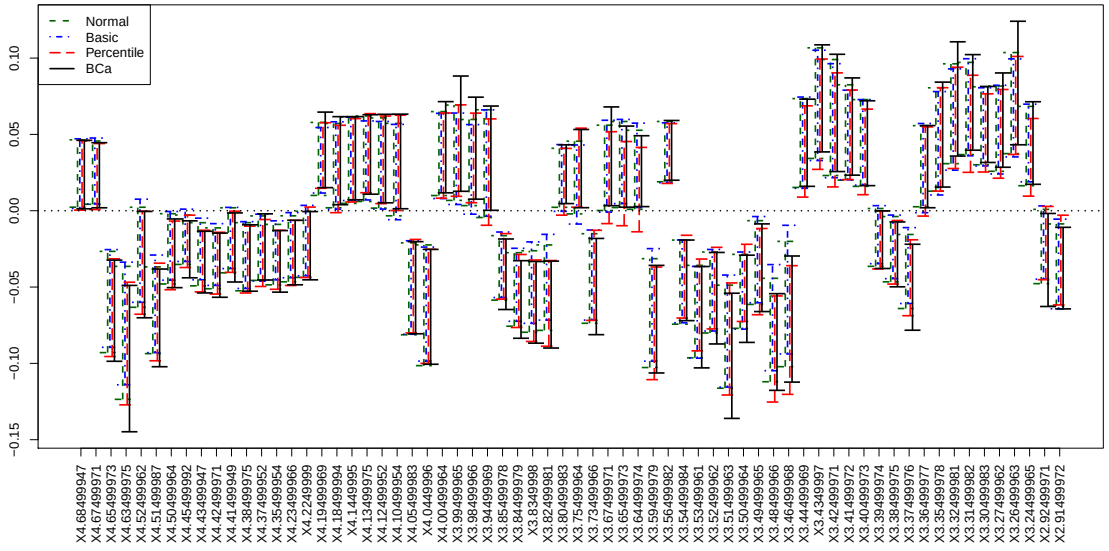

FIGURE 5. *Intervalles de confiance bootstrap à 95%, variables significatives BC_a, 3 composantes BR*

 ISSN: 2102-6238

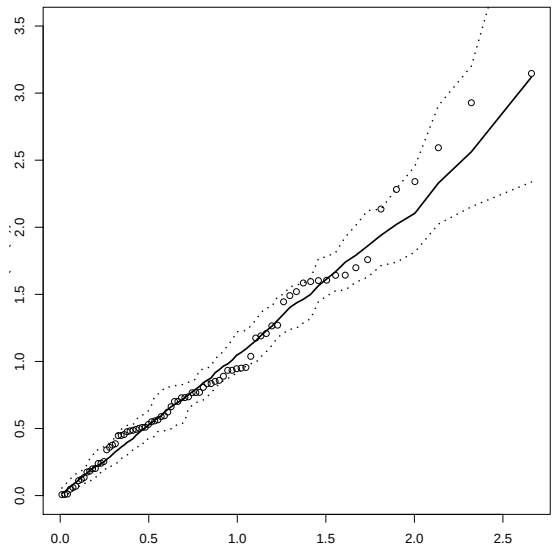

FIGURE 6. *Résidus et enveloppe simulée*

79 prédicteurs significatifs pour la technique bootstrap BC_a permettent une séparation linéaire franche entre les sujets sains, en vert foncé, et les sujets malades dont l'état est d'autant plus grave que la couleur se rapproche du rouge. Pour permettre une évaluation graphique de la qualité prédictive du modèle, nous représentons les valeurs observées en fonction des valeurs prédites sur l'échelle logit, Figure 8, et sur l'échelle originale Figure 9. Remarquons que, comme le passage de l'un à l'autre des deux graphiques se fait à l'aide d'un transformation continue, bijective et croissante des axes, le coefficient de corrélation de Kendall entre les valeurs observées et les valeurs prédites sera identique pour les deux échelles. Pour les deux représentations graphiques, nous mesurons une association au sens du coefficient de corrélation de Kendall entre les valeurs observées et les valeurs prédites égale à $0,72$ pour une p-valeur de $7,03E-19$, donc significative au seuil de 1%.

Notons toutefois que les données de cet exemple sont en fait formées par le mélange d'individus sains pour lesquels le taux d'infiltration est nul et d'individus malades pour lesquels ces taux sont non-nuls et varient de 0 à 1. Ainsi l'utilisation d'un modèle de régression PLS basé sur la combinaison d'un modèle de régression PLS logisitique et d'un modèle de régression logistique Bêta ou d'un modèle de régression Bêta avec inflation de 0 introduit par Ospina and Ferrari (2012) serait vraisemblablement judicieuse et est actuellement en cours de réalisation par les auteurs.

7. Conclusion et perspectives

Notre objectif a été de proposer une extension de la régression PLS aux modèles de régression Bêta, puis de la mettre à la disposition des utilisateurs du langage libre R.

Nous offrons ainsi la possibilité de travailler, pour modéliser des taux ou des proportions, avec des prédicteurs colinéaires, difficulté inévitable dans le cas de la modélisation des mélanges ou lors de l'analyse de spectres, de l'étude de données génétiques, protéomiques ou métabonomiques.

 ISSN: 2102-6238

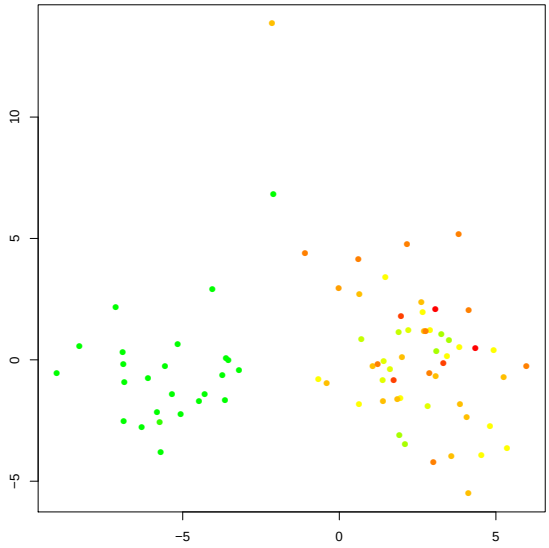

FIGURE 7. *Représentation des individus sur le plan formé par les deux premières composantes*

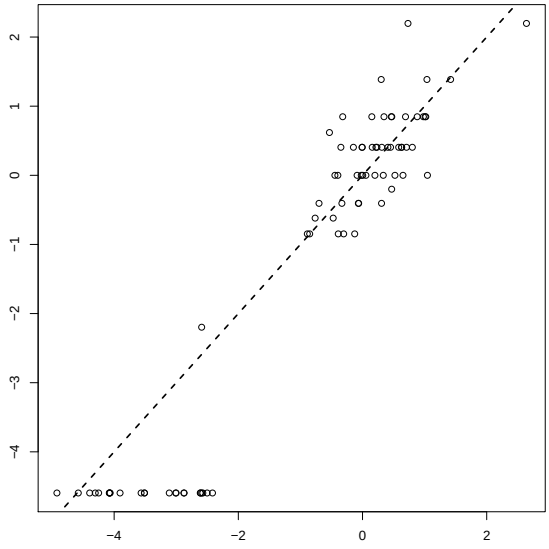

FIGURE 8. *Valeurs observées en fonction des valeurs prédites sur l'échelle* logit

 ISSN: 2102-6238

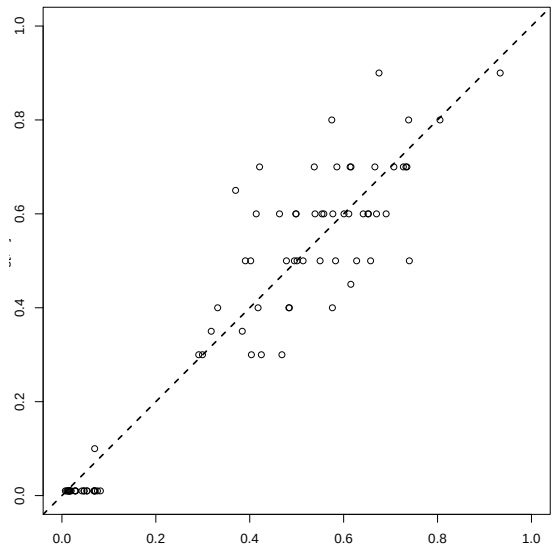

FIGURE 9. *Valeurs observées en fonction des valeurs prédites sur l'échelle originale*

De plus, la régression Bêta PLS peut être aussi appliquée à des jeux de données incomplets. Il est également possible dans ce cas, comme dans celui des données complètes, de sélectionner le nombre de composantes par validation croisée « repeated k-fold cross-validation ».

Enfin, nous proposons des techniques bootstrap afin de, par exemple, tester la significativité de chacun des prédicteurs présents dans le jeu de données et ainsi valider les modèles construits.

L'étude de deux jeux de données réels a permis aux outils proposés de démontrer leur efficacité.

Remerciements : Les auteurs sont reconnaissants aux deux arbitres pour leurs commentaires et suggestions qui ont permis d'améliorer la qualité de plusieurs points de l'article ainsi qu'à Jean-Pierre Gauchi pour sa disponibilité et ses remarques pertinentes et constructives.

Références

Atkinson, A. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68(1) :13–20.

Atkinson, A. (1985). *Plots, Transformations and Regression : An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press, Oxford.

Bastien, P. (2008). Deviance residuals based PLS regression for censored data in high dimensional setting. *Chemometrics and Intelligent Laboratory Systems*, 91(1) :78–86.

Bastien, P., Esposito Vinzi, V., and Tenenhaus, M. (2005). Pls generalised linear regression. *Computational Statistics & Data Analysis*, 48(1) :17–46.

Canty, A. and Ripley, B. (2009). `boot` : Bootstrap R (S-Plus) Functions. R package version 1.2-37.

Cribari-Neto, F. and Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, 34(2) :1–24.

Davison, A. and Hinkley, D. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

DiCiccio, T. and Efron, B. (1996). Bootstrap confidence intervals (with discussion). *Statistical Science*, 11 :189–228.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Ferrari, S., , and Cribari-Neto, F. (2004). Beta Regression for Modeling Rates and Proportions. *Journal of Applied Statistics*, 31(7) :799–815.

 ISSN: 2102-6238

Grün, B., Kosmidis, I., and Zeileis, A. (2012). Extended Beta Regression in R : Shaken, Stirred, Mixed and Partitioned. *Journal of Statistical Software*, 48(11) :1–25.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, 2 :211–228.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, volume 2. Wiley, New York, 2nd edition.

Kosmidis, I. and Firth, D. (2010). A Generic Algorithm for Reducing Bias in Parametric Estimation. *Journal of Chemometrics*, 4 :1097–1112.

Kraemer, N. and Sugiyama, M. (2011). The Degrees of Freedom of Partial Least Squares Regression. *Journal of the American Statistical Association*, 106(494) :697–705.

Li, B., Morris, J., and Martin, E. (2002). Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 64 :79–89.

McCullagh, P. and Nelder, J. (1995). *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, 2nd edition.

Meyer, N., Maumy-Bertrand, M., and Bertrand, F. (2010). Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives : application aux donnés d'allélotypage. *Journal de la Société Française De Statistique*, 151(2) :1–18.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10 :65–80.

Naes, T. and Martens, H. (1985). Comparison of prediction methods for multicollinear data. *Communications in Statistics – Simulation and Computation*, 14 :545–576.

Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3) :370–384.

Ospina, R. and Ferrari, S. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(1) :1609–1623.

Piotto, M., Moussallieh, F.-M., Neuville, A., Bellocq, J.-P., Elbayed, K., and Namer, I. (2012). Towards real-time metabolic profiling of a biopsy specimen during a surgical operation by 1h high resolution magic angle spinning nuclear magnetic resonance : a case report. *Journal of Medical Case Reports*, 6(1).

Simas, A., Barreto-Souza, W., and Rocha, A. (2010). Improved Estimators for a General Class of Beta Regression Models. *Computational Statistics & Data Analysis*, 54(2) :348–366.

Tenenhaus, M. (1998). *La régression PLS : Théorie et Pratique*. Technip, Paris.

Wold, H. (1966). Estimation of principal component and related models by iterative least squares. In Krishnaiah, P., editor, *Multivariate Analysis*, pages 391–420. Academic Press, New York.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression : a basic tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58 :109–130.

 ISSN: 2102-6238