Oil & Gas Science and Technology – *Rev. IFP*, Vol. 56 (2001), No. 4, pp. 373-388

La mise en œuvre des techniques de bootstrap pour la prévision économétrique : application à l'industrie automobile

S. Juan[1] et F. Lantz[2]

1 Renault, 1, avenue du Golf, 78288 Guyancourt Cedex - France
2 École du pétrole et des moteurs, 228-232, avenue Napoléon Bonaparte, 92852 Rueil-Malmaison Cedex - France
e-mail : sandrine.juan@renault.com - frederic.lantz@ifp.fr

Résumé — L'application des méthodes de bootstrap aux modèles de régression permet d'obtenir des approximations de la distribution des coefficients ainsi que la distribution des erreurs de prédiction. Dans cet article, nous nous intéressons à l'application des techniques de bootstrap pour déterminer des intervalles de prédiction à partir d'une modélisation économétrique où les régresseurs sont des données. Nous abordons différents problèmes liés à cette application : la détermination du nombre de réplications, le choix de la méthode de calcul de l'estimateur des moindres carrés ordinaires (pseudo-inverse ou inverse), ainsi que l'algorithme de tri de la statistique considérée. Ces investigations proviennent des besoins de prédiction des coûts dans l'industrie automobile dès la phase d'avant-projet du développement d'un nouveau véhicule. Généralement, les échantillons sont de faible taille et les termes d'erreur n'ont pas forcément une distribution gaussienne. Ainsi, l'utilisation des techniques de bootstrap permet d'améliorer les intervalles de prédiction en retranscrivant la distribution originale des données. Deux exemples (moteur et réservoir) illustrent la mise en œuvre de ces techniques.
Mots-clés : bootstrap, pseudo-inverse, algorithme de tri, prévision économétrique, industrie automobile.

Abstract* — *Application of Bootstrap Techniques in Econometrics: the Example of Cost Estimation in the Automotive Industry *— Bootstrap methods applied in regression models help to approximate the distributions of the coefficients and the prediction errors. In this paper, we apply bootstrap techniques to determine prediction intervals from econometric models when the regressors are known. We investigate problems associated with their application: determining the number of replications, choosing the method to calculate the least-squares estimator (pseudo-inverse or inverse) and sorting algorithm of the statistic of interest. This investigation arises from the need in the automotive industry to predict costs in the early phases of development of a new vehicle. Generally, the sample size is small and the model's error term of the model is not Gaussian. Consequently, bootstrap techniques strongly improve prediction intervals by reflecting the original distribution of the data. Two examples (engine and fuel tank) illustrate the technique.*
Keywords: bootstrap, pseudo-inverse, sorting methods, econometric forecast, car industry.

INTRODUCTION

L'économétrie trouve une grande partie de ses applications industrielles dans l'obtention de prévisions. La détermination des intervalles de confiance des coefficients ainsi que des intervalles de prédiction dépend des hypothèses inhérentes aux méthodes d'estimation et, en particulier, des hypothèses sur la distribution du terme d'erreur du modèle de régression. Lorsque celles-ci ne sont plus vérifiées, les intervalles de prédiction standard ne peuvent plus être utilisés.

Le « bootstrap » proposé par Efron (1979) fournit une approximation d'une distribution inconnue par une distribution empirique obtenue par un processus de ré-échantillonnage. L'application des méthodes de bootstrap aux modèles de régression donne ainsi une approximation de la distribution des coefficients (Freedman, 1981) et la distribution des erreurs de prédiction lorsque les régresseurs sont des données (Stine, 1985) ou des variables aléatoires (McCullough, 1996).

Notre propos concerne la mise en œuvre du bootstrap pour établir des intervalles de prédiction sur des modèles économétriques lorsque les régresseurs sont connus. Ces investigations ont été suscitées par les besoins de prédiction des coûts dans l'industrie automobile dès les premières phases de développement d'un nouveau véhicule. En effet, la détermination anticipée des coûts fait partie de la stratégie d'offre des constructeurs automobiles. Elle permet, de manière générale, d'aider à la conception d'une nouvelle voiture autour d'un prix cible et, en particulier, de réaliser des comparaisons entre plusieurs solutions techniques.

Parmi les différentes approches envisageables pour effectuer des prévisions de coûts, l'utilisation de la modélisation économétrique est particulièrement adaptée dans les premières étapes d'un projet automobile car elle ne nécessite pas d'information détaillée. Cependant, elle soulève des difficultés liées à la faible taille des échantillons de données et à la distribution inconnue des termes d'erreur des modèles de régression. Dans ce contexte, le bootstrap autorise l'utilisation d'une approche économétrique à des fins prédictives.

Dans la première section, nous rappelons brièvement le principe du bootstrap sur les modèles de régression. Nous abordons ensuite, dans la section 2, plusieurs problèmes liés à sa mise en œuvre : détermination du nombre de réplications, choix de la méthode de calcul de l'estimateur des moindres carrés (pseudo-inverse ou inverse) et algorithme de tri de la statistique d'intérêt. La section suivante est consacrée à l'estimation des coûts au stade des avant-projets dans l'industrie automobile et à plusieurs applications du bootstrap *(section 3)*. Ainsi, après l'estimation du coût d'un moteur, qui ne dépend que d'une variable continue, nous présentons un modèle économétrique du coût d'un réservoir à carburant où intervient une variable binaire. Le résumé des résultats obtenus et quelques voies d'investigation forment la conclusion.

1 LES TECHNIQUES DE BOOTSTRAP SUR LES MODÈLES DE RÉGRESSION

Le bootstrap est une technique de ré-échantillonnage basée sur des tirages aléatoires avec remise dans les données constituant un échantillon. Utilisées pour approcher la distribution inconnue d'une statistique par sa distribution empirique, les méthodes de bootstrap sont mises en œuvre afin d'améliorer la précision des estimations statistiques. Des présentations détaillées de cette approche sont proposées, notamment par Hall (1992) ainsi que Efron et Tibshyrani (1993).

L'utilisation du bootstrap sur les modèles de régression a initialement été abordée par Freedman (1981). Jeong et Maddala (1993), Vinod (1993) et Veall (1998) offrent des synthèses des nombreux développements et applications des techniques de bootstrap dans le domaine de l'économétrie qui sont ensuite apparus. Horowitz (1997) s'intéresse aux performances théoriques et numériques du bootstrap en économétrie. Nous rappelons brièvement le principe de cette méthode de ré-échantillonnage ainsi que son application aux modèles de régression en annexe.

1.1 Le bootstrap des résidus et les intervalles de confiance bootstrap

Le modèle de régression linéaire multiple est noté :

$$Y = X\beta + u \tag{1}$$

où Y est un vecteur $(n, 1)$, X une matrice (n, p), β le vecteur des coefficients à estimer $(p, 1)$ et u le vecteur des erreurs aléatoires $(n, 1)$. Un rang d'observations i ($i = 1, \ldots, n$) de la matrice X, correspondant à une ligne, est noté X_i $(1, p)$.

L'estimateur des paramètres β obtenu par la méthode des moindres carrés ordinaires (MCO) s'exprime comme :

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2}$$

et les résidus comme $\hat{u} = Y - X\hat{\beta}$.

Pour la suite de nos développements, nous avons retenu une approche en termes de bootstrap des résidus plutôt qu'une approche en termes de bootstrap par paires car nous ne sommes pas confrontés à un problème d'hétéroscédasticité[1] (Flachaire, 1998).

Le modèle théorique bootstrap est le suivant :

$$Y^* = X\hat{\beta} + u^* \tag{3}$$

où u^* est un terme aléatoire issu des résidus $\hat{u}$ de la régression initiale. À chaque itération b ($b = 1, \ldots, B$), un échantillon $\{y_i^*\}_{i=1}^n$, de dimension $(n, 1)$, est constitué à partir du modèle bootstrap (3).

(1) L'*hétéroscédasticité* désigne l'absence d'égalité des variances des erreurs pour tous les rangs d'observation. Elle s'oppose à l'*homoscédasticité*. L'hypothèse d'homoscédasticité est nécessaire pour que l'estimateur MCO des coefficients soit efficace.

Les résidus MCO étant plus petits que les erreurs qu'ils estiment, le terme aléatoire du modèle théorique bootstrap est construit à partir des résidus transformés suivants, qui sont de même norme que les termes erreurs u_i :

$$\tilde{u}_i = \frac{\hat{u}_i}{\sqrt{(1-h_i)}} - \frac{1}{n}\sum_{s=1}^{n}\frac{\hat{u}_s}{\sqrt{(1-h_s)}}$$

où h_i est l'élément diagonal (i, i) de la matrice $X\,(X^T X)^{-1}\,X^T$. En effet, les erreurs u et les résidus $\hat{u}$ sont liés par la relation :

$$\hat{u} = (1 - X\,(X^T X)^{-1}\,X^T)\,u$$

Le modèle théorique bootstrap s'exprime donc comme :

$$y_i^*(b) = X_i\hat{\beta} + \tilde{u}_i^*(b)\,, \quad i = 1,...,n \tag{4}$$

où $\tilde{u}_i^*(b)$ est ré-échantillonné à partir des $\tilde{u}_i$.

Soit la variable aléatoire z_j définie comme :

$$z_j = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)}$$

où $s(\hat{\beta}_j)$ désigne l'écart type estimé du coefficient. L'intervalle de confiance standard de β_j découle de l'hypothèse selon laquelle z_j est distribuée selon une loi de Student à n-p degrés de liberté. Ainsi, pour un niveau de confiance $1 - 2\alpha$, cet intervalle de confiance prend la forme suivante :

$$\left[\hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(1-\alpha),n-p}\,,\ \hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(\alpha),n-p}\right] \tag{5}$$

où t est la valeur des quantiles α et $1 - \alpha$ de la distribution de Student à n-p degrés de liberté.

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 d'obtention d'intervalles de confiance. Pour un niveau $1 - 2\alpha$, l'intervalle de confiance percentile pour le paramètre β_j est donné par :

$$\left[\hat{\beta}_j^*(\alpha B)\,,\ \hat{\beta}_j^*((1-\alpha)B)\right] \tag{6}$$

où $\hat{\beta}_j^*(\alpha B)$ représente la αB-ième valeur (respectivement $\hat{\beta}_j^*((1-\alpha)B)$ la $(1-\alpha B)$-ième valeur) de la liste ordonnée des B réplications bootstrap. Les valeurs seuils sont donc choisies telles que α % des réplications ont fourni des $\hat{\beta}_j^*$ plus petits (grands) que la borne inférieure (supérieure) de l'intervalle de confiance percentile.

La procédure bootstrap percentile-t consiste à estimer la fonction de répartition de z_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_j^*. Cette table est nommée *table bootstrap*. Les z_j^* sont définies comme :

$$z_j^* = \frac{\hat{\beta}_j^* - \hat{\beta}_j}{s^*(\hat{\beta}_j^*)} \tag{7}$$

Notons, en comparaison avec la méthode percentile, un calcul supplémentaire dans cette approche. En effet, pour chacune des réplications bootstrap, il est nécessaire de calculer l'écart type estimé bootstrap $s^*(\hat{\beta}_j^*)$.

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

$$\#\left\{z_j^*(b) \le \hat{t}^{(\alpha)}\right\} / B = \alpha$$

Finalement, l'intervalle de confiance percentile-t pour β_j s'écrit :

$$\left[\hat{\beta}_j - s(\hat{\beta}_j)\cdot\hat{t}^{(1-\alpha)}\,,\ \hat{\beta}_j - s(\hat{\beta}_j)\cdot\hat{t}^{(\alpha)}\right] \tag{8}$$

Ainsi, l'intervalle de confiance percentile-t est l'analogue bootstrap de l'intervalle de confiance standard.

En résumé, l'intervalle de confiance percentile-t substitue, aux valeurs critiques de la loi de Student utilisées dans l'intervalle standard, les valeurs seuils de la table bootstrap. Notons que ces dernières peuvent être très différentes. Cette différence est d'autant plus importante que la distribution (inconnue) des erreurs est éloignée 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 $\hat{\beta}_j$. 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 $\hat{\beta}_j$. Cette prise en compte d'une possible asymétrie constitue un avantage important des intervalles de confiance bootstrap.

1.2 Les intervalles de prédiction bootstrap

À la suite de Stine (1985) et Breiman (1992), notre cadre de travail nous conduit à utiliser le bootstrap pour la construction des intervalles de prédiction sur des modèles de régression avec des régresseurs fixes, dont les valeurs sont connues (prévision non conditionnelle). Notons cependant que la construction des intervalles de prédiction bootstrap sur des modèles avec des régresseurs stochastiques est proposée par McCullough (1996).

Pour un nouveau rang *f* d'observation des variables explicatives X_f, la prédiction de coût $\hat{y}_f$ est calculée à partir du modèle de régression :

$$\hat{y}_f = X_f\hat{\beta}$$

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 $1 - 2\alpha$, cet intervalle de prédiction standard s'écrit :

$$\left[\hat{y}_f - s_f \cdot t_{(1-\alpha),n-p}\ ,\ \hat{y}_f - s_f \cdot t_{(\alpha),n-p}\right] \qquad (9)$$

où, en notant *s* l'écart type résiduel, s_f est l'écart type estimé de l'erreur de prédiction, défini comme :

$$s_f = s\sqrt{1 + X_f(X^TX)^{-1}X_f^T}$$

L'utilisation du bootstrap, pour préciser les intervalles de prédiction, conduit à étudier la distribution de l'erreur de prédiction. Aussi, afin de conserver le même processus générateur de données (PGD) pour les estimations des coefficients et des prédictions, les intervalles de prédiction bootstrap sont obtenus avec la procédure du bootstrap des résidus. De manière similaire à la construction des intervalles de confiance, il existe deux principales méthodes de construction des intervalles de prédiction bootstrap : l'approche percentile et l'approche percentile-*t*.

1.2.1 L'intervalle de prédiction percentile

La méthode percentile consiste à utiliser l'approximation bootstrap de la distribution de l'erreur de prédiction : $e_f = \hat{y}_f - y_f$, pour construire un intervalle de prédiction de y_f.

Les réplications bootstrap de la future valeur y_f^*, pour le nouveau rang d'observations X_f, sont générées suivant le même modèle (4) :

$$y_f^* = X_f\hat{\beta} + \tilde{u}_f^* \qquad (10)$$

Le terme d'erreur $\tilde{u}_f^*$ est issu, comme les $\tilde{u}^*$, d'un tirage avec remise dans la distribution empirique des résidus transformés.

Pour chacune des *B* réplications bootstrap, nous calculons l'estimateur bootstrap. Ainsi, la prévision et l'erreur de prédiction bootstrap s'écrivent respectivement :

$$\hat{y}_f^*(b) = X_f\hat{\beta}^*(b) \qquad (11)$$

$$e_f^*(b) = \hat{y}_f^*(b) - y_f^*(b)$$

En utilisant l'équation (9), nous pouvons réécrire l'erreur de prédiction bootstrap comme :

$$e_f^* = \hat{y}_f^* - \hat{y}_f - \tilde{u}_f^* \qquad (12)$$

Cette dernière dépend donc, par nature, de la prédiction MCO initiale $\hat{y}_f$.

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

Un intervalle de prédiction percentile est finalement de la forme suivante :

$$\left[\hat{y}_f - G^{*-1}(1-\alpha); \hat{y}_f - G^{*-1}(\alpha)\right] \qquad (13)$$

1.2.2 L'intervalle de prédiction percentile-*t*

De manière identique à l'intervalle de confiance, la construction de l'intervalle de prédiction avec la méthode percentile-*t* implique le calcul, pour chaque échantillon bootstrap, de l'estimateur 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. Il s'écrit :

$$s_f^* = s^* \cdot \sqrt{(1 + h_f)} \qquad (14)$$

où s^* est l'estimateur bootstrap de l'écart type des termes erreurs et où $h_f = X_f(X^TX)^{-1}X_f^T$.

La procédure percentile-*t* consiste à construire les statistiques z_f^*, telles que :

$$z_f^* = \frac{e_f^*}{s_f^*} = \frac{\hat{y}_f^* - \hat{y}_f - \tilde{u}_f^*}{s_f^*} \qquad (15)$$

La distribution bootstrap de z_f^* définit l'intervalle de prédiction bootstrap percentile-*t*. Les quantiles $z_{f(1-\alpha)}^*$ et $z_{f(\alpha)}^*$ remplacent ainsi les valeurs critiques de la distribution de Student, prises en compte dans l'intervalle de prédiction standard *(éq. (9))*.

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

$$\left[\hat{y}_f - s_f \cdot z_{f(1-\alpha)}^*; \hat{y}_f - s_f \cdot z_{f(\alpha)}^*\right] \qquad (16)$$

Notons que, comme pour l'intervalle de confiance des coefficients, le quantile $1 - \alpha$ de la distribution de z_f^* définit la borne inférieure de l'intervalle de prédiction et inversement pour le quantile α.

Une distribution symétrique de z_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_f^* possède une queue de distribution plus longue vers la droite, les quantiles $z^*_{f(1-\alpha)}$ et $z^*_{f(\alpha)}$ 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 MCO. 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.

2 LA MISE EN ŒUVRE DES MÉTHODES

Les développements que nous proposons pour la mise en œuvre du bootstrap s'inscrivent dans le prolongement de Efron et Tibshyrani (1993), Booth et Sarkar (1998) pour la détermination du nombre de réplications bootstrap. À la suite de McCullough et Vinod (1996), nous nous intéressons au choix de la méthode de calcul de l'estimateur MCO par pseudo-inverse ainsi qu'à l'algorithme de tri de la statistique d'intérêt.

2.1 Le nombre de réplications bootstrap *B*

Efron et Tibshyrani (1993) préconisent d'effectuer un nombre de réplications bootstrap de l'ordre de 25 pour le calcul de l'écart type d'un estimateur et de l'ordre du millier pour les intervalles de confiance bootstrap.

La construction de l'intervalle de confiance (IC)[2] bootstrap implique les fractiles 2,5 % et 97,5 % de la distribution empirique bootstrap de la statistique z^*. 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 avant d'obtenir des fractiles qui varient peu.

Intuitivement, il paraît logique que l'estimation des fractiles d'une distribution nécessite un nombre plus élevé d'échantillons bootstrap que le calcul de l'écart type de l'estimateur, par exemple. En effet, elle dépend de la queue de distribution, où peu d'échantillons apparaissent. La question revient alors à déterminer le nombre de réplications bootstrap à partir duquel la valeur du fractile peut être considérée comme stable.

Le processus mis en œuvre consiste, pour un certain nombre de valeurs de *B*, à effectuer un ensemble de simulations (noté *k*), visant à juger de la stabilité des résultats, lorsque les racines (valeurs de départ) du générateur de nombres pseudo-aléatoires sont différentes. $k = 100$ simulations ont été réalisées, ce qui semble raisonnable compte tenu de nos investigations. Par ailleurs, les simulations ont été effectuées pour les nombres de réplications bootstrap suivants : $B = 20, 30, 100, 500, 1000, 5000$ et 10 000.

Plutôt que d'étudier la variabilité de chacun des fractiles 2,5 % et 97,5 % de la distribution séparément, nous considérons l'étendue de l'intervalle entre ces derniers, que nous nommons *intervalle de confiance de* z^*. Ainsi, nos travaux s'appuient sur l'analyse de la variabilité des étendues de ces IC, sur 100 simulations, en fonction de *B*.

En résumé, pour chacune des *k* simulations, $k = 1, \ldots, 100$, les *B* réplications bootstrap fournissent la distribution empirique de z^*, à partir de laquelle sont extraits les fractiles d'intérêt. Pour un nombre de réplications *B* donné, les simulations permettent donc d'obtenir 100 IC de z^* et leurs étendues. Leur variabilité est ensuite étudiée.

Les comparaisons des distributions des étendues sont effectuées deux à deux, pour des valeurs croissantes de *B*. Pour ce faire, les caractéristiques de valeurs centrales (moyenne, médiane) et de dispersion sont calculées et trois critères sont examinés : l'égalité des médianes[3], l'égalité des variances et le coefficient de variation (CV).

Les deux premiers critères font l'objet de tests statistiques, respectivement le test non paramétrique de Wilcoxon d'égalité des médianes et le test de Fisher d'égalité des variances de deux échantillons indépendants. L'évolution du CV en fonction du nombre de réplications fait l'objet de l'analyse du dernier critère. L'application et l'interprétation de cette procédure sont présentées sur l'exemple du modèle d'estimation de coût des moteurs.

2.2 Le calcul de l'estimateur des paramètres de la régression

Le calcul de l'estimateur MCO est effectué, classiquement, à partir de la formule (2) en inversant la matrice $X^T X$. L'inversion de cette matrice soulève des problèmes d'instabilité numérique lorsque la matrice *X* des variables explicatives est mal conditionnée (Belsley *et al*., 1980). Il est préférable de calculer l'estimateur MCO à partir de la décomposition en valeurs singulières :

$$\underset{(n,\,p)}{X} = \underset{(n,\,p)}{U}\ \underset{(p,\,p)}{D}\ \underset{(p,\,p)}{V^T} \qquad (17)$$

D est la matrice diagonale des valeurs singulières de *X*. *U* est la matrice orthogonale des *p* vecteurs propres associés aux *p* valeurs propres non nulles de $X X^T$, et *V* est la matrice orthogonale des vecteurs propres de $X^T X$. En notant

(2) Le processus, présenté pour la construction de l'IC bootstrap, s'applique de manière identique pour la construction de l'intervalle de prédiction bootstrap.

(3) Nous avons retenu la médiane comme indicateur de valeur centrale car, contrairement à la moyenne, elle est insensible aux variations des valeurs extrêmes de la distribution.

$X^+ = V D^+ U^T$ la pseudo-inverse de X, l'estimateur MCO s'exprime comme :

$$\hat{\beta} = X^+ y \tag{18}$$

et sa matrice de variance-covariance estimée s'écrit :

$$\hat{V}(\hat{\beta}) = s^2 V D^{-2} V^T$$

La matrice de projection qui lie les résidus et le terme d'erreur s'exprime simplement comme :

$$I_n - X (X^T X) X^T = I_n - U U^T \tag{19}$$

Nous avons comparé les performances obtenues en calculant classiquement l'estimateur MCO par inversion de matrice *(éq. (2))* et par pseudo-inverse *(éq. (18)).* Après un ensemble de tests sur Matlab (Juan, 1999), les deux méthodes ont été programmées directement en Fortran. L'algorithme le mieux adapté pour l'inversion de la matrice $X^T X$, qui est définie positive, repose sur une décomposition de Cholesky (Seak, 1972). Une analyse de ses performances numériques est fournie dans Lantz (1983). Le calcul de la pseudo-inverse X^+ a été réalisé en reprenant l'algorithme de Golub et Reinsch, qui est décrit dans Forsythe *et al.* (1977).

La multicolinéarité a moins de conséquences sur le conditionnement de la matrice X lorsque l'on considère des données normées ou centrées et réduites (Belsley *et al*., 1980 ; Belsley, 1984 ; Erkel-Rousse, 1995). Nous avons testé l'incidence de ces transformations pour différentes dimensions de la matrice des variables explicatives (n = 10, …, 50 ; p = 2, …, 10). Pour chaque dimension nous avons généré 1000 matrices X telles que X_1^T suit une loi uniforme (0, 1), et les vecteurs j suivants sont construits comme $X_j^T = X_{j-1}^T + \nu$ où ν suit une loi uniforme (0, 0,001). Nous avons mesuré les erreurs de calcul comme la somme des valeurs absolues des écarts entre $X X^+$ ou $(X^T X)(X^T X)^{-1}$ et la matrice identité. Le tableau 1 résume ces écarts pour différentes dimensions (n, p)[4]. Ils conduisent à estimer le vecteur des paramètres par pseudo-inverse sur une matrice X normée à 1 ou, le cas échéant, à le calculer par inversion de matrice sur des données centrées et réduites lorsque p est faible.

(4) L'ensemble des résultats est disponible auprès des auteurs.

Nous avons ensuite comparé les temps de calcul pour ces deux méthodes sur le bootstrap des résidus (en utilisant l'algorithme de tri décrit dans la section suivante). Nous avons généré 1000 échantillons de données pour différentes dimensions (n, p), comme précédemment. Pour chaque échantillon, nous avons appliqué un bootstrap des résidus et la méthode percentile avec 1000 réplications. Le tableau 2 donne les temps de calcul obtenus sur un micro-ordinateur de type PC (Pentium II, 300 MHz). L'utilisation d'une pseudo-inverse divise les temps de calcul par un facteur proche de deux en évitant le centrage des données.

TABLEAU 2

Temps de calcul (s) pour le bootstrap des résidus et la méthode percentile sur 1000 réplications

Computation time (s) for residual bootstrap associated to the percentile method with 1000 replications

Dimension de X	**X^+ Données normées à 1**	**$(X^T X)^{-1}$ Données centrées et réduites**
(10, 2)	0,0103	0,0208
(50, 4)	0,0518	0,0972
(90, 6)	0,1001	0,1823

2.3 L'algorithme de tri des réplications

Les deux algorithmes de tri les plus répandus sont le tri par comparaison de tous les éléments et le tri par partition. Cette dernière méthode est privilégiée pour de grands échantillons puisqu'elle requiert de l'ordre de $B \log_2 B$ opérations pour trier les B réplications, alors que la méthode classique nécessite de l'ordre de $B^2/2$ opérations.

Le tri modifié que nous proposons consiste à ne s'intéresser qu'à la détermination des fractiles α et $1 - \alpha$. Ainsi, dans la phase de tri, nous traitons uniquement $\alpha B + 1$ réplications bootstrap, à la différence de l'algorithme classique, dans lequel la totalité des réplications *(B)* est triée. Il peut être décomposé en deux étapes.

TABLEAU 1

Erreur de calcul sur les produits de matrices $X X^+$ et $(X^T X)(X^T X)^{-1}$

Computation error on $X X^+$ and $(X^T X)(X^T X)^{-1}$

Dimension de X	**$X X^+$ Données normées à 1**	**$(X^T X)(X^T X)^{-1}$ Données normées à 1**	**$(X^T X)(X^T X)^{-1}$ Données centrées et réduites**
(50, 2)	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$
(50, 4)	$< 10^{-10}$	$0{,}61 \times 10^{-6}$	$0{,}11 \times 10^{-6}$
(50, 8)	$< 10^{-10}$	$0{,}54 \times 10^{-5}$	$0{,}59 \times 10^{-6}$
(50, 10)	$< 10^{-10}$	$0{,}50 \times 10^{-4}$	$0{,}50 \times 10^{-5}$

La première consiste à construire un vecteur S_1 de dimension $\alpha B + 1$, constitué des $\alpha B + 1$ premières statistiques $\hat{\theta}^*(b)$[5], classées par ordre croissant. Ainsi, $S_1(1)$ et $S_1(\alpha B + 1)$ correspondent respectivement à la plus petite et la plus grande valeur des $\alpha B + 1$ premières statistiques bootstrap. S_1 est ensuite dupliqué dans un second vecteur S_2, de dimension identique.

Dans une deuxième étape, chacune des $B - (\alpha B + 1)$ statistiques suivantes est comparée avec les éléments de S_1. Le vecteur S_1 est utilisé pour la recherche des plus petits éléments, parmi les B réplications bootstrap, suivant la procédure suivante. Cette dernière est appliquée pour chaque itération bootstrap b, $b = \alpha B + 1, \ldots, B$. $\hat{\theta}^*(b)$ est tout d'abord comparé au plus grand élément de S_1. S'il est plus grand ou égal à celui-ci, on passe à la comparaison avec S_2. Dans le cas contraire, $\hat{\theta}^*(b)$ est alors comparé à chacun des éléments du vecteur S_1. S'il est inférieur ou égal à l'élément k de S_1, $S_1(k)$ est remplacé par $\hat{\theta}^*(b)$ et, pour chaque indice supérieur à k de ce vecteur, les éléments sont décalés d'un rang, vers le rang supérieur. Finalement, le vecteur S_1 contient les $\alpha B + 1$ plus petites valeurs des B statistiques $\hat{\theta}^*(b)$. Ainsi, le quantile α correspond à l'élément $S_1(\alpha B)$.

De manière similaire on compare les $B - (\alpha B + 1)$ statistiques bootstrap avec les éléments de S_2. Il s'agit donc du vecteur utilisé pour la recherche des plus grands éléments, parmi les B réplications bootstrap. Ainsi, le quantile $1 - \alpha$ de la distribution bootstrap correspond à l'élément $S_2(1)$. Notons qu'il est nécessaire de classer $\alpha B + 1$, réplications bootstrap, et non pas αB pour extraire le quantile $1 - \alpha$. En effet, ce dernier correspond au premier élément du vecteur S_2, de dimension $(\alpha B + 1)$.

Dans cette version améliorée de l'algorithme, deux étapes sont distinguées : le tri des $\alpha B + 1$ premières réplications bootstrap puis, pour chaque réplication suivante, les comparaisons avec les éléments de S_1 et S_2. Le nombre de comparaisons de la première étape est égal à $[\alpha B (\alpha B + 1)]/2$. Celui de la seconde étape peut être encadré par un cas minimum, où chaque réplication est toujours inférieure à $S_2(1)$ et supérieure à $S_1(\alpha B + 1)$, et un cas maximum où elle est toujours comparée aux $\alpha B + 1$ éléments de S_1 et de S_2. Il est donc compris entre $[B - (\alpha B + 1)] \times 2$ et $[B - (\alpha B + 1)] \times (\alpha B + 1) \times 2$.

Le nombre total de comparaisons, dans la version améliorée de l'algorithme de tri, est compris entre $[(\alpha B + 1) \times (\alpha B - 1)] + B$ et $(\alpha B + 1) \times (B - 1)$, il est d'ordre B^2. Asymptotiquement, le nombre d'opérations de cet algorithme est du même ordre que pour un tri classique. Cependant, pour $B = 5000$ et $\alpha = 0{,}025$, le nombre d'opérations effectuées est de l'ordre de 12 millions avec l'algorithme classique, et compris entre 20 000 et 600 000 avec la version améliorée de l'algorithme. Ainsi, cette dernière permet de diviser le nombre de comparaisons au moins par 20.

(5) Elles correspondent aux valeurs de la statistique bootstrap, pour les $b = 1, \ldots, \alpha B + 1$ premières itérations.

3 APPLICATIONS À L'ESTIMATION DES COÛTS DANS LE SECTEUR AUTOMOBILE

Le marché automobile est dans une situation de forte concurrence. La réduction des coûts constitue un objectif majeur des firmes pour renforcer leur compétitivité. Ainsi, dans les pays industrialisés, les marchés automobiles sont saturés et la compétition entre les constructeurs se fait désormais par les prix. Verboven (1996) propose une analyse du marché automobile européen en termes de concurrence oligopolistique. Dans les pays émergents où les ventes sont en progression, le pouvoir d'achat des consommateurs est plus faible et les industriels doivent maîtriser leurs coûts.

L'estimation des coûts est effectuée à tous les stades, de l'avant-projet à la production des véhicules. En effet, dès les premières phases de conception d'une nouvelle automobile (c'est-à-dire environ 36 mois avant la fabrication), les choix techniques opérés fixent une large partie des coûts futurs (Juan, 1999). La prédiction des coûts revêt donc un enjeu majeur.

L'économétrie fournit une méthode d'estimation et de prédiction des coûts particulièrement utile à cette fin. En effet, les modèles économétriques permettent d'expliquer les coûts en fonction des quelques paramètres techniques qui constituent la seule information disponible au démarrage d'un projet.

Cependant, l'estimation de tels modèles soulève des difficultés liées à la faible taille des échantillons de données et de la distribution asymétrique des erreurs dans les modèles de régression. Les méthodes de bootstrap permettent de pallier ces difficultés en fournissant une approximation de la distribution des erreurs de prédiction par leur distribution empirique. Nous illustrons ceci au travers de deux exemples. Au préalable, nous définissons le coût étudié et nous présentons les données utilisées.

3.1 Les données et la forme générale des modèles

La voiture est constituée de plusieurs milliers de pièces et accessoires, formant des sous-ensembles (ou fonctions) tels que le moteur, la caisse, la direction, etc. Pour un modèle de véhicule donné, il existe différentes versions, en termes de niveau d'équipement, de motorisation, etc. Ainsi, pour effectuer des prévisions de coûts pour tous les véhicules du modèle considéré, des modèles d'estimation de coût sont élaborés au niveau des sous-ensembles. L'entité dont le coût est modélisé correspond donc le plus souvent à un sous-ensemble de pièces assemblées.

Par ailleurs, il importe de préciser le périmètre des coûts étudiés. Notre travail porte sur le coût de production, nommé *prix de revient de fabrication* (PRF) *hors amortissements*. Ce dernier, représentatif des dépenses engagées par l'entreprise pendant la phase de production du bien, est composé des achats (de matières et de pièces œuvrées extérieures) et d'une valeur de transformation. Le PRF représente environ 60 % du

coût complet d'un véhicule (ce dernier incluant les coûts de garantie, de logistique, etc.). Notons que le PRF hors amortissements n'inclut pas les amortissements du capital.

La spécificité des modèles de coût implique de préciser le contexte de production industrielle. En premier lieu, les coûts des différents éléments de la base de données sont normalisés, pour un volume de fabrication moyen, représentant les conditions de production « habituelles » pour la famille de produits. Nous nous affranchissons ainsi des effets d'échelle en travaillant sur des coûts unitaires de fabrication.

En second lieu, l'état de la technologie est considéré comme donné. En effet, les changements technologiques se traduisent principalement par de nouvelles machines, plus performantes. Or, le périmètre du coût étudié n'inclut pas les amortissements du capital.

Ainsi, le cadre d'analyse nous amène à spécifier des modèles de coût pour un volume de production normalisé et un état donné du système productif. Par ailleurs, nos besoins en modélisation nous conduisent à utiliser des données en « coupe transversale » du coût de chacun des éléments de la base de données à l'instant *t*. Les estimations sont donc effectuées dans un cadre statique et n'intègrent pas les phénomènes de progrès technique.

Le coût est donc expliqué par les caractéristiques techniques pertinentes du produit, à un instant donné de la technologie, à des fins de prédiction de coût pour de nouveaux produits. Notre démarche est donc différente de celle utilisée dans l'application des techniques de bootstrap sur les modèles de frontières de production (Simar, 1992). En effet, dans ce cas, il s'agit d'évaluer l'efficience d'une unité de production par rapport à une frontière efficace, à l'aide du bootstrap. Finalement, notons que si aucun *a priori* n'est posé quant à la forme de la relation liant le coût aux descripteurs, les formes linéaires ou multiplicatives sont le plus souvent retenues.

3.2 Le modèle d'estimation de coût des moteurs

L'exemple porte sur le périmètre du moteur, avec ses composants électriques. De manière très schématique, le moteur est composé du carter cylindre, de l'attelage mobile (vilebrequin, volant moteur, etc.), de pièces assurant la distribution (arbre à cames, soupapes, etc.), et de la culasse (y compris son couvercle). Les équipements électriques comprennent le démarreur, la bobine et les bougies d'allumage, l'alternateur, etc.

Le moteur constitue, avec l'assemblage final du véhicule, une fonction dont la fabrication est, le plus souvent, interne à l'entreprise. Les données de coûts sont donc des PRF hors amortissements. Par ailleurs, le moteur représente une partie non négligeable (de l'ordre de 20 %) du PRF d'un véhicule.

Il existe différentes familles de moteurs, distinguées principalement par le mode de carburation (essence ou diesel), le mode d'injection (directe, monopoint, etc.) et la cylindrée. L'étude porte sur des moteurs de cylindrée supérieure à 1700 cm^3, ces derniers formant un échantillon de travail homogène de 15 moteurs.

Notons, dans le périmètre technique du moteur, la présence des composants supplémentaires (support, tuyaux, etc.), imposés par des équipements tels que la direction assistée ou le conditionnement d'air. Ceux-ci sont présents ou non sur le moteur, en fonction des prestations à assurer par le véhicule sur lequel ils sont destinés à être montés.

Dans l'équation économétrique, le coût d'un moteur est exprimé comme une fonction de sa cylindrée. Le modèle possède une variable explicative et un terme constant ; il est estimé par MCO. Le R^2 est égal à 0,97 et les coefficients estimés[6] de la régression sont significativement différents de zéro, pour un risque de première espèce de 5 %. Le test de White $F(2, 12) = 0{,}66$ ne permet pas de rejeter l'hypothèse d'homoscédasticité.

Les résidus transformés de la régression $\tilde{u}$, à partir desquels sont constitués les échantillons dans la procédure de bootstrap des résidus, sont illustrés dans la figure 1. Nous remarquons une queue de distribution plus longue vers la droite, traduisant la présence de résidus positifs élevés. En effet, le résidu « extrême », correspondant à un moteur de 1783 cm^3, que nous notons $\hat{u}_6$, possède la valeur la plus élevée de l'ensemble des résidus. Ce moteur, ainsi qu'un second dans l'échantillon, est équipé du conditionnement d'air et correspond aux résidus les plus élevés de la régression.

(6) Pour des raisons de confidentialité, les valeurs des paramètres estimés ne sont pas reproduites.

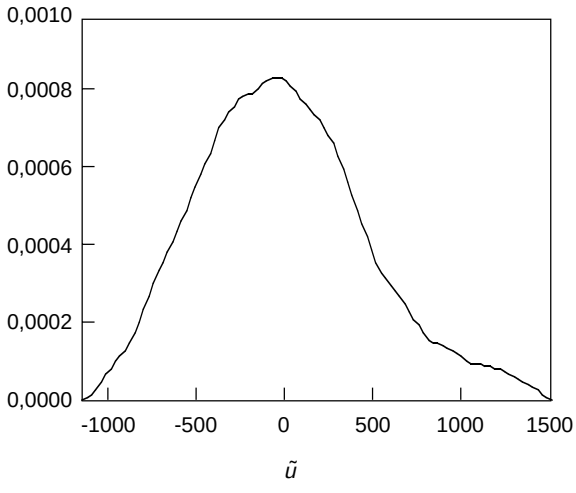

Figure 1

Estimateur à noyau de la densité des résidus transformés.
Kernel density of transformed residuals.

La procédure de détermination du nombre de réplications bootstrap nécessaires est ensuite mise en œuvre *(section 2.1)*. Les résultats présentés portent sur l'exemple du coefficient de la cylindrée z_1^*.

3.2.1 Le nombre de réplications *B*

Les résultats des tests de Wilcoxon pour l'égalité des médianes et des tests de Fisher pour l'égalité des variances des étendues des intervalles de confiance de z_1^* ne permettent pas d'établir de manière certaine une valeur de *B* à partir de laquelle nous pourrions accepter l'hypothèse d'égalité des médianes et des variances des IC *(tableau 3)*.

L'évolution du CV en fonction du nombre de réplications est représentée sur la figure 2. La décroissance du CV est forte, pour les faibles valeurs de *B*, puis s'affaiblit progressivement. Ainsi, à partir de $B = 5000$, il se stabilise. Ce troisième critère nous permet finalement de sélectionner le nombre de réplications bootstrap $B = 5000$, à partir duquel

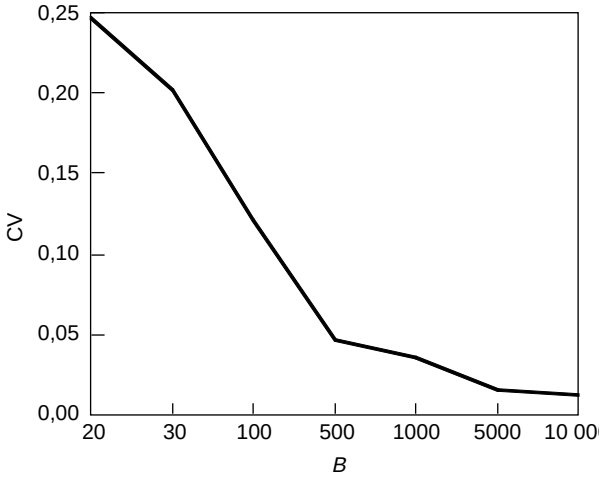

Figure 2

Coefficient de variation des étendues des IC de z_1^*.
Coefficient of variation of CI ranges of z_1^*.

TABLEAU 3

Tests des étendues des IC de z_1^*
Tests of CI ranges of z_1^*

B	Médiane des étendues	Test de Wilcoxon	P-marg.	Écart type des étendues	Test de Fisher	P-marg.
20	3,910			0,973		
30	4,307	3,676	0,000	0,889	1,197	0,372
100	4,224	0,864	0,388	0,520	2,924	0,000
500	4,273	1,130	0,258	0,203	6,565	0,000
1000	4,304	0,458	0,647	0,150	1,830	0,003
5000	4,301	0,030	0,976	0,067	5,019	0,000
10 000	4,317	1,604	0,109	0,051	1,727	0,007

TABLEAU 4

Prévisions MCO et intervalles de prédiction standard et bootstrap (FF)
OLS predictions and standard bootstrap prediction intervals (FF)

Cyl. (cm^3)	Prévision MCO	Intervalle de prédiction standard			
		2,5 %	97,5 %	Étendue	Forme*
1900	9133,42	8160,67	10106,20	1945,53	1,00

Cyl. (cm^3)	Intervalle de prédiction percentile				Intervalle de prédiction percentile			
	2,5 %	97,5 %	Étendue	Forme	2,5 %	97,5 %	Étendue	Forme
1900	8419,66	10179,57	1759,91	1,47	8370,00	10271,97	1901,96	1,49

* La forme est définie comme : $\frac{\sup - \hat{y}_f}{\hat{y}_f - \inf}$, avec inf et sup correspondant respectivement aux bornes inférieure et supérieure de l'intervalle de prédiction bootstrap, $\hat{y}_f$ étant la prévision MCO.

les étendues des IC de z_1^* (et donc les fractiles 2,5 % et 97,5% de la distribution) peuvent être considérées comme stables. De manière similaire, B = 5000 réplications sont retenues pour la construction des intervalles de prédiction bootstrap.

3.2.2 La prédiction de coût

L'ajustement de la régression est maintenant utilisé pour prévoir les PRF d'un nouveau moteur, de cylindrée égale à 1900 cm^3. Nous reportons, pour les intervalles de prédiction bootstrap, les deux méthodes de construction : percentile et percentile-*t (tableau 4).* Les intervalles percentiles possèdent une étendue nettement plus réduite que les intervalles percentiles-*t* ou standard. Ainsi, la méthode percentile conduit à des intervalles trop « optimistes » (trop petits) et n'apparaît pas pertinente pour ce type d'investigation. En effet, l'erreur de prédiction n'est pas une statistique pivot et l'inférence bootstrap peut s'avérer erronée dans ce cas. Notons que les intervalles de prédiction bootstrap sont décalés vers les plus fortes valeurs de coût (forme supérieure à 1) par rapport aux intervalles standard.

3.2.3 Le repérage du résidu « extrême » dans les réplications bootstrap pour la prédiction

Afin d'expliquer l'asymétrie des intervalles de prédiction bootstrap, nous étudions l'impact du tirage du résidu extrême $\hat{u}_6$ dans le PGD bootstrap sur la distribution de l'erreur de prédiction bootstrap. Pour ce faire, nous examinons, pour chacune des réplications, le résidu $\tilde{u}_f^*$ du modèle théorique bootstrap de la prédiction et vérifions s'il correspond ou non au résidu $\hat{u}_6$. Le tableau 5 présente, pour ces deux possibilités, les valeurs moyennes et écarts types de z_f^* (la statistique bootstrap de l'erreur de prédiction normée), pour les 5000 réplications. La figure 3 illustre, pour la prédiction du coût d'un moteur de 1900 cm^3, la distribution de l'erreur de prédiction bootstrap, en distinguant les cas où le résidu extrême est tiré.

TABLEAU 5

Les caractéristiques de z_f^*

Characteristics of z_f^*

Moteur	z_f^*	$\tilde{u}_f^* \neq \hat{u}_6$	$\tilde{u}_f^* = \hat{u}_6$	Total
1900 cm^3	Moyenne	0,144	– 2,472	– 0,022
	Écart type	0,896	0,745	1,092
	Effectifs	4682	318	5000

La distribution bootstrap de l'erreur de prédiction normée z_f^* paraît fortement asymétrique vers les valeurs négatives. De plus, nous constatons *(fig. 3)* que cette asymétrie est due au tirage du résidu extrême $\hat{u}_6$ dans le modèle théorique de prévision bootstrap. En effet, comme $e_f^* = \hat{y}_f^* - \hat{y}_f - \tilde{u}_f^*$, si $\tilde{u}_f^* = \hat{u}_6$ (forte valeur positive), l'erreur de prédiction bootstrap est fortement négative, ce que nous retrouvons dans l'histogramme. Ainsi, l'intervalle de prédiction résultant est décalé vers les plus grandes valeurs de coût, puisque le fractile 2,5 % de la distribution détermine la borne supérieure de l'intervalle et vice versa pour le fractile 97,5 %. Notons que ce phénomène se répète pour la distribution de la statistique z_f^* effectuée pour d'autres prédictions (Juan, 1999).

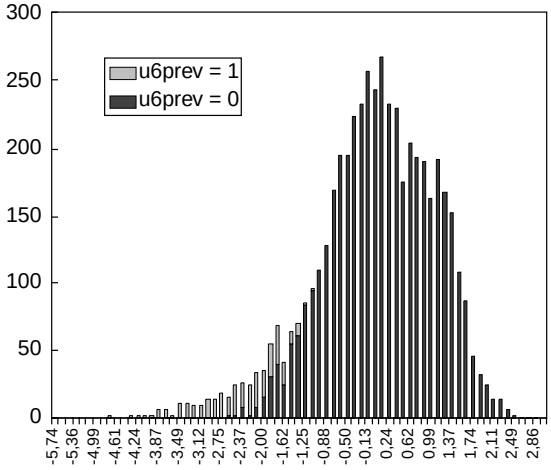

Figure 3

Distribution de la statistique z_f^* pour un moteur de 1900 cm^3.

Distribution of statistic z_f^* *for a 1900 cm^3 engine.*

3.3 Le modèle d'estimation de coût des réservoirs à carburant

Cet exemple porte sur un modèle de coût des réservoirs dans lequel intervient une variable binaire. Le périmètre technique du produit étudié est le réservoir à carburant, équipé de la jauge à carburant et de la pompe d'aspiration. Le réservoir est réalisé en plastique, selon un procédé de soufflage de la matière. Il fait partie d'un sous-ensemble plus vaste du véhicule : le circuit à carburant.

Le réservoir équipé constitue un exemple de fonction du véhicule dont la conception et la fabrication sont entièrement externalisées et incombent aux fournisseurs équipementiers. Le réservoir est donc entièrement une « pièce œuvrée extérieure ». Le prix de revient du réservoir, pour le constructeur, est le prix d'achat. Ce dernier comporte la marge du fournisseur. Elle peut cependant être considérée constante, au regard des conditions de vente correspondant à des volumes constants, que ce soit pour le constructeur automobile ou pour les fournisseurs, dont les produits constituent notre échantillon.

Par ailleurs, il importe de distinguer les réservoirs diesels des réservoirs essence. La discrimination s'explique techniquement par la présence d'une pompe intégrée dans

l'ensemble d'aspiration sur les réservoirs essence, qui n'équipe pas les réservoirs diesel. De plus, l'essence est beaucoup plus volatile que le gazole et les normes de dépollution imposent une perméabilité maximale du réservoir à respecter. Ces contraintes impliquent donc un traitement anti-évaporation spécifique. Notons que le coût du procédé de fluoration dépend essentiellement de la norme de dépollution et, marginalement, de la capacité du réservoir.

Ainsi, les réservoirs essence présentent un surcoût par rapport à leurs homologues diesel. De plus, ce surcoût est variable. En effet, des contraintes d'architecture et de conception peuvent impliquer, selon le véhicule, un double puits (un pour le jaugeage et un pour l'aspiration).

Le modèle économétrique explique le coût en fonction de la capacité du réservoir et d'une variable muette « carburant » qui vaut zéro si le réservoir contient du gazole et un s'il contient de l'essence. Il est estimé par MCO. Le R^2 = 0,95 laisse apparaître qu'une large part de la variance du coût est expliquée par ces deux variables ; les paramètres estimés des variables capacité et carburant sont significatifs, pour un risque de première espèce de 5 % ; le test de White $F(3, 11) = 1,29$ ne permet pas de rejeter l'hypothèse d'homoscédasticité.

Par ailleurs, nous avons effectué un test de Fisher, afin de juger de la validité d'un modèle global, par rapport à deux régressions distinctes, sur chaque type de réservoir (essence et diesel) : $F(1, 11) = 0,54$, qui est inférieur à la valeur critique au seuil 5 %. Ainsi, le test ne permet pas de rejeter l'hypothèse nulle d'une seule régression pour les réservoirs essence et diesel.

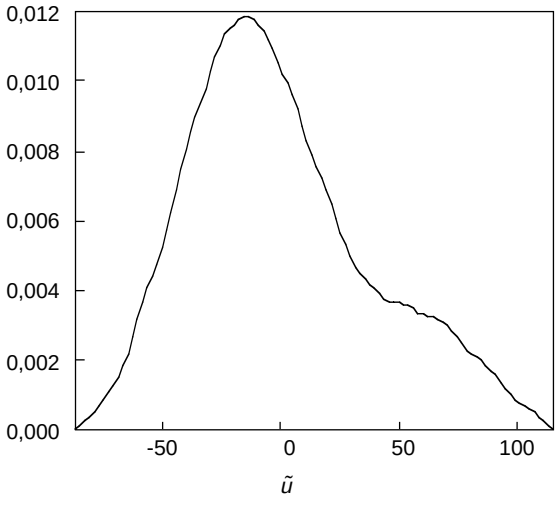

Figure 4

Estimateur à noyau de la densité des résidus transformés.
Kernel density of transformed residuals.

La figure 4 présente l'estimateur à noyau de la densité des résidus transformés $\tilde{u}$, sans distinction du type de réservoir. La queue de distribution est nettement plus longue vers la droite, traduisant la présence de résidus positifs importants. En effet, les résidus « extrêmes » correspondant aux observations n^{os} 2 et 8 (réservoirs essence) possèdent des valeurs élevées. D'un point de vue technique, il s'avère que ceux-ci, contrairement aux autres réservoirs essence, sont munis d'un double puits, imposé par les contraintes d'architecture du véhicule.

3.3.1 Les procédures de bootstrap des résidus classiques et stratifiés

Nous étudions le processus de construction des échantillons bootstrap, qui, en présence d'une variable muette dans la régression, présente certaines spécificités. En effet, la procédure de bootstrap des résidus « classiques », telle que nous l'avons présentée dans la section 1.2, ne respecte pas la structure de l'échantillon initial. Dans le PGD bootstrap, les résidus (essence ou diesel) sont affectés aléatoirement, à la capacité d'un réservoir diesel par exemple. Intuitivement, cette procédure ne semble pas satisfaisante puisque la démarche sous-jacente au bootstrap consiste à générer des échantillons artificiels le plus proches possible de l'échantillon initial. Dès lors, en présence d'une (ou plusieurs) variable(s) explicative(s) de type qualitatif dans le modèle, nous avons adopté une version adaptée du bootstrap des résidus en stratifiant l'échantillon bootstrap. La construction de celui-ci est détaillée ci-dessous.

Soit $\tilde{u}$ le vecteur de dimension n = 15 des résidus transformés de la régression. Ce vecteur est scindé en deux sous-échantillons : $\tilde{u}_1$, de taille n_1 = 8, composé des résidus associés aux réservoirs essence, et $\tilde{u}_2$, de taille n_2 = 7, des sept résidus associés aux réservoirs diesel. Nous effectuons respectivement huit (sept) tirages aléatoires avec remise dans $\tilde{u}_1$ ($\tilde{u}_2$) pour constituer $\tilde{u}_1^*$ ($\tilde{u}_2^*$). La concaténation des deux vecteurs $\tilde{u}_1^*$ et $\tilde{u}_2^*$ forme ensuite le vecteur des résidus bootstrap ré-échantillonnés « par strates » : $\tilde{u}^*$ de dimension n = 15. Les deux procédures de bootstrap des résidus (classiques et stratifiés) sont mises en œuvre et comparées lors de la construction d'intervalles de prédiction bootstrap.

3.3.2 La prédiction de coût

Les prévisions de coût et leurs intervalles sont calculés pour des réservoirs diesel et essence, de capacité de 60 litres *(tableaux 6, 7 et 8)*.

Les intervalles de prédiction obtenus avec la procédure de bootstrap des résidus classiques sont proches, en termes d'étendue, des intervalles standard. Leurs formes sont asymétriques vers les plus fortes valeurs de coûts, ceci de manière identique pour les réservoirs essence et diesel. Ainsi, cette procédure, qui réaffecte de manière aléatoire les résidus,

TABLEAU 6

Prévisions MCO et intervalles de prédiction standard et bootstrap (FF)

OLS predictions and standard bootstrap prediction intervals (FF)

Réservoir	**Prévision MCO**	**Intervalle de prédiction standard**			
		2,5 %	**97,5 %**	**Étendue**	**Forme**
60 l diesel	260,24	182,74	337,74	157,00	1,00
60 l essence	507,50	430,37	584,63	154,26	1,00

TABLEAU 7

Intervalles de prédiction bootstrap des résidus classiques (FF)

Bootstrap prediction intervals of classic residuals (FF)

Réservoir	**Intervalle de prédiction percentile**				**Intervalle de prédiction percentile-*t***			
	2,5 %	**97,5 %**	**Étendue**	**Forme**	**2,5 %**	**97,5 %**	**Étendue**	**Forme**
60 l diesel	203,20	342,22	139,02	1,44	203,33	354,27	150,94	1,65
60 l essence	449,90	588,30	138,40	1,40	451,73	599,26	147,53	1,64

TABLEAU 8

Intervalles de prédiction bootstrap des résidus stratifiés (FF)

Bootstrap prediction intervals of stratified residuals (FF)

Réservoir	**Intervalle de prédiction percentile**				**Intervalle de prédiction percentile-*t***			
	2,5 %	**97,5 %**	**Étendue**	**Forme**	**2,5 %**	**97,5 %**	**Étendue**	**Forme**
60 l diesel	220,52	308,91	88,39	1,22	215,33	312,98	97,65	1,17
60 l essence	441,50	594,94	153,44	1,32	446,19	619,38	173,19	1,82

TABLEAU 9

Les caractéristiques de z_f^* pour un réservoir de 60 litres

Characteristics of z_f^* for a 60 liter tank

	Réservoir 60 litres	z_f^*	$\tilde{u}_f^* \neq (\hat{u}_2$ **ou** $\hat{u}_8)$	$\tilde{u}_f^* = (\hat{u}_2$ **ou** $\hat{u}_8)$	**Total**
Bootstrap des résidus classiques	Diesel	Moyenne	0,256	– 1,984	– 0,028
		Écart type	0,793	0,761	1,086
		Effectif	4365	635	5000
	Essence	Moyenne	0,265	– 2,062	– 0,033
		Écart type	0,790	0,895	1,119
		Effectif	4360	640	5000
Bootstrap des résidus stratifiés	Diesel	Moyenne	0,010	0,002	0,008
		Écart type	0,733	0,772	0,743
		Effectif	3723	1277	5000
	Essence	Moyenne	0,603	– 1,974	– 0,035
		Écart type	0,685	0,897	1,337
		Effectif	3763	1237	5000

retranscrit l'asymétrie de leur distribution indifféremment sur les deux types de réservoirs. Or, cette asymétrie, causée par les réservoirs essence, n'a pas lieu d'être reportée sur les réservoirs diesel.

La procédure de bootstrap des résidus stratifiés fournit des intervalles de prédiction dont l'étendue est réduite (de l'ordre de 100 FF) pour les réservoirs diesel et plus large (175 FF) pour les réservoirs essence. De plus, ces intervalles sont symétriques pour les réservoirs diesel et fortement asymétriques vers la droite pour les réservoirs essence.

3.3.3 Le repérage des résidus « extrêmes » dans les réplications bootstrap pour la prédiction

L'impact du tirage des résidus extrêmes sur la distribution de l'erreur de prédiction bootstrap est étudié, en vérifiant, pour chaque réplication, si le résidu $\tilde{u}_f^*$ du modèle théorique bootstrap de la prédiction correspond à $\hat{u}_2$ ou $\hat{u}_8$. Le tableau 9 présente, suivant $\tilde{u}_f^*$, les valeurs moyennes et écarts types de la statistique bootstrap de l'erreur de prédiction normée, pour 5000 réplications bootstrap.

La figure 5 illustre la distribution empirique bootstrap de l'erreur de prédiction lorsque la procédure de bootstrap des résidus est stratifiée. La distribution bootstrap de l'erreur de prédiction normée z_f^* paraît fortement asymétrique vers les valeurs négatives. Cette asymétrie est due au tirage des résidus extrêmes $\hat{u}_2$ ou $\hat{u}_8$, dans le modèle théorique de prévision bootstrap *(fig. 5)*. En effet, comme $e_f^* = \hat{y}_f^* - \hat{y}_f - \tilde{u}_f^*$, si $\tilde{u}_f^* = \hat{u}_2$ ou $\hat{u}_2$ (fortes valeurs positives), l'erreur de prédiction bootstrap est fortement négative. Ce phénomène se reproduit pour la distribution de la statistique z_f^* d'autres prédictions (Juan, 1999).

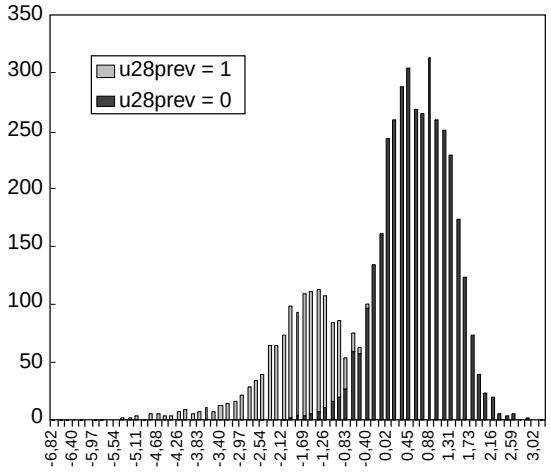

Figure 5
Distribution de la statistique z_f^* pour un réservoir essence de 60 litres (bootstrap des résidus stratifiés).
Distribution of statistic z_f^* for a 60 liter gasoline tank (bootstrap of stratified residuals).

CONCLUSION

L'application des méthodes de bootstrap sur les modèles de régression fournit une approximation de la distribution des erreurs de prédiction par leur distribution empirique lorsque celle-ci est inconnue. Le bootstrap est ainsi particulièrement utile lorsque les échantillons de données sont de petite taille et qu'il n'est pas possible de formuler l'hypothèse d'une distribution gaussienne du terme d'erreur. Le nombre de réplications peut être déterminé à partir des coefficients de variation de l'étendue de l'intervalle de confiance des coefficients, ou de l'intervalle de prédiction lorsque ceux-ci deviennent peu variants.

La mise en œuvre du bootstrap invite à privilégier le calcul des paramètres estimés en utilisant la pseudo-inverse de la matrice des variables explicatives. Cette méthode de calcul s'avère robuste en présence de multicolinéarité. Elle est également performante en termes de temps de calcul car il est suffisant de réduire les données, qui sont alors normées à 1, avant de calculer l'estimateur, plutôt que de les centrer et de les réduire comme pour une inversion de matrice.

Un algorithme modifié permet de trier rapidement la distribution empirique de la statistique d'intérêt. Seules les queues de distribution sont triées et les éléments sont comparés aux valeurs extrêmes de celles-ci qui correspondent aux fractiles retenus.

La prévision des coûts au stade des avant-projets dans l'industrie automobile soulève des difficultés liées à la faible taille des échantillons de données et à l'asymétrie qui caractérise les distributions de coût. Deux applications permettent d'apprécier l'apport du bootstrap.

Le premier exemple développé illustre l'utilisation des techniques de bootstrap sur un modèle simplifié de coût d'un moteur. L'analyse des intervalles de prédiction montre que le bootstrap permet de retranscrire l'asymétrie de la distribution des résidus dans les intervalles de prédiction. En effet, ces derniers sont décalés, par rapport aux intervalles standard, vers les plus fortes valeurs de coûts et autorisent ainsi, pour la prédiction du coût d'un nouveau moteur, des valeurs plus élevées. L'utilisation des techniques de bootstrap permet donc une meilleure retranscription de l'information contenue dans l'échantillon initial pour les intervalles de prédiction.

Le second exemple a permis d'exposer une utilisation des techniques de bootstrap adaptée, dans le cas d'une modélisation en présence de variables muettes. Cette méthode permet de construire des intervalles de prédiction symétriques pour les réservoirs diesels, et asymétriques vers les plus fortes valeurs de coûts pour les réservoirs essence. Ceci résulte de l'asymétrie, à la fois de la distribution de l'erreur de prédiction et de celle du surcoût lié au carburant essence. Ainsi, l'utilisation de la procédure de bootstrap des résidus stratifiés permet, lorsque l'information n'est pas disponible, de prendre en compte, dans les intervalles de confiance et de prédiction, des surcoûts éventuels pour les réservoirs essence, imposés par les contraintes d'architecture du véhicule.

Des développements dans l'utilisation du bootstrap sont envisagés, notamment la prise en compte d'éventuelles non-linéarités dans la spécification des fonctions de coût. En effet, bien que ceci n'ait pas été mis en évidence dans les exemples traités, il faut envisager la non-linéarité de la relation comme une alternative à un modèle linéaire avec terme d'erreur asymétrique. L'utilisation du bootstrap concerne alors les tests de spécifications ainsi que l'estimation des paramètres du modèle.

RÉFÉRENCES

Belsley, D. (1984) Demeaning Conditioning Diagnostic Through Centering. *The American Statistician*, **38**, 2, 73-93.

Belsley, D., Kuh, E. et Welsch, R. (1980) *Regression Diagnostics, Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York.

Booth, J.G. et Sarkar, S. (1998) Monte-Carlo Approximation of Bootstrap Variances. *The American Statistician*, **52**, 4, 354-357.

Breiman, L. (1992) The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error. *Journal of the American Statistical Association*, **87**, 738-754.

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

Efron, B. et Tibshyrani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Erkel-Rousse, H. (1995) Détection de la multicolinéarité dans un modèle linéaire ordinaire : quelques éléments pour un usage averti des indicateurs de Belsley, Kuh et Welsch. *Revue de statistique appliquée*, **18**, 4, 19-42.

Flachaire, E. (1998) Les méthodes du bootstrap et l'inférence robuste à l'hétéroscédasticité, *Thèse*, université de la Méditerranée, GREQAM.

Forsythe, G.E., Malcolm, M.A. et Moler, C.B. (1977) *Computer Methods for Mathematical Computations*, Prentice-Hall.

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

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

Horowitz, J.L. (1997) *Bootstrap Methods in Econometrics: Theory and Numerical Performance, Advances in Economics and Econometrics: Theory and Application*, **3**, Cambridge University Press, 188-222.

Jeong, J. et Maddala, G.S. (1993) A Perspective on Application of Bootstrap Methods in Econometrics, in *Handbook of Statistics*, **11**, North-Holland.

Juan, S. (1999) Les modélisations économétriques d'estimation de coût dans l'industrie automobile : l'apport des techniques de bootstrap. *Thèse*, ENSPM-Université de Bourgogne.

Lantz, F. (1983) *Mise en œuvre de la régression linéaire : calcul et stabilité de la matrice (X'X) inverse*, Les Cahiers du 3e cycle économétrie, université de Paris X, 34-47.

McCullough, B.D. (1996) Estimating Forecast Intervals When the Exogenous Variable is Stochastic. *Journal of Forecasting*, **15**, 293-304.

McCullough, B.D. et Vinod, H. (1993) Implementing the Single Bootstrap: Some Computational Considerations. *Computational Economics*, **6**, 1-15.

Seaks, T. (1972) Computer Algorithms—Syminv: An Algorithm for the Inversion of a Positive Definite Matrix by the Cholesky Decomposition. *Econometrica*, **40**, 5.

Simar, L. (1992) Estimating Efficiencies from Frontier Models with Panel Data: A Comparison of Parametric, Non Parametric and Semi-Parametric Methods with Bootstrapping. *The Journal of Productivity Analysis*, **3**, 171-203.

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

Veall, M.R. (1998) Applications of the Bootstrap, in *Handbook of Applied Economic Statistics*, **155**, Aman U., Giles D.E.A.

Verboven, F. (1996) International Price Discrimination in the European Car Market. *Rand Journal of Economics*, **27**, 240-268.

Vinod, H. (1993) Bootstrap Methods: Applications in Econometrics, in *Handbook of Statistics*, **11**, North-Holland, 629-661.

Manuscrit final reçu en juin 2001

ANNEXE

PRINCIPE DU BOOTSTRAP ET APPLICATION AU MODÈLE DE RÉGRESSION

Le principe

Le principe du bootstrap consiste, en répétant un grand nombre de fois le ré-échantillonnage dans les données d'origine, à construire la fonction de répartition empirique bootstrap d'une statistique d'intérêt. Cette dernière approche alors de manière satisfaisante la vraie distribution de la statistique qui, elle, est inconnue. Le processus de construction de la fonction de répartition empirique bootstrap d'un estimateur est détaillé ci-dessous.

Soit un échantillon *i.i.d.*, $\{y_i\}_{i=1}^n$ d'une variable aléatoire y de loi F inconnue. Nous cherchons à déterminer la loi $\hat{\theta}(F)$ d'un estimateur $\hat{\theta}$ d'un paramètre θ de F. L'objectif poursuivi est donc de construire une table statistique[1] de valeurs approchées de $\hat{\theta}(F)$. Pour ce faire, nous disposons des n observations de l'échantillon, donc de la fonction de répartition empirique $\hat{F}_n$.

En théorie, il est possible de construire une table de $\hat{\theta}(\hat{F}_n)$: T, conditionnelle aux $\{y_i\}_{i=1}^n$, en calculant $\hat{\theta}$ sur chacun des n échantillons tirés avec remise dans $\{y_i\}_{i=1}^n$. Cependant, il existe n^n tels échantillons et cette procédure ne peut être mise en œuvre que lorsque n est très petit. En pratique, nous allons tirer B échantillons ($b = 1, \ldots, B$), nommés *échantillons bootstrap*, pour construire une table extraite de T. Sur chacun des échantillons bootstrap, nous calculons la valeur $\hat{\theta}^*(b)$ de la statistique $\hat{\theta}$. La table bootstrap de $\hat{\theta}^*$ est donc une sous-table de T, conditionnelle à l'échantillon de données. Cette notion de dépendance aux données est importante pour la compréhension du processus bootstrap. La table bootstrap ne s'applique, en effet, que pour l'échantillon initial. Pour un nouvel échantillon, il est donc nécessaire de construire une nouvelle table bootstrap, propre à cet échantillon.

D'après le théorème de Glivenko-Cantelli (G-C), pour un échantillon de variables aléatoires *i.i.d.* de loi F inconnue, lorsque la taille de l'échantillon n tend vers l'infini, la fonction de répartition empirique de l'échantillon converge uniformément presque sûrement vers la loi F. Ainsi, la table bootstrap de $\hat{\theta}^*$, conditionnelle aux $\{y_i\}_{i=1}^n$, fournit une bonne approximation de la loi $\hat{\theta}(F)$.

Les méthodes bootstrap sur les modèles de régression

Le modèle de régression linéaire multiple est noté :

$$Y = X\beta + u \qquad \text{(A-1)}$$

où Y est un vecteur (n, 1), X une matrice (n, p), β le vecteur des coefficients à estimer (p, 1) et u le vecteur des erreurs aléatoires (n, 1). Un rang d'observations i ($i = 1, \ldots, n$) de la matrice X, correspondant à une ligne, est noté X_i (1, p). Les paramètres estimés par la méthode des moindres carrés ordinaires (MCO) $\hat{\beta}$ et les résidus $\hat{u}$ sont définis comme :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

et :

$$\hat{u} = Y - X\beta$$

Le bootstrap des résidus

Le modèle théorique bootstrap est le suivant :

$$Y^* = X\hat{\beta} + u^* \qquad \text{(A-2)}$$

où $\hat{\beta}$ est l'estimateur MCO et u^* est un terme aléatoire issu des résidus $\hat{u}$ de la régression initiale, dont nous décrivons la construction ci-dessous.

L'application de la procédure bootstrap consiste à répéter B fois les étapes suivantes :

– à chaque itération b ($b = 1, \ldots, B$), un échantillon $\{y_i^*\}_{i=1}^n$, de dimension (n, 1), est constitué à partir du modèle bootstrap (A-2). Nous disposons alors d'un nouveau couple (Y^*, X) à partir duquel nous pouvons réaliser une estimation des paramètres de la régression ;

– les résidus MCO étant plus petits que les erreurs qu'ils estiment, une transformation est nécessaire pour élaborer le terme aléatoire du modèle théorique bootstrap. Ainsi, à la suite de Freedman (1981), ce dernier est construit avec les résidus transformés suivants ($\hat{u}_i$ est divisé par un facteur proportionnel à la racine de sa variance), qui sont de même norme que les termes erreurs u_i :

$$\tilde{u}_i = \frac{\hat{u}_i}{\sqrt{(1 - h_i)}} - \frac{1}{n}\sum_{s=1}^{n} \frac{\hat{u}_s}{\sqrt{(1 - h_s)}}$$

Le modèle théorique bootstrap est donc le suivant :

$$y_i^*(b) = X_i\hat{\beta} + \tilde{u}_i^*(b), \quad i = 1, \ldots, n \qquad \text{(A-3)}$$

où $\tilde{u}_i^*(b)$ est ré-échantillonné à partir des $\tilde{u}_i$. Les nouvelles variables dépendantes « bootstrappées » $y_i^*(b)$ sont donc construites à partir des valeurs calculées $\hat{y}_i$ et des $\tilde{u}_i^*(b)$:

$$y_i^*(b) = \hat{y}_i + \tilde{u}_i^*(b)$$

– la procédure d'estimation par MCO est appliquée sur le modèle de régression (A-3) afin d'obtenir l'estimateur bootstrap. Pour le b-ième échantillon, ce dernier s'écrit :

$$\hat{\beta}^*(b) = (X^T X)^{-1} X^T Y^*(b) \qquad \text{(A-4)}$$

(1) Par table statistique, nous désignons une version empirique (sur les B réplications bootstrap) de la distribution d'échantillonnage de $\hat{\theta}$.

Les première et deuxième étapes sont répétées B fois ($b = 1, \ldots, B$). Les fonctions de répartition empirique bootstrap de $\hat{\beta}^*$ et des statistiques issues de $\hat{\beta}^*$ sont ensuite construites.

Lors de la mise en œuvre d'une procédure de bootstrap sur les résidus, les variables explicatives X sont considérées comme fixées. Cette procédure est donc valide si les X sont réellement fixées et si les erreurs vérifient les hypothèses classiques des MCO[2]. Par conséquent, elle n'est pas correcte si ces dernières sont hétéroscédastiques. En appliquant la procédure de bootstrap des résidus pour construire le modèle théorique bootstrap, différents termes erreurs sont associés à différentes variables explicatives. Ainsi, la procédure bootstrap de ré-échantillonnage des résidus n'est pas en mesure de respecter cette relation. Dans de tels cas, le modèle théorique bootstrap est construit suivant une procédure différente, nommée *bootstrap par paires*.

Le bootstrap par paires

Cette seconde approche bootstrap des modèles de régression consiste à ré-échantillonner directement dans les données d'origine, à partir des paires (y_i, X_i). Notons cependant que le retirage simultané de (y_i, X_i) introduit une corrélation entre les régresseurs et les erreurs du processus générateur de données (PGD) bootstrap. Ainsi, le bootstrap par paires, sous cette forme[3] simple, ne respecte pas l'hypothèse d'exogénéité des régresseurs dans le PGD bootstrap.

L'application de la procédure bootstrap par paires consiste à répéter B fois les étapes suivantes :

– à chaque itération b ($b = 1, \ldots, B$), le vecteur Y^* et la matrice des variables explicatives X^* sont construits, en effectuant n tirages aléatoires avec remise[4] de paires (y_i, X_i) dans l'échantillon d'origine. Ainsi, si le terme erreur u_i associé à X_i a une grande variance, la relation sera préservée dans l'échantillon bootstrap ;
– une estimation par MCO des coefficients du modèle de régression bootstrap est ensuite réalisée :

$$\hat{\beta}^*(b) = (X^{*T}(b)X^*(b))^{-1}X^{*T}(b)Y^*(b) \qquad \text{(A-5)}$$

Notons, qu'à la différence de la procédure bootstrap des résidus, la matrice des variables explicatives $X^*(b)$ est différente à chaque itération b.

Les B réplications $\hat{\beta}^*$ fournissent alors la fonction de répartition empirique bootstrap. Ainsi, les B réplications bootstrap indépendantes, obtenues suivant les procédures de bootstrap présentées ci-dessus, fournissent un échantillon aléatoire des $\hat{\beta}^*$ qui est utilisé pour estimer la distribution bootstrap de $\hat{\beta}$. Cette dernière permet alors la construction des intervalles de confiance bootstrap des paramètres du modèle de régression.

(2) Les erreurs ont une espérance mathématique nulle, sont homoscédastiques et non-autocorrélées.

(3) Le *wild bootstrap* est une méthode adaptée en présence d'hétéroscédasticité, qui respecte l'hypothèse d'exogénéité des régresseurs.

(4) Notons que Freedman (1981) envisage des échantillons bootstrap de taille m différente de n.