Apport du bootstrap à la régression PLS : application à la prédiction de la qualité des gazoles

S. Aji, S. Tavolaro, F. Lantz, A. Faraj

Oil & Gas Science and Technology – Rev. IFP, Vol. 58 (2003), No. 5, pp. 599-608

Apport du *bootstrap* à la régression PLS : application à la prédiction de la qualité des gazoles

S. Aji[1], S. Tavolaro[1], F. Lantz[2], A. Faraj[1]

1 Institut français du pétrole, 1 et 4, avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex - France
2 École du pétrole et des moteurs, 228-232, avenue Napoléon-Bonaparte, 92852 Rueil-Malmaison Cedex - France
e-mail : salaheddine.aji@ifp.fr - santiago.tavolaro@ifp.fr - frederic.lantz@ifp.fr - abdelaziz.faraj@ifp.fr

Résumé — L'objectif de la modélisation que nous développons s'intègre dans un processus de contrôle de qualité des produits élaborés en raffinerie. Nous construisons un modèle de prédiction statistique des propriétés chimiques des gazoles à partir des spectres proches infrarouges (PIR) de distillats moyens. Le grand nombre de variables explicatives ainsi que leur forte multicolinéarité préconisent une modélisation par régression PLS *(partial least squares)*. La détermination des intervalles de prédiction nous amène à utiliser des techniques de *bootstrap*. Pour les mettre en œuvre, nous utilisons les propriétés du modèle PLS en tant que méthode de régression linéaire multiple à partir des composantes PLS orthogonales. Nous donnons ensuite des approximations de la distribution des coefficients ainsi que de la distribution des erreurs de prédiction.

Abstract* — *Contribution of Bootstrap Techniques to PLS Regression: Application to the Prediction Model of Gas-Oil Quality Control — *The objective of our modelling approach is a part of the quality control process of refining products. We build a prediction model of gas oil chemical properties from near infrared spectroscopy (NIR) of mid-distillates. The large number of explanatory variables and their high level of multicolinearity leads to the use of PLS (partial least squares) regression. Then, bootstrap techniques are used to determine prediction intervals. We consider the PLS model as a multiple linear regression on PLS orthogonal components to implement these bootstrap methods. Thus, we approximate the coefficients and prediction errors distributions.*

INTRODUCTION

La détermination des propriétés des gazoles repose classiquement sur un ensemble d'analyses. La construction d'un modèle de prédiction de ces propriétés à partir des spectres proches infrarouges (PIR) permet d'envisager une réduction des coûts et des délais dans le processus de contrôle de qualité des produits élaborés en raffinerie. La quantification des incertitudes inhérentes au modèle statistique ainsi établi est nécessaire pour permettre aux chimistes de le valider.

Ce travail fait suite à Aji *et al.* (2003) dont il permet d'approfondir la méthodologie statistique. Généralement, le nombre de variables longueurs d'onde caractérisant les spectres est très élevé par rapport au nombre d'observations. Par ailleurs, ces variables présentent souvent une forte multicolinéarité. La régression PLS (partial least squares) est une méthode d'analyse des données particulièrement adaptée à l'étude des relations, souvent complexes, entre deux tableaux de données *X* et *Y* (Wold *et al.*, 1983).

Les termes d'erreur du modèle n'ont pas forcement une distribution gaussienne. Nous nous intéressons ici à l'application des techniques *bootstrap* pour déterminer des intervalles de prédiction à partir d'une modélisation par régression PLS. Nous abordons différents problèmes liés à cette application et notamment la détermination du nombre de réplications.

Le bootstrap (Efron, 1979) fournit une approximation de la distribution inconnue d'un estimateur par une distribution empirique obtenue à partir d'une procédure de rééchantillonnage basée sur des tirages aléatoires avec remise dans les données. Son utilisation permet d'améliorer l'estimation des intervalles de prédiction (Stine, 1985). Au cours de ce travail, nous nous sommes intéressés aux intervalles bootstrap (percentile et percentile-*t*).

Dans la première section, nous présentons le problème de modélisation statistique de la qualité des gazoles et les caractéristiques de la base d'étalonnage utilisée pour la prédiction. Ces données nous permettent d'étudier les performances et les caractéristiques essentielles du modèle de régression PLS appliqué à un problème de caractérisation des produits élaborés en raffinerie. La deuxième section est consacrée à la détermination du modèle de prédiction. La quantification des incertitudes liées au modèle PLS en utilisant les techniques bootstrap constitue la troisième section. Pour finir, nous déterminons l'incidence du nombre de réplications bootstrap sur les intervalles de prédictions.

Dans l'Annexe, nous présentons un aperçu de la méthode de régression PLS. Nous abordons ensuite les techniques bootstrap pour ces modèles, les différents problèmes liés à leur mise en œuvre ainsi que les intervalles de confiance s'y référant.

1 MODÈLE D'ESTIMATION PLS DE LA QUALITÉ DES GAZOLES

Le modèle de prédiction statistique a pour avantage d'estimer les propriétés chimiques des gazoles (au nombre de 11) sans procéder à des expériences en laboratoire. On explique ainsi la qualité d'un gazole à partir de son spectre PIR. Cette approche renvoie à une importante littérature pour laquelle on pourra se reporter à Marteau *et al.* (1992).

Pour illustrer nos propos, nous présentons ici les résultats d'une des propriétés, à savoir la concentration en pourcentage poids total-aromatique par ultraviolet. Cette propriété a été auparavant décomposée en trois classes, dont nous présentons les résultats de la première (valeurs comprises entre 0 et 45 %). La figure 1 illustre les $n = 69$ observations pour les 2232 variables longueurs d'onde caractérisant les spectres (la bande spectrale étant fixée entre 4900-9200 cm^{-1}).

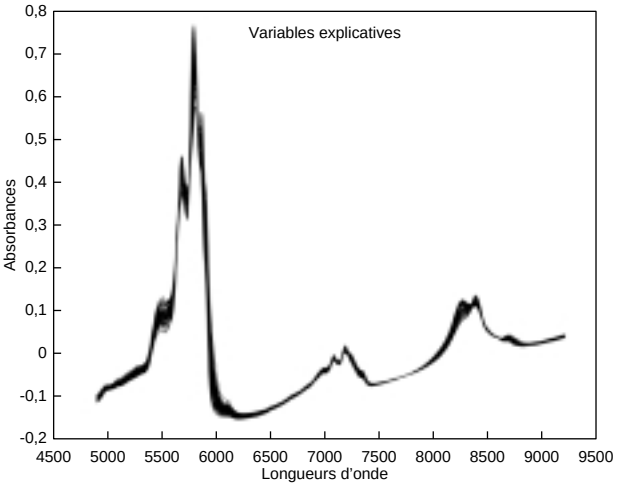

Figure 1

Base d'étalonnage : spectres proches infrarouges des échantillons de gazole.

Near infrared spectra of calibration base samples.

Les variables explicatives présentent une forte multicolinéarité. Celle-ci se traduit souvent par le fait que les variances des estimateurs des moindres carrés ordinaires (MCO) sont élevées et que le modèle est instable. De plus, dans le jeu de données que nous avons traité, le nombre d'observations est largement inférieur au nombre de variables explicatives. La régression PLS (moindres carrés partiels) est alors une alternative très efficace aux particularités du problème et des données à exploiter. C'est une méthode d'analyse des données spécifiquement construite pour l'étude des relations entre un ensemble de variables réponses *Y* et un ensemble de variables explicatives *X* lorsque la multicolinéarité est forte ou lorsque

le nombre de variables explicatives et/ou réponses est élevé par rapport au nombre d'individus.

La régression PLS effectue une analyse en composantes principales pour chaque ensemble de variables *X* et *Y*, sous la contrainte que les composantes de *X* soient fortement corrélées à celles de *Y*. L'algorithme NIPALS *(voir Annexe)* permet de déterminer séquentiellement les paramètres de la régression PLS. Outre les avantages précédents, elle permet d'exprimer le modèle sous-jacent directement à partir des données initiales *X* et *Y*, mais aussi à partir des composantes PLS.

On distingue habituellement le cas où la variable à expliquer est un vecteur (régression PLS1) de celui où il s'agit d'une matrice (régression PLS2). Dans cet article, chacune des propriétés étudiées est estimée individuellement (PLS1), sachant que le principe est le même pour la régression PLS2.

2 NOMBRE DE COMPOSANTES PLS POUR LA PRÉDICTION

Une des particularités de la régression PLS est que le modèle obtenu dépend du nombre de composantes h ($h = 1, \ldots,$ *rang (X)* pour la PLS1). Il intervient lors du déroulement de l'algorithme NIPALS. Ce nombre est alors considéré comme un paramètre qu'il faut choisir avec soin. Pour décider de la dimension du modèle, deux considérations sont à prendre en compte : la qualité de l'ajustement aux données (pourcentage de variance expliquée) et la qualité de prédiction (somme des carrés des erreurs de prédiction ou *PRESS (PRediction Error Sum of Squares)*) par validation croisée (Wold, 1978).

2.1 Pourcentage de variance expliquée par les composantes PLS

Les pourcentages de variance expliquée de *X* et de *Y* obtenus pour un nombre de composantes de 1 à 10 sont présentés dans le tableau 1. À partir de celui-ci, nous constatons que le modèle de régression PLS avec six composantes peut être satisfaisant. Il explique 98,38 % de la variance de *X* et 98,80 % de la variance de *Y*. Ce résultat montre la qualité des composantes PLS qui permettent d'expliquer aussi bien les variables explicatives que la variable réponse. Il met en évidence la forte multicolinéarité présentée par les variables caractérisant les spectres PIR.

TABLEAU 1

Pourcentage de variance expliquée par les composantes PLS

Proportion of variance explained by PLS components

Nombre de composantes	Variables explicatives *X*		Variable réponse *Y*	
	% de variance	% cumulé	% de variance	% cumulé
1	54,48	54,48	58,43	58,43
2	20,87	75,34	34,16	92,59
3	15,18	90,52	1,92	94,51
4	4,42	94,94	2,03	96,53
5	1,97	96,91	1,33	97,87
6	1,47	98,38	0,93	98,80
7	0,25	98,63	0,24	99,04
8	0,33	98,96	0,05	99,09
9	0,39	99,35	0,03	99,12
10	0,22	99,57	0,08	99,20

2.2 La validation croisée

La validation croisée est une méthode générale de sélection des modèles de prédiction. Elle détermine le modèle qui a la plus petite erreur de prédiction parmi plusieurs candidats. L'idée de base est de séparer l'échantillon initial en deux sous-échantillons. Le modèle de prédiction est construit à partir du premier (apprentissage) et le deuxième est utilisé pour en tester la qualité (validation). Au cours de ce travail, nous utilisons la méthode de validation croisée *jackknife*. Celle-ci consiste à construire les modèles pour tous les nombres de composantes h possibles sur n–1 observations, puis à tester le modèle sur l'observation exclue. On réitère n fois la procédure, obtenant ainsi les $PRESS_h$ relatifs à l'estimation de la propriété considérée. Nous remarquons sur la figure 2 que l'erreur entre les modèles à 6 et 7 composantes se stabilise. Ce résultat confirme le choix du nombre de composantes obtenu à partir du pourcentage de variance expliquée.

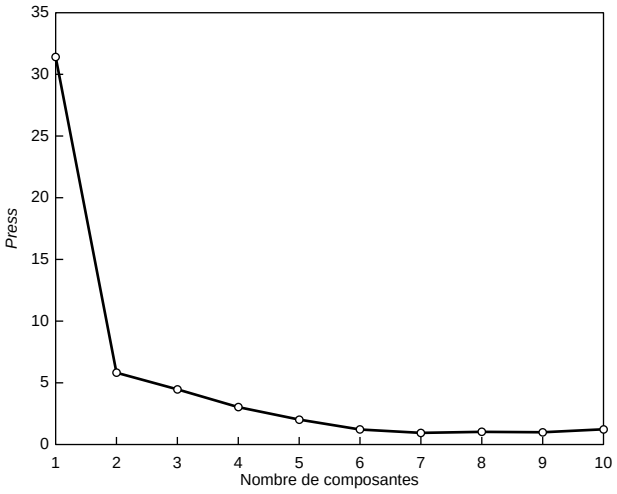

Figure 2

Évolution de la somme des carrés des erreurs de prévision $PRESS_h$.

Prediction error sum of squares ($PRESS_h$).

3 QUANTIFICATION DES INCERTITUDES DU MODÈLE DE PRÉDICTION

Une fois le modèle de prédiction déterminé, il nous reste à juger de sa qualité pour le valider. Pour cela, nous nous

sommes intéressés à l'erreur entre la variable réponse *Y* (mesure des chimistes) et les prédictions correspondantes. Par ailleurs, il faut associer un intervalle de confiance aux valeurs prédites par le modèle. Cependant, les distributions des prédictions étant inconnues, deux approches sont alors envisagées. La plus communément utilisée établit une hypothèse de normalité des distributions (intervalle de prédiction standard). L'autre approche, sans hypothèse, s'appuie sur une distribution empirique bootstrap.

Les composantes PLS de *X* ayant été déterminées, le modèle prédictif de *Y* s'exprime sous deux formes équivalentes. La première (modèle PLS) est directement fonction des variables explicatives *X*. La deuxième est construite par une régression linéaire multiple à partir des composantes PLS orthogonales. Ceci nous permet de mettre en œuvre les techniques bootstrap pour le modèle de régression linéaire sur les composantes PLS. Nous déterminons ainsi les intervalles de prédiction. Par l'intermédiaire d'une matrice de passage, nous obtenons les intervalles des coefficients du modèle PLS exprimés à partir de *X (voir Annexe)*.

Les intervalles de confiance envisagés sont les intervalles bootstrap percentile et percentile-*t* ainsi que l'intervalle de confiance standard. L'intervalle percentile-*t* commet une erreur asymptotique en $O(n^{-3/2})$, le percentile et le standard en $O(n^{-1})$ (Hall, 1992) avec, pour ce dernier, l'hypothèse gaussienne de la distribution inconnue. Nous remarquons ainsi la supériorité des intervalles bootstrap pour notre étude, qui sont construits sans établir d'hypothèse et plus particulièrement, la méthode percentile-*t* pour sa vitesse de convergence.

3.1 Intervalles de confiance des coefficients de régression PLS

Nous nous sommes intéressés aux intervalles des coefficients pour nous assurer qu'ils étaient correctement contrôlés. Nous présentons dans la figure 3 les intervalles de confiance percentile-*t* des coefficients pour un niveau de confiance à 95 % *(voir Annexe A.2.1)*. Les longueurs d'onde les plus explicatives à la prédiction de *Y* sont les valeurs élevées en valeurs absolues. Inversement, les valeurs proches de zéro contribuent faiblement à l'estimation de *Y*.

Nous constatons que, plus les coefficients sont élevés en valeur absolue, plus leur intervalle de confiance est large. L'intervalle percentile-*t* permet de prendre en compte l'asymétrie de l'intervalle *(voir Annexe A.2)*. Les intervalles standard sont symétriques et supposent la normalité des distributions. Il s'ensuit que les bornes de ces deux méthodes sont d'autant plus différentes que l'asymétrie de l'intervalle bootstrap est forte. Le coefficient d'asymétrie[1] de la figure 4 met en évidence ce phénomène.

3.2 Les intervalles de prédiction

La qualité prédictive du modèle *(voir fig. 5)* permet au chimiste de valider le modèle PLS de prédiction selon la précision des résultats attendus. L'asymétrie des intervalles percentile-*t* comparés aux intervalles standard, symétriques par construction *(fig. 6)*, montre l'intérêt des techniques

1 $\frac{\text{(borne supérieure de l'intervalle – valeur estimée)}}{\text{étendue de l'intervalle}}$

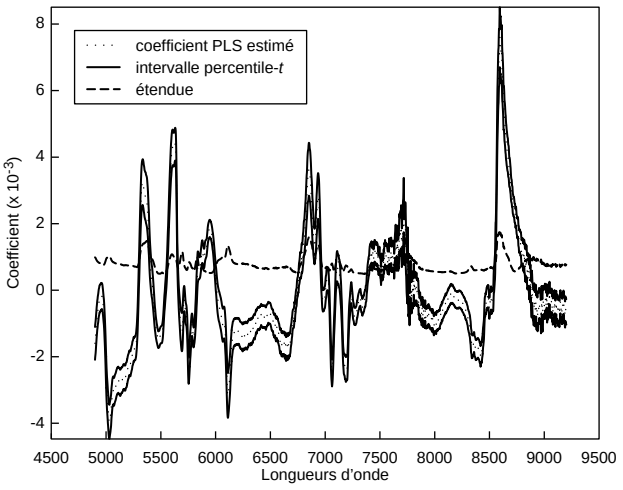

Figure 3

Intervalle de confiance percentile-*t* à 95 % des coefficients de régression.

Percentile-t confidence interval of the regression coefficients.

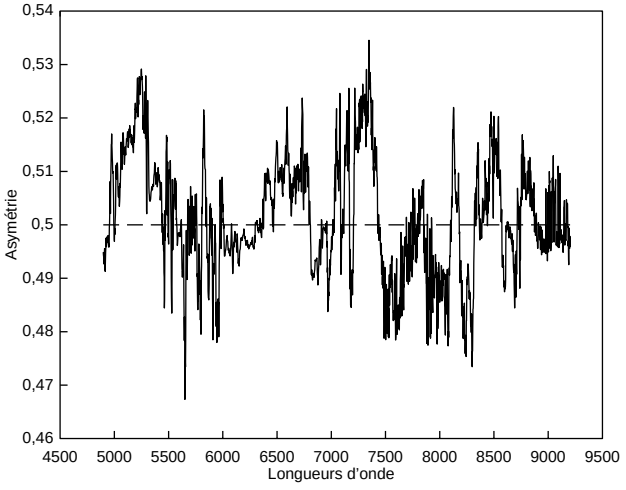

Figure 4

Coefficient d'asymétrie de l'intervalle percentile-*t* des coefficients de régression.

Asymmetric coefficient of percentile-t confidence interval of the regression coefficients.

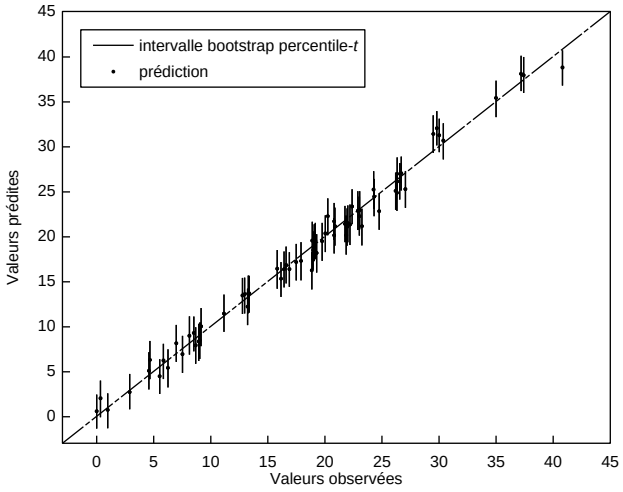

Figure 5

Intervalle percentile-*t* à 95 % des prédictions.
Percentile-t prediction interval.

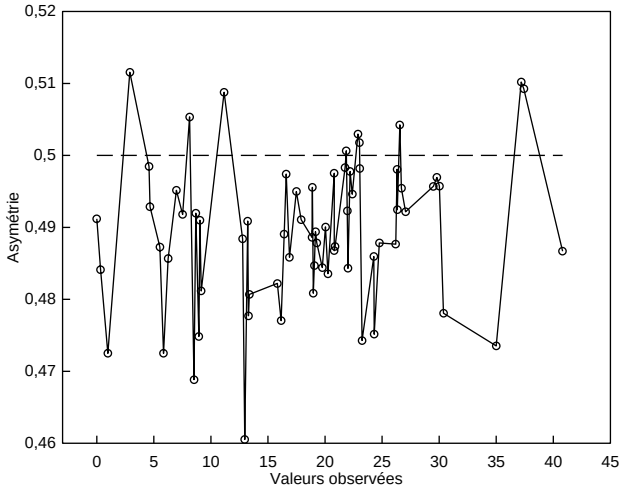

Figure 6

Coefficients d'asymétrie de l'intervalle percentile-*t* des prédictions.
Asymmetric coefficients of percentile-t prediction interval.

bootstrap *(voir Annexe A.2.2)*. La plupart des observations montrent que les coefficients d'asymétrie des intervalles percentile-*t* sont inférieurs à 0.5. Les prédictions sont donc plus proches de leur borne supérieure.

Plus la différence entre les intervalles bootstrap et standard augmente, plus les distributions inconnues des prédictions sont différentes d'une distribution gaussienne. Elle se répercute directement sur les étendues *(fig. 7)*. La faible différence entre les intervalles percentile et percentile-*t* s'explique par leur vitesse de convergence. Ainsi la différence des deux intervalles bootstrap est d'autant plus petite que le nombre d'observations est grand.

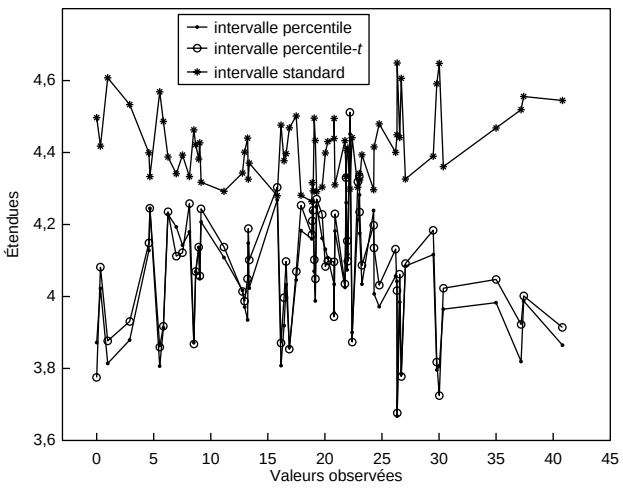

Figure 7

Étendues des intervalles de prédiction.
Confidence prediction intervals ranges.

La classe (0-45 %) de la propriété exposée contient beaucoup d'observations en comparaison aux deux autres classes (30-70 % et 55-100 %) qui n'en ont qu'une vingtaine chacune. Bien évidemment, les interprétations des méthodes précédemment exposées restent les mêmes pour ces deux dernières. Cependant, les différences de résultats entre intervalles percentile et percentile-*t* sont accentuées pour les échantillons de plus petite taille et conduisent à privilégier les intervalles percentile-*t*.

4 LE NOMBRE D'ÉCHANTILLONS BOOTSTRAP

La construction des intervalles de confiance bootstrap pour un niveau à $(1-2\alpha)$ d'une prédiction repose sur les fractiles α et $(1-\alpha)$ de la distribution empirique bootstrap. Nous nous sommes intéressés à l'incidence du nombre de réplications sur la détermination des fractiles afin de définir le nombre minimum de réplications nécessaires pour obtenir des fractiles qui varient peu.

Pour un nombre de réplications B fixe, l'erreur due au rééchantillonnage sur le niveau de confiance de l'intervalle est en $O(B^{-1})$. Si l'on choisit B, tel que α ou $(1-\alpha)$ soit un multiple de $(B+1)^{-1}$, alors cette erreur n'est plus qu'en $O((nB)^{-1})$ (Hall, 1986). Dans ce cas, l'erreur d'estimation de la version empirique bootstrap, à partir de B réplications, estimant la version bootstrap est négligeable. On a ainsi $(B+1)$ valeurs de la fonction de répartition empirique à partir des B valeurs bootstrap : 0, $1/B$, $2/B$, …, $(B-1)/B$, 1. Le fractile α (resp. fractile $(1-\alpha)$) est déterminé par la $\alpha(B+1)$ième (resp. $(1-\alpha)(B+1)$ième) valeur de la liste ordonnée des B réplications bootstrap.

Le processus mis en œuvre consiste, pour un certain nombre de valeurs de *B*, à effectuer un ensemble de simulations, notées *k*, visant à juger de la stabilité des résultats. Plutôt que d'étudier la variabilité de chacun des fractiles séparément, nous considérons l'étendue de l'intervalle entre ces derniers. Pour un nombre de réplications *B* donné, les simulations permettent donc d'obtenir *k* intervalles de confiance et leur étendue. La figure 8 présente l'écart type de $k = 30$ étendues percentile-*t* pour plusieurs valeurs de *B* vérifiant $(B + 1)^{-1}$ multiple de $\alpha = 0{,}025$ (pour un intervalle bilatéral à 95 %) et pour chacune des prédictions. Nous remarquons que l'évolution de l'écart type entre $B = 2199$ et $B = 2799$ se stabilise. On considère alors le gain d'approximation comme faible pour $B > 2199$ par rapport au nombre de réplications supplémentaires.

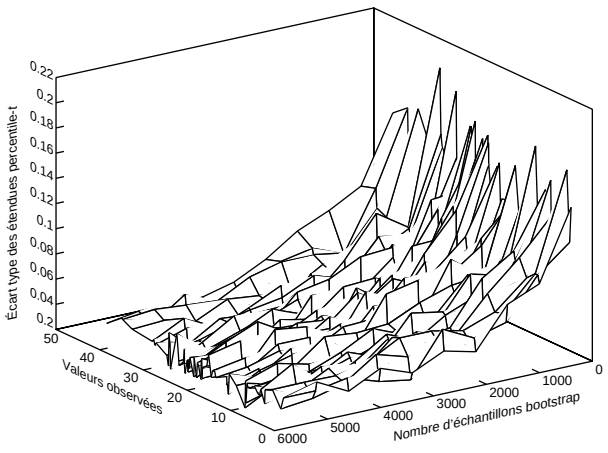

Figure 8

Évolution de l'écart type des étendues de l'intervalle percentile-*t* des prédictions en fonction du nombre de réplications bootstrap.

Standard deviation of the percentile-t prediction intervals ranges versus the number of the bootstrap replications.

CONCLUSION

La régression PLS est une méthode d'analyse des données bien adaptée à la modélisation de la qualité des gazoles. Il s'agit d'effectuer des analyses en composantes principales des variables *X* et *Y*, sous la contrainte que les composantes principales des *X* soient fortement corrélées aux composantes principales des *Y*.

Au cours de ce travail, nous nous sommes limités aux deux intervalles bootstrap percentile et percentile-*t*. Ces deux intervalles ont des propriétés bien connues et définies (Hall, 1992). Les deux approches bootstrap ne suggèrent aucune hypothèse sur la distribution des paramètres du modèle. La méthode percentile-*t* possède d'excellentes propriétés de convergence et est considérée comme la plus pertinente pour les intervalles envisagés. Bien que l'intervalle percentile converge plus lentement, nous avons été amenés à le construire pour sa facilité d'implémentation. Les résultats numériques montrent la cohérence des approches bootstrap.

La faible taille des bases de données d'étalonnage des modèles utilisés pour le contrôle de qualité des produits élaborés en raffinerie et le fait que les termes d'erreur n'ont pas forcement une distribution gaussienne donnent l'avantage aux intervalles bootstrap.

RÉFÉRENCES

Aji, S., Zanier-Szydlowski, N. et Faraj, A. (à paraître) Partial Least Squares Modelling for the Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy, *Oil & Gas Science and Technology - Revue de l'Institut français du pétrole*.

Efron, B. (1979) Bootstrap Methods: Another Look at the Jacknife. *Annals of Statistics*, **7**, 1-26.

Freedman, D.A. (1981) Bootstrapping Regression Models. *Annals of Statistics*, 9, 1218-1228.

Geladi, P. et Kowalski, B.R. (1986) Partial Least-Squares Regression - A Tutorial. *Analytica Chimica Acta*, 185, 1-7.

Hall, P. (1986) On the Number of Bootstrap Simulations Required to Construct a Confidence Interval. *Annals of Statistics*, **14**, 1453-1462.

Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*, Springer Verlag, New York.

Juan, S. et Lantz, F. (2001) La mise en œuvre des techniques de bootstrap pour la prévision économétrique : application à l'industrie automobile. *Oil & Gas Science and Technology - Revue de l'Institut français du pétrole*, **56**, 4, 373-388.

Marteau, P., Obriot J., Zanier-Szydlowski, N., Ruffier-Meray, V. et Behar, E. (1992) Use of Near Infrared Spectroscopy as a Tool for *in-situ* Measurements for the Phase Diagram Determination of Some Solute-Methane Mixtures, Near Infrared Spectroscopy. *Bridging the Gap between Data Analysis and NIR Applications*, K.I. Hildrum *et al.* Eds., 389-394. Ellis Horwood Ltd., Chichester.

Stine, R.A. (1985) Bootstrap Prediction Intervals for Regression. *Journal of the American Statistical Association*, 80, 1026-1031.

Tenenhaus, M. (1998) *La régression PLS : théorie et pratique*, Éditions Technip.

Wold, S. (1978) Cross-Validatory Estimation of the Number of Components Analysis. *Technometrics*, **20**, 4, 397-405.

Wold, S., Albano, C., Dunn, III W.J., Esbensen, K., Hellberge, S., Johansson, E. et Sjostrom, H. (1983) Pattern Recognition: Finding and Using Regularities. In: *Multivariate Data in Proc. IUFOST Conf. "Food Research and Data Analysis", Applied Science Publications*, Martens J. Ed., London.

Manuscrit définitif reçu en juillet 2003

ANNEXE

A.1 RÉGRESSION PLS

La régression PLS décompose la matrice des variables explicatives X en produit bilinéaire plus un résidu :

$$X = t_1 p_1' + x_1 \quad (1)$$

où $t_1 \in \Re^n$ et $p_1 \in \Re^p$ désignent respectivement la première composante PLS fortement corrélée à Y et le vecteur directeur correspondant. Ils peuvent être calculés en utilisant l'algorithme NIPALS *(cf. tableau A.1)*.

TABLEAU A.1

Régression PLS1 selon le principe de l'algorithme NIPALS

PLS1 regression according to NIPALS algorithm (NonLinear Estimation by Iterative Partial Least Square)

Étapes	Résumé des étapes
0	Centrer et réduire X et Y : $x_0 = X$, $y_0 = Y$
1	Pour $h = 1, 2, ..., rang\ (X)$.
2	Calculer w_h par régression de x_{h-1} sur y_{h-1} : $w_h = x'_{h-1} y_{h-1} / y'_{h-1} y_{h-1}$
3	Normer w_h à 1 : $w_h = w_h / \|w_h\|$
4	Calculer les composantes PLS t_h : $t_h = x_{h-1} w_h$
5	Calculer c_h par régression de y_{h-1} sur t_h : $y_{h-1} = y_h + t_h c'_h$ avec : $c_h = (t'_h t_h)^{-1} t'_h y_{h-1}$
6	Répéter les étapes 2 à 5 jusqu'à convergence de w_h
7	Calculer p_h par régression de x_{h-1} sur t_h : $x_{h-1} = x_h + t_h p'_h$ avec : $p_h = x'_{h-1} t_h (t'_h t_h)^{-1}$
8	Calculer les résidus x_h et y_h : $x_h = x_{h-1} - t_h p'_h$, $y_h = y_{h-1} - t_h c_h$
9	$h$ième équation de régression : $Y = t_1 c_1 + ... + t_h c_h + y_h$

On effectue ensuite une régression simple de Y sur t_1 :

$$Y = t_1 c_1 + y_1 \quad (2)$$

où c_1 est le coefficient de régression. Les matrices des résidus sont calculées par :

$$y_1 = Y - t_1 c_1$$

$$x_1 = X - t_1 p_1'$$

Si le pouvoir explicatif de cette régression est faible, on remplace X et Y par x_1 et y_1 respectivement, et l'on répète la même procédure comme pour les premières composantes. Dans le tableau A.1 nous présentons les étapes de l'algorithme PLS1 selon les principes de l'algorithme NIPALS qui permet de déterminer séquentiellement les coefficients de la régression PLS.

Le modèle de régression de Y sur les composantes PLS $(t_1, \ldots, t_h)$ est :

$$Y = t_1 c_1 + t_2 c_2 + ... + t_h c_h + y_h \quad (3)$$

où y_h représente la partie de Y qui n'est pas expliquée par les composantes PLS $(t_1, ..., t_h)$.

Posons : $C_h = (c_1, ..., c_h)'$, $T_h = (t_1, ..., t_h)$, $W_h = (w_1, ..., w_h)$ et $P_h = (p_1, ..., p_h)$.

Les composantes T_h s'expriment en fonction des variables explicatives X par (Tenenhaus, 1998) :

$$T_h = X W_h (P_h' W_h)^{-1}$$

En utilisant les notations précédentes, (3) devient :

$$Y = X W_h (P_h' W_h)^{-1} C_h + y_h \quad (4)$$

Nous en déduisons que l'estimateur de la matrice β_h des coefficients de régression PLS de Y sur X, réalisé en utilisant h composantes et exprimé en fonction de X, s'écrit :

$$\hat{\beta}_h = W_h (P_h' W_h)^{-1} C_h \quad (5)$$

Une des propriétés attractives de la régression PLS est que les composantes $(t_1, ..., t_h)$ sont orthogonales. On a donc :

$$c_h = (t_h' t_h)^{-1} t_h' y_{h-1} = (t_h' t_h)^{-1} t_h' Y$$

et :

$$C_h = (T_h' T_h)^{-1} T_h' Y \quad (6)$$

Ainsi, le vecteur y_h des résidus de la régression MCO de y_{h-1} sur t_h est aussi le vecteur des résidus de la régression MCO de Y sur $(t_1, ..., t_h)$:

$$y_h = Y - (t_1 c_1 + ... + t_h c_h) = y_{h-1} - t_h c_h \quad (7)$$

A.2 TECHNIQUES BOOTSTRAP POUR LES MODÈLES DE REGRESSION PLS

Le bootsrap (Efron, 1979) est une technique de rééchantillonnage basée sur des tirages aléatoires avec remise dans les données. L'utilisation du bootstrap sur les modèles de régression des MCO a initialement été abordée par Freedman (1981). Au cours de ce travail, nous nous intéressons à une technique bootstrap des résidus pour les modèles de régression PLS.

En utilisant les notations précédentes, le modèle de régression PLS à h composantes est noté :

$$Y = X\beta_h + y_h \quad (8)$$

L'estimateur $\hat{\beta}_h$ de la matrice β_h est donné par (5) et celui des résidus y_h est donné par :

$$\hat{y}_h = Y - X\hat{\beta}_h = Y - T_h C_h'$$

Dans cet article, nous nous intéressons à une technique de bootstrap des résidus. Le processus générateur de données bootstrap est le suivant :

$$Y_h^* = X\hat{\beta}_h + y_h^* \tag{9}$$

où y_h^* est un terme aléatoire issu des résidus y_h de la régression PLS initiale.

Y_h^* désigne le vecteur colonne dont les n termes sont la somme de $X\hat{\beta}_h$ et des résidus « bootstrappés » pour h composantes.

Sous les hypothèses standard de la régression MCO de Y sur T_h, on a :

$$\mathrm{Var}(C_h) = \sigma_h^2 (T_h' T_h)^{-1} \tag{10}$$

$$E(y_{h,i})^2 = \sigma_h^2 (1 - u_{h,i}) \tag{11}$$

où σ_h^2 et $u_{h,i}$ désignent respectivement la variance théorique des termes d'erreur et le $i$ième élément diagonal de la matrice de projection orthogonale :

$$\Lambda_h = T_h (T_h' T_h)^{-1} T_h' \tag{12}$$

Le rééchantillonnage à partir des résidus y_h dans la procédure bootstrap sous-estime la variabilité des erreurs (Juan *et al.*, 2001). Le terme aléatoire du modèle théorique bootstrap est construit à partir des résidus transformés suivants :

$$\tilde{y}_{h,i} = \frac{y_{h,i}}{\sqrt{1-u_{h,i}}} - \frac{1}{n}\sum_{s=1}^{n} \frac{y_{h,s}}{\sqrt{1-u_{h,s}}}$$

Les résidus ainsi transformés sont nommés résidus standardisés. Nous divisons $y_{h,i}$ — pour $i = 1, ..., n$ — par un facteur proportionnel à la racine de sa variance. Notons qu'il convient de centrer les résidus $\frac{y_{h,i}}{\sqrt{1-u_{h,i}}}$. Ces derniers n'ont, en effet, aucune raison de l'être par opposition aux résidus $y_{h,i}$ qui le sont par construction. Les résidus $\tilde{y}_{h,i}$ sont de même norme que les termes erreurs $y_{h,i}$. Ils ont tous la même variance et sont recentrés.

Finalement, le modèle bootstrap, sur lequel seront effectuées les estimations, est le suivant :

$$Y_h^* = X\hat{\beta}_h + \tilde{y}_h^* \tag{13}$$

où $\tilde{y}_h^*$ est rééchantillonné à partir de $\tilde{y}_h$.

À chaque itération b ($b = 1, \ldots, B$), un échantillon Y_h^{*b} est constitué à partir des valeurs calculées $\hat{Y}_h$ et des résidus $\tilde{y}_h^{*b}$:

$$Y_h^{*b} = \hat{Y}_h + \tilde{y}_h^{*b} \tag{14}$$

La projection PLS de Y_h^{*b} sur les composantes PLS ($t_1, \ldots, t_h$) permet d'obtenir les estimateurs bootstrap pour le $b$ième échantillon. Notons qu'il faut centrer et réduire Y_h^{*b} par la moyenne et la variance de Y avant de faire la projection. Les étapes de la projection PLS s'écrivent sous la forme :

$$\begin{cases} Y_h^{*b} = t_1 c_1^{*b} + y_1^{*b} \\ y_1^{*b} = t_2 c_2^{*b} + y_2^{*b} \\ \ldots, \\ y_{h-1}^{*b} = t_h c_h^{*b} + y_h^{*b} \end{cases}$$

Les coefficients c_h^{*b} sont fournis par :

$$c_h^{*b} = (t_h' t_h)^{-1} t_h' y_{h-1}^{*b} \tag{15}$$

Cependant, l'orthogonalité des composantes t_h permet aisément d'obtenir l'estimation de $C_h^{*b} = (c_1^{*b}, ..., c_h^{*b})'$ par :

$$C_h^{*b} = (T_h' T_h)^{-1} T_h' Y_h^{*b} \tag{16}$$

L'estimateur bootstrap $\hat{\beta}_h^{*b}$ de la matrice des coefficients est ensuite fourni par :

$$\hat{\beta}_h^{*b} = W_h (P_h' W_h)^{-1} C_h^{*b} \tag{17}$$

Les résidus estimés bootstrap pour le modèle de régression PLS à h composantes s'obtiennent par :

$$\begin{aligned} y_h^{*b} &= y_{h-1}^{*b} - t_h c_h^{*b} \\ &= Y_h^{*b} - T_h C_h^{*b} \\ &= Y_h^{*b} - X\hat{\beta}_h^{*b} \\ &= Y_h^{*b} - \hat{Y}_h^{*b} \end{aligned}$$

A.2.1 Intervalle de confiance des coefficients de la régression PLS

A.2.1.1 Intervalle de confiance standard

Soit $z_{h,j}$ la variable aléatoire définie par :

$$z_{h,j} = \frac{\hat{\beta}_{h,j} - \beta_{h,j}}{s(\hat{\beta}_{h,j})}$$

où $s(\hat{\beta}_{h,j})$ désigne l'écart type estimé du coefficient de régression. D'après (5) et (10) $s^2(\hat{\beta}_{h,j})$ est donné par le $j$ième élément diagonal de la matrice de variance covariance de $\hat{\beta}_h$ donnée par :

$$\mathrm{Var}(\hat{\beta}_h) = \hat{\sigma}_h^2 \Gamma_h (T_h' T_h)^{-1} \Gamma'_h \tag{18}$$

où : $\Gamma_h = W_h (P_h' W_h)^{-1}$ et $\hat{\sigma}_h^2 = \dfrac{1}{n-h} \left\| Y - \hat{Y}_h \right\|^2$

Un intervalle de confiance standard de $\beta_{h,j}$ découle de l'hypothèse selon laquelle $z_{h,j}$ est distribuée selon une loi de Student à (n-h) degrés de liberté. Ainsi, pour un niveau de

confiance $(1-2\alpha)$, cet intervalle de confiance prend la forme suivante :

$$[\hat{\beta}_{h,j} - s(\hat{\beta}_{\alpha,j})t_{1-\alpha,n-h}, \hat{\beta}_{h,j} - s(\hat{\beta}_{h,j})t_{\alpha,n-h}] \qquad (19)$$

où $t_{\alpha,n-h}$ et $t_{1-\alpha,n-h}$ désignent respectivement les quantiles α et $(1-\alpha)$ de la distribution de Student à $(n-h)$ degrés de liberté.

A.2.1.2 Intervalle de confiance percentile

Les intervalles de confiance bootstrap sont construits à partir des deux approches percentile et percentile-*t*. La première méthode, basée uniquement sur les estimations bootstrap, est la méthode la plus simple pour obtenir les intervalles de confiance. Pour un niveau $(1-2\alpha)$, l'intervalle de confiance percentile pour le coefficient $\beta_{h,j}$ est donné par :

$$[\hat{\beta}^{*\lfloor \alpha B \rfloor}_{h,j},\ \hat{\beta}^{*\lceil (1-\alpha)B \rceil}_{h,j}] \qquad (20)$$

où $\hat{\beta}^{*\lfloor \alpha B \rfloor}_{h,j}$ et $\hat{\beta}^{*\lceil (1-\alpha)B \rceil}_{h,j}$ désignent respectivement la $\lfloor \alpha B \rfloor$ième [2] et la $\lceil (1-\alpha)B \rceil$ième valeur de la liste ordonnée des B réplications bootstrap. Les valeurs seuil sont donc choisies telles que α % des réplications ont fourni des $\hat{\beta}^{*}_{h,j}$ plus petits (resp. grands) que la borne inférieure (resp. supérieure) de l'intervalle de confiance percentile.

A.2.1.3 Intervalle de confiance percentile-*t*

La procédure bootstrap percentile-*t* consiste à estimer la fonction de répartition de $z_{h,j}$ directement à partir des données. Cela revient à construire une table statistique à partir de la fonction de répartition empirique des B réplications bootstrap $z^{*b}_{h,j}$ définies comme :

$$z^{*b}_{h,j} = \frac{\hat{\beta}^{*b}_{h,j} - \hat{\beta}_{h,j}}{s^{*}(\hat{\beta}^{*b}_{h,j})}$$

où $s(\hat{\beta}^{*b}_{h,j})$ désigne l'écart type estimé bootstrap du coefficient de régression. D'après (10, 16 et 17), $s^2(\hat{\beta}^{*b}_{h,j})$ est donné par le $j$ième élément diagonal de la matrice de variance covariance de $\hat{\beta}^{*b}_{h}$ donnée par :

$$\mathrm{Var}(\hat{\beta}^{*b}_{h}) = \hat{\sigma}^{2*b}_{h}\Gamma_h(T_h'T_h)^{-1}\Gamma'_h \qquad (21)$$

où :
$$\hat{\sigma}^{2*b}_{h} = \frac{1}{n-h}\left\|Y^{*b}_{h} - \hat{Y}^{*b}_{h}\right\|^2$$

et :
$$\Gamma_h = W_h(P_h'W_h)^{-1}$$

Soit $\hat{F}(z^{*}_{h,j})$ la fonction de répartition empirique des $z^{*b}_{h,j}$. Le fractile à α %, $\hat{F}^{-1}(z^{*}_{h,j},\alpha)$ est estimé par la valeur $\hat{t}(\alpha)$ telle que :

$$\#\left\{z^{*b}_{h,j} \le \hat{t}(\alpha)\right\} / B = \alpha$$

2 $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) désigne la partie entière inférieure (resp. supérieure) de x.

Alors, l'intervalle de confiance percentile-*t* pour $\beta_{h,j}$ s'écrit :

$$[\hat{\beta}_{h,j} - s(\hat{\beta}_{h,j})\hat{t}(1-\alpha), \hat{\beta}_{h,j} - s(\hat{\beta}_{h,j})\hat{t}(\alpha)] \qquad (22)$$

Remarque

L'intervalle de confiance percentile-*t* substitue aux valeurs critiques de la loi de Student utilisées dans l'intervalle standard les valeurs seuil de la table bootstrap. Notons que ces dernières peuvent être différentes. Cette différence est d'autant plus importante que la distribution inconnue des erreurs est différente de la loi normale. De plus, nous remarquons que les valeurs des quantiles α et $(1-\alpha)$ de la distribution de Student, symétriques par nature, entraînent directement la symétrie de l'intervalle de confiance standard autour de l'estimation $\hat{\beta}_{h,j}$. Par opposition, les valeurs $\hat{t}(\alpha)$ et $\hat{t}(1-\alpha)$ de la table bootstrap peuvent être asymétriques et permettent alors des intervalles de confiance asymétriques autour de $\hat{\beta}_{h,j}$. Cette prise en compte d'une possible asymétrie constitue un avantage important des intervalles de confiance bootstrap.

A.2.2 Intervalle de confiance des prédictions

A.2.2.1 Intervalle de prédiction standard

La prédiction $\hat{Y}$ de la variable réponse est calculée à partir du modèle de régression :

$$\hat{Y}_h = X\ \hat{\beta}_h \qquad (23)$$

L'intervalle de prédiction standard découle, comme les intervalles de confiance des coefficients de la régression, de l'hypothèse de normalité des erreurs. Ainsi, pour un niveau de confiance $(1-2\alpha)$ l'intervalle de prédiction standard de Y_i s'écrit :

$$[\hat{Y}_{h,i} - s(y_{h,i})t_{1-\alpha,n-h}, \hat{Y}_{h,i} - s(y_{h,i})t_{\alpha,n-h}] \qquad (24)$$

où :
$$s^2(y_{h,i}) = \hat{\sigma}^2_h(1 - T_{h,i}(T_h'T_h)^{-1}T'_{h,i}) \qquad (25)$$

est l'écart type estimé de l'erreur de prédiction et :

$$\hat{\sigma}^2_h = \frac{1}{n-h}\left\|Y - \hat{Y}_h\right\|^2$$

Soit X_f la matrice de nouvelles observations de variables explicatives, la prédiction des variables réponses $\hat{Y}_{h,f}$ est calculée à partir du modèle de régression :

$$\hat{Y}_{h,f} = X_f\hat{\beta}_h \qquad (26)$$

L'intervalle standard s'obtient comme précédemment :

$$[\hat{Y}_{h,f} - s(y_{h,f})t_{1-\alpha,n-h}, \hat{Y}_{h,f} - s(y_{h,f})t_{\alpha,n-h}] \qquad (27)$$

où :
$$s^2(y_{h,f}) = \hat{\sigma}^2_h(1 + T_{h,f}(T_h'T_h)^{-1}T'_{h,f}) \qquad (28)$$

$T_{h,f}$ représente les composantes PLS de X_f données par :

$$T_{h,f} = X_f W_h(P_h'W_h)^{-1} \qquad (29)$$

A.2.2.2 Intervalle de prédiction percentile

La méthode percentile consiste à utiliser l'approximation bootstrap de la distribution de l'erreur de prédiction $y_h = Y - \hat{Y}_h$ pour construire un intervalle de prédiction de Y.

Les réplications bootstrap des valeurs Y_h^{*b} sont générées suivant le modèle :

$$Y_h^{*b} = X\hat{\beta}_h + \tilde{y}_h^{*b} \tag{30}$$

Pour chacune des B réplications bootstrap, nous calculons l'estimateur bootstrap $\hat{\beta}_h^{*b}$ donné par (17). Ainsi, la prévision et l'erreur de prédiction bootstrap pour la $i$ième observation s'écrivent respectivement :

$$\hat{Y}_{h,i}^{*b} = X_i\hat{\beta}_h^{*b}$$
$$y_{h,i}^{*b} = Y_{h,i}^{*b} - \hat{Y}_{h,i}^{*b} = \hat{Y}_{h,i} + \tilde{y}_{h,i}^{*b} - \hat{Y}_{h,i}^{*b}$$

Les B réplications bootstrap de l'erreur de prédiction fournissent la distribution empirique G_i^* de $y_{h,i}^*$ Les quantiles de cette distribution empirique, notés $G_i^{*-1}(\alpha)$ et $G_i^{*-1}(1-\alpha)$, sont alors utilisés pour construire un intervalle de prédiction bootstrap.

L'intervalle de prédiction percentile de Y_i est alors de la forme :

$$[\hat{Y}_{h,i} + G_i^{*-1}(\alpha),\ \hat{Y}_{h,i} + G_i^{*-1}(1-\alpha)] \tag{31}$$

Pour de nouvelles observations de variables explicatives X_f, les réplications bootstrap de la valeur $Y_{h,f}^*$ sont générées suivant le même modèle que précédemment :

$$Y_{h,f}^* = \hat{Y}_{h,f} + \tilde{y}_h^* = X_f\hat{\beta}_h + \tilde{y}_h^*$$

Pour chacune des B réplications bootstrap, nous calculons l'estimateur bootstrap $\hat{\beta}_h^{*b}$ donné par (17). Ainsi, la prévision et l'erreur de prédiction bootstrap s'écrivent respectivement :

$$\hat{Y}_{h,f}^{*b} = X_f\hat{\beta}_h^{*b}$$
$$y_{h,f}^{*b} = Y_{h,f}^{*b} - \hat{Y}_{h,f}^{*b}$$

Les B réplications bootstrap de l'erreur de prédiction fournissent la distribution empirique G_f^* de $Y_{h,f}^*$.

L'intervalle de prédiction percentile est alors de la forme :

$$[\hat{Y}_{h,f} + G_f^{*-1}(\alpha), \hat{Y}_{h,f} + G_f^{*-1}(1-\alpha)] \tag{32}$$

A.2.2.3 Intervalle de prédiction percentile-*t*

De manière identique à l'intervalle de confiance des coefficients, la construction de l'intervalle de prédiction avec la méthode percentile-*t* implique le calcul pour chaque échantillon bootstrap de l'écart type. Ainsi, pour établir des intervalles de prédiction percentile-*t*, l'estimateur bootstrap de l'écart type de prédiction est nécessaire pour chacune des réplications.

La procédure percentile-*t* consiste à construire, pour chacune des réplications, les statistiques $z_{h,i}^{*b}$ telles que :

$$z_{h,i}^{*b} = \frac{y_{h,i}^{*b}}{s(y_{h,i}^{*b})} = \frac{\hat{Y}_{h,i} + \tilde{y}_{h,i}^{*b} - \hat{Y}_{h,i}^{*b}}{s(y_{h,i}^{*b})}$$

où : $$s^2(y_{h,i}^{*b}) = \hat{\sigma}_h^{2*b}(1 - T_{h,i}(T_h'T_h)^{-1}T_{h,i}') \tag{33}$$

et : $$\hat{\sigma}_h^{2*b} = \frac{1}{n-h}\left\|Y_h^{*b} - \hat{Y}_h^{*b}\right\|^2 \tag{34}$$

La distribution bootstrap de $z_{h,i}^*$ définit l'intervalle de prédiction bootstrap percentile-*t*. Les quantiles $z_{h,i}^{*\lceil(1-\alpha)B\rceil}$ et $z_{h,i}^{*\lfloor\alpha B\rfloor}$ remplacent ainsi les valeurs critiques de la distribution de Student prises en compte dans l'intervalle de prédiction standard donné par (24).

L'intervalle de prédiction percentile-*t* s'écrit donc :

$$[\hat{Y}_h - s(y_{h,i})z_{h,i}^{*\lceil(1-\alpha)B\rceil},\ \hat{Y}_h - s(y_{h,i})z_{h,i}^{*\lfloor\alpha B\rfloor}] \tag{35}$$

où $s(y_{h,i})$ est donnée à partir de (25).

Pour de nouvelles observations de variables explicatives X_f, la procédure percentile-*t* consiste à construire pour chacune des réplications les statistiques $z_{h,f}^{*b}$ telles que :

$$z_{h,f}^{*b} = \frac{y_{h,f}^{*b}}{s(y_{h,f}^{*b})}$$

où : $$s^2(y_{h,f}^{*b}) = \hat{\sigma}_h^{2*b}(1 + T_{h,f}(T_h'T_h)^{-1}T_{h,f}')$$

et $T_{h,f}$ est donnée par (29).

L'intervalle de prédiction percentile-*t* s'écrit donc :

$$[\hat{Y}_f - s(y_{h,f})z_{h,f}^{*\lceil(1-\alpha)B\rceil},\ \hat{Y}_f - s(y_{h,f})z_{h,f}^{*\lfloor\alpha B\rfloor}] \tag{36}$$

où : $$s^2(y_{h,f}) = \hat{\sigma}_h^2(1 - T_{h,f}(T_h'T_h)^{-1}T_{h,f}')$$

Notons que, comme pour l'intervalle de confiance des coefficients, le quantile $(1-\alpha)$ de la distribution de $z_{h,f}^*$ définit la borne inférieure de l'intervalle de prédiction et inversement pour le quantile α.

Une distribution symétrique de $z_{h,f}^*$ implique donc la symétrie de l'intervalle de prédiction percentile-*t*. Cependant, dans le cas contraire, l'asymétrie est retranscrite de manière inversée pour ce dernier. Par exemple, si $z_{h,f}^*$ possède une queue de distribution plus longue vers la droite, les fractiles $z_{h,f}^{*\lceil(1-\alpha)B\rceil}$ et $z_{h,f}^{*\lfloor\alpha B\rfloor}$ sont décalés vers les valeurs élevées des erreurs de prédiction bootstrap, comparativement aux quantiles correspondants d'une distribution symétrique. L'intervalle de prédiction percentile-*t* résultant est donc décalé vers la gauche, asymétrique autour de la valeur prédite. Ainsi, sa construction implique une sorte de correction automatique du biais et permet l'acceptation, pour un niveau de confiance donné, de valeurs prédites plus faibles que l'intervalle de prédiction standard, symétrique.