Université de Montréal

Inférence suite à la sélection d'un modèle en régression linéaire multiple

par

Didier Garriguet

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

août 1999

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Inférence suite à la sélection d'un modèle en régression linéaire multiple

présenté par

Didier Garriguet

a été évalué par un jury composé des personnes suivantes :

Urs Maag

(président-rapporteur)

Christian Léger

(directeur de recherche)

Martin Bilodeau

(membre du jury)

Mémoire accepté le :

Novembre 1999

SOMMAIRE

Lorsque nous désirons effectuer une régression linéaire, nous devons fréquemment choisir certaines variables explicatives parmi un grand nombre afin d'établir la relation de dépendance avec la variable expliquée. Cette étape préliminaire à l'inférence statistique est la sélection de modèle.

Suite à cette sélection, la façon classique de faire de l'inférence consiste à oublier les variables non sélectionnées et à ne considérer que les variables choisies afin d'estimer les coefficients et de construire les intervalles de confiance du modèle de régression. Cette façon de procéder ne tient pas compte du côté aléatoire de la sélection qui est basée sur des observations. En particulier, nous fixons implicitement l'intervalle de confiance des variables exclues du modèle comme étant l'ensemble $\{0\}$. Les pourcentages de couverture de chacun des coefficients peuvent être alors différents de la valeur prescrite.

Dans ce mémoire, nous étudions trois méthodes de rééchantillonnage afin de palier à ce problème. Il s'agit du rééchantillonnage des résidus, des paires d'observations et du sous-échantillonnage. La première méthode avait été considérée par Carignan (1996). Toutes ces méthodes tiennent compte du côté aléatoire de la sélection. En particulier, les intervalles de confiance des variables exclues du modèle peuvent être différents de $\{0\}$. Plusieurs simulations sont effectuées afin de comparer ces méthodes entre elles et entre la méthode dite classique.

Nous concluons que le rééchantillonnage donne des pourcentages de couverture pouvant être beaucoup plus élevés que les pourcentages de couverture de la

méthode classique. Parmi les méthodes de rééchantillonnage, nous concluons que le rééchantillonnage des paires d'observations est la meilleure méthode en terme de pourcentages de couverture pour une méthode de sélection convergente quelque soit la taille du jeu de données. Pour une méthode de sélection non convergente, nous concluons que pour un petit jeu de données, le rééchantillonnage des résidus est préférable. Pour un jeu de données plus grand, l'économie du temps de calcul nous fait pencher davantage du côté du sous-échantillonnage.

REMERCIEMENTS

Un travail de cette envergure ne se fait pas sans la collaboration de plusieurs personnes. Je voudrais profiter de cette occasion pour les remercier. Tout d'abord, je tiens à remercier mon directeur de recherches, Christian Léger, autant pour sa rigueur et ses conseils que pour son soutien financier par l'entremise du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et de *Mathematics of Information Technology and Complex Systems* (MITACS). Des fleurs également pour Miguel Chagnon qui m'a permis, avec le poste de coadministrateur du laboratoire de statistique, d'ajouter à ma formation une dimension informatique qui me fut et me sera profitable autant pour ce mémoire que mes projets futurs. Je tiens à souligner à cet égard le soutien financier de tous les professeurs de statistique qui contribuent à la bonne marche du laboratoire. Finalement, je ne peux terminer sans glisser un mot sur mes compagnons de classe. Leur contribution à ce mémoire fut négligeable, mais leur présence fit de ces deux années une expérience des plus plaisante.

Table des matières

Table des figures

Liste des tableaux

INTRODUCTION

Lorsque nous voulons établir la relation de dépendance entre une variable de réponse et une ou plusieurs variables explicatives, nous pouvons utiliser la régression linéaire multiple. Nous estimons les coefficients associés à ces variables grâce à nos observations. Certaines variables peuvent être exclues du modèle sans perte d'information et de "précision". Au contraire, la "précision" de ces coefficients dépend, entre autre, du nombre de variables indépendantes contenues dans le modèle. La sélection de certaines variables et l'exclusion des autres variables est une étape préliminaire importante à notre modélisation permettant d'augmenter la précision de l'estimation des coefficients et de diminuer les coûts des analyses subséquentes. Cependant, l'élimination d'une variable indépendante importante peut entraîner une estimation biaisée des coefficients de régression et de la variance des erreurs.

La façon "classique" de faire de l'inférence suite à la sélection d'un modèle est d'utiliser les variables indépendantes choisies afin d'estimer les coefficients de la régression linéaire et de construire les intervalles de confiance comme s'il n'y avait jamais eu de sélection à l'aide des données. Ainsi, les variables non sélectionnées ont implicitement pour intervalle de confiance l'ensemble nul. Cette façon de procéder ne tient pas compte du caractère aléatoire de la sélection de modèle basé sur les observations. L'inférence statistique classique peut alors entraîner des pourcentages de couverture réels très différents de la valeur prescrite. Pour remédier à la situation, Carignan (1996), a proposé d'utiliser la méthode

du rééchantillonnage. Sa façon d'appliquer le rééchantillonnage, en sélectionnant un premier modèle duquel sera généré un grand nombre d'observations bootstrap, que nous appellerons le rééchantillonnage des résidus, permet effectivement de corriger la situation. Les simulations qu'il a effectuées en utilisant cette méthode lui ont donné des pourcentages de couverture supérieurs à ceux obtenus par inférence classique.

Il existe d'autres façons d'appliquer le rééchantillonnage. Une première consiste à rééchantillonner avec remise les paires d'observations en sélectionnant un échantillon de taille égale au nombre d'observations. Une seconde consiste à sélectionner sans remise un nombre plus restreint d'observations.

Nous voulons déterminer laquelle de ces méthodes, que nous appellerons le rééchantillonnage des paires d'observations et le sous-échantillonnage, est la meilleure selon le nombre d'observations, petit ou grand, utilisé. Nous désirons également regarder l'influence de la méthode de sélection utilisée. Cette dernière peut ou non converger vers le vrai modèle de régression. Nous considérerons les méthodes convergentes de Ducharme et de BIC, toutes deux basées sur la détermination du "meilleur" sous-ensemble de variables parmi tous les modèles possibles. Nous considérerons également les méthodes non convergentes du C_p de Mallows, également basée sur la sélection du "meilleur" sous-ensemble, et les méthodes de sélection séquentielles d'addition et de retrait par étapes.

Les simulations conduites à partir de ces deux nouvelles méthodes nous ont permis de conclure que, selon la convergence de la méthode de sélection choisie et le nombre d'observations utilisés, le rééchantillonnage des paires ou le sous-échantillonnage étaient préférables au rééchantillonnage des résidus.

Au chapitre 1, nous présenterons les modèles de régression linéaire et les méthodes de sélection que nous utiliserons tout au long de notre étude. Nous considérerons un modèle où les variables explicatives sont fixes et un modèle où elles sont aléatoires. Nous exposerons les méthodes de sélection selon qu'elles convergent ou non et qu'elles soient construites selon une procédure séquentielle ou en choisissant le meilleur sous-ensemble de variables indépendantes. Au chapitre 2, nous développerons les applications du rééchantillonnage des paires d'observations, du rééchantillonnage des résidus et du sous-échantillonnage pour le calcul des intervalles de confiance suite à la sélection de modèle. Finalement, au chapitre 3, nous présenterons les résultats de nos différentes simulations utilisant les trois méthodes de rééchantillonnage.

Chapitre 1

MÉTHODES DE SÉLECTION DE MODÈLE EN RÉGRESSION LINÉAIRE

La régression linéaire permet d'établir la relation de dépendance linéaire entre une variable de réponse et une ou plusieurs variables indépendantes. Un sous-ensemble de ces variables peut être suffisant pour expliquer cette dépendance. Nous utilisons différentes méthodes de sélection afin de choisir le sous-ensemble optimal.

Nous commencerons par introduire le principal estimateur en régression linéaire classique et le cas particulier où nous travaillons avec le "vrai" modèle. Par la suite, nous définirons la sélection de modèle et la convergence de la sélection. Nous terminerons ce chapitre en présentant sommairement les méthodes de sélection séquentielle et du meilleur sous-ensemble selon qu'elles convergent ou non.

1.1. RÉGRESSION LINÉAIRE CLASSIQUE ET VRAI MODÈLE DE RÉGRESSION

Considérons le vecteur $\underset{\sim}{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip-1})'$ de variables indépendantes et la variable de réponse correspondante y_i pour $i = 1, 2, \ldots, n$. Formons la matrice X dont chacune des n lignes est un vecteur $\underset{\sim}{x}_i'$. Formons également le vecteur $\underset{\sim}{y}$,

dont chacune des n composantes est la variable y_i. Le couple $(\underset{\sim}{x}_i', y_i)$ sera la $i^{\text{ième}}$ ligne de la matrice formée de X et $\underset{\sim}{y}$.

Une première façon de modéliser la relation entre la variable de réponse y et le vecteur de variables indépendantes $\underset{\sim}{x}$ est

Modèle 1.1.1.

$$E(y_i \mid \underset{\sim}{x}_i) = \underset{\sim}{x}_i'\underset{\sim}{\beta}, \qquad i = 1, \ldots, n,$$

où

- $\underset{\sim}{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})$ *est un vecteur de paramètres inconnus;*
- $(\underset{\sim}{x}_i', y_i)$ *sont i.i.d. de loi multivariée* F *de moyenne* $\underset{\sim}{0}$ *et de matrice de variance* Σ.

Une autre façon de représenter cette relation est de considérer la matrice X comme étant fixe. Dans ce cas, on obtient le modèle

Modèle 1.1.2.

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

où

- X *est une matrice de constantes de plein rang* $p \leq n$;
- $\underset{\sim}{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})$ *est un vecteur de paramètres inconnus;*
- $\underset{\sim}{\epsilon}$ *est un vecteur d'erreurs i.i.d. de distribution* F *de moyenne 0 et de variance finie* σ^2.

Afin d'estimer le vecteur $\underset{\sim}{\beta}$, nous utiliserons la méthode des moindres carrés qui selon le théorème de Gauss-Markov (Neter, Kutner, Nachtsheim et Wasserman, 1996, p.20) nous donnera un estimateur, $\underset{\sim}{\hat{\beta}}$, sans biais et à variance minimale parmi tous les estimateurs linéaires sans biais.

L'estimateur des moindres carrés du vecteur des paramètres est

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1}X'\underset{\sim}{y} \tag{1.1.1}$$

et sa variance est

$$Var(\underset{\sim}{\hat{\beta}}) = \sigma^2(X'X)^{-1}. \tag{1.1.2}$$

Nous estimerons cette variance par:

$$\widehat{Var}(\underset{\sim}{\hat{\beta}}) = \hat{\sigma}^2(X'X)^{-1}$$

$$\text{où } \hat{\sigma}^2 = (\underset{\sim}{y} - X\underset{\sim}{\hat{\beta}})'(\underset{\sim}{y} - X\underset{\sim}{\hat{\beta}})/(n-p). \tag{1.1.3}$$

Si on suppose la normalité des erreurs ou la multinormalité de $(\underset{\sim}{x}, y)$, Sampson (1974) démontre qu'en utilisant une réalisation de n vecteurs $(\underset{\sim}{x}, y)$, les estimés obtenus par maximum de vraisemblance sont les mêmes pour le modèle 1.1.1 et 1.1.2. Par conséquent, nous aurons également les mêmes estimés par la méthode des moindres carrés.

Il est toujours possible de diviser la matrice X en deux matrices de plein rang X_1 et X_2 respectivement de taille $n \times p_1$ et $n \times p_2$ où $p_1 + p_2 = p$. Le modèle 1.1.2 devient alors:

Modèle 1.1.3.

$$\underset{\sim}{y} = (X_1 \quad X_2)\begin{pmatrix} \underset{\sim}{\beta_1} \\ \underset{\sim}{\beta_2} \end{pmatrix} + \underset{\sim}{\epsilon}$$

où $\underset{\sim}{\beta_1}$ *et* $\underset{\sim}{\beta_2}$ *sont des vecteurs de paramètres respectivement de taille* $p_1 \times 1$ *et* $p_2 \times 1$ *et* $\underset{\sim}{\epsilon}$ *est un vecteur d'erreurs i.i.d.* F *de moyenne 0 et de variance finie* σ^2.

De la même façon, séparer le vecteur $\underset{\sim}{x}$ en 2 vecteurs $\underset{\sim}{x}_1$ et $\underset{\sim}{x}_2$ de taille $p_1 \times 1$ et $p_2 \times 1$ nous donnera le modèle

Modèle 1.1.4.

$$E(y_i \mid \underset{\sim}{x}_{i1}, \underset{\sim}{x}_{i2}) = \underset{\sim}{x}'_{i1}\underset{\sim}{\beta}_1 + \underset{\sim}{x}'_{i2}\underset{\sim}{\beta}_2, \qquad i = 1, \ldots, n,$$

où $\underset{\sim}{\beta}_1$ *et* $\underset{\sim}{\beta}_2$ *sont des vecteurs de paramètres de taille* $p_1 \times 1$ *et* $p_2 \times 1$ *et* $E(y_i \mid \underset{\sim}{x}_{i1}, \underset{\sim}{x}_{i2})$ *représente l'espérance conditionnelle de* y_i *suite à la réalisation des 2 vecteurs* $\underset{\sim}{x}_{i1}$ *et* $\underset{\sim}{x}_{i2}$.

Cette division des variables explicatives en deux sous-ensembles est particulièrement utile lorsque l'on veut considérer différents sous-modèles. Ces sous-modèles veulent expliquer la relation linéaire entre X et $\underset{\sim}{y}$ en utilisant certaines variables explicatives et en excluant les autres. En d'autres termes, nous fixons implicitement la valeur du coefficient des variables exclues à 0. En permutant les colonnes de la matrice X nous pouvons regrouper tous ces coefficients en un seul vecteur, $\underset{\sim}{\beta}_2$. Nous fixerons également de façon implicite la variance des coefficients des variables exclues à 0. Ensuite, nous calculerons explicitement les estimés des coefficients non nuls, $\underset{\sim}{\beta}_1$, et nous estimerons leur variance en utilisant les moindres carrés. Nous obtiendrons explicitement

$$\underset{\sim}{\hat{\beta}}_1 = (X_1'X_1)^{-1}X_1'\underset{\sim}{y} \tag{1.1.4}$$

et l'estimé de sa variance

$$\widehat{Var}(\underset{\sim}{\hat{\beta}}_1) = \hat{\sigma}_1^2(X_1'X_1)^{-1} \tag{1.1.5}$$

$$\text{où } \hat{\sigma}_1^2 = SSE_1/(n - p_1) \tag{1.1.6}$$

$$\text{avec } SSE_1 = (\underset{\sim}{y} - X_1\underset{\sim}{\hat{\beta}}_1)'(\underset{\sim}{y} - X_1\underset{\sim}{\hat{\beta}}_1). \tag{1.1.7}$$

La matrice X_1 est l'ensemble des colonnes de la matrice X correspondant aux variables indépendantes incluses dans le sous-modèle.

Implicitement, nous obtenons également

$$\underset{\sim}{\hat{\beta}}_2 = \underset{\sim}{0}_{p_2 \times 1} \tag{1.1.8}$$

un vecteur de p_2 zéros et l'estimé de sa variance

$$\widehat{Var}(\underset{\sim}{\hat{\beta}}_2) = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}_{p_2 \times p_2} \tag{1.1.9}$$

une matrice de 0 de dimension $p_2 \times p_2$.

Définissons maintenant le "vrai" modèle de régression. Nous allons considérer tous les modèles contenant une constante. La première colonne de la matrice X, x_{i0} pour $i = 1, 2, \ldots, n$, sera un vecteur de 1 de dimension $n \times 1$ et les $p-1$ autres colonnes, les variables explicatives observées. Dans le cas du modèle 1.1.1, nous allons considérer que la première composante de tous les vecteurs $\underset{\sim}{x}$, x_0, est 1. Soient $\Omega = \{0, 1, ..., p-1\}$ l'ensemble contenant les indices de toutes les variables explicatives et $\mathcal{A}$, l'ensemble des 2^{p-1} sous-ensembles de Ω contenant l'indice 0.

Si nous considérons une matrice X fixe divisible en deux matrices, alors nous pouvons définir le "vrai" modèle de la façon suivante:

Modèle 1.1.5. *Soit $A_0 \in \mathcal{A}$ représentant le "vrai" modèle de taille p_0 et $A_1 = \Omega \backslash A_0$, les variables qui ne font pas partie du "vrai" modèle, de taille p_1. Le "vrai" modèle de régression peut alors s'écrire*

$$\underset{\sim}{y} = (X_{A_0} \quad X_{A_1}) \begin{pmatrix} \underset{\sim}{\beta}_{A_0} \\ \underset{\sim}{\beta}_{A_1} \end{pmatrix} + \underset{\sim}{\epsilon}$$

où X_{A_0} et X_{A_1} sont les matrices dont les colonnes correspondent respectivement aux éléments de A_0 et A_1, $\underset{\sim}{\beta}_{A_0} \neq \underset{\sim}{0}$, $\underset{\sim}{\beta}_{A_1} = \underset{\sim}{0}$ et $\underset{\sim}{\epsilon}$ est un vecteur d'erreurs i.i.d. F de moyenne 0 et de variance finie σ^2.

Nous pouvons également considérer comme précédemment le modèle aléatoire 1.1.1. Dans ce cas, le "vrai" modèle peut alors s'écrire

Modèle 1.1.6.

$$E(y_i \mid \underset{\sim}{x}_{iA_0}, \underset{\sim}{x}_{iA_1}) = \underset{\sim}{x}'_{iA_0}\underset{\sim}{\beta}_{A_0} + \underset{\sim}{x}'_{iA_1}\underset{\sim}{\beta}_{A_1}, \qquad i = 1, \ldots, n,$$

où $\underset{\sim}{x}_{iA_0}$ et $\underset{\sim}{x}_{iA_1}$ sont les vecteurs dont les éléments correspondent à ceux de A_0 et A_1, $\underset{\sim}{\beta}_{A_0} \neq \underset{\sim}{0}$ et $\underset{\sim}{\beta}_{A_1} = \underset{\sim}{0}$.

Notons qu'il sera toujours possible de permuter *a posteriori* les colonnes de la matrice X afin que les p_0 premières colonnes correspondent aux indices A_0. L'estimation du vecteur des paramètres et de la matrice de variance s'effectue alors par moindres carrés en utilisant les équations (1.1.4) à (1.1.9) avec $X_1 = X_{A_0}$, $X_2 = X_{A_1}$, $\underset{\sim}{\beta}_1 = \underset{\sim}{\beta}_{A_0}$ et $\underset{\sim}{\beta}_2 = \underset{\sim}{\beta}_{A_1}$ où $\underset{\sim}{\beta}_1 \neq \underset{\sim}{0}$ et $\underset{\sim}{\beta}_2 = \underset{\sim}{0}$.

1.2. SÉLECTION DE MODÈLES

Plusieurs variables explicatives peuvent être utilisées en régression linéaire pour modéliser une variable dépendante. Idéalement, nous désirons obtenir un modèle contenant chacune des variables appartenant au "vrai" modèle et uniquement celles-ci. Faisons l'hypothèse que toutes ces variables font partie du choix des variables qui s'offrent à nous. Plusieurs techniques existent afin d'éliminer les variables superflues. Ces techniques, que nous verrons plus loin, se divisent en deux grandes catégories: les méthodes de sélection séquentielle et les méthodes de sélection du meilleur sous-ensemble. Miller (1990) et Thompson (1978a, b) font un bon survol sur l'ensemble de ces techniques. Suite à la sélection d'un modèle,

nous serons à même de déterminer si ce modèle est le "vrai", s'il est biaisé ou trop grand.

Soient $\mathcal{A}_1 = \{A \in \mathcal{A} \mid A \not\supseteq A_0\}$ l'ensemble de tous les modèles ne contenant pas A_0 et $\mathcal{A}_2 = \{A \in \mathcal{A} \mid A \supseteq A_0\}$ l'ensemble de tous les modèles contenant A_0.

Définition 1.2.1. *Un modèle de régression est appelé le "vrai" modèle si ce modèle contient toutes les variables dont le coefficient est non nul et uniquement celles-ci.*

Définition 1.2.2. *Un modèle de régression est appelé modèle biaisé si les indices des coefficients non nuls font partie de l'ensemble* $\mathcal{A}_1$.

Définition 1.2.3. *Un modèle de régression est appelé modèle trop grand si les indices des coefficients non nuls font partie de l'ensemble* $\mathcal{A}_2 - \{A_0\}$.

La qualité d'une méthode de sélection peut être définie par la probabilité qu'a cette méthode de choisir le "vrai" modèle. On dira qu'une méthode de sélection est convergente si elle respecte les deux conditions définies par Nishii (1984). Soit $p_n(A) = P\{\hat{A} = A \mid A \in \mathcal{A}\}$, la probabilité que le modèle sélectionné soit A, où $\hat{A}$ est l'ensemble des indices des variables sélectionnées par une méthode.

Condition 1.2.1.

$$a) \quad \lim_{n\to\infty} np_n(A) = 0 \quad \textit{pour } A \in \mathcal{A}_1$$

$$b) \quad \lim_{n\to\infty} p_n(A) = 0 \quad \textit{pour } A \in \mathcal{A}_2 - \{A_0\}$$

En d'autres termes, la probabilité de sélectionner un modèle autre que le vrai modèle est asymptotiquement nulle. De plus, la condition (a) implique que la probabilité de sélectionner un modèle biaisé convergera plus rapidement vers 0 que celle de choisir un modèle trop grand.

1.3. MÉTHODES DE SÉLECTION SÉQUENTIELLE

Ces méthodes sont largement utilisées en pratique. Nous considérerons les méthodes d'addition par étapes, de retrait par étapes et pas-à-pas. Simples et rapides, elles peuvent donner d'excellents résultats. Elles sont cependant limitées. Berk (1978) comparent les méthodes d'addition par étapes et de retrait par étapes de façon théorique et en utilisant des jeux de données. Il conclut que la méthode d'addition par étapes fait particulièrement bien pour des modèles contenant peu de variables et la méthode de retrait par étapes pour des modèles contenant beaucoup de variables. Cependant, ces méthodes ne sélectionneront peut-être pas des modèles où deux variables prises séparément n'ont pas d'effet, mais qui en ont un si elles sont prises conjointement.

Méthode d'addition par étapes

La méthode d'addition par étape consiste à choisir le meilleur modèle d'une seule variable, c'est-à-dire la variable ayant la plus forte corrélation avec la variable de réponse y, et de lui ajouter des variables une à la fois jusqu'à ce qu'une règle d'arrêt soit satisfaite. A chaque étape, on teste, à l'aide de la statistique F, si le coefficient de la nouvelle variable est nul sachant que le modèle possède les variables déjà sélectionnées. On ajoute la variable qui possède la plus grande statistique F jusqu'à ce que cette statistique descende sous un niveau fixé à l'avance. Ceci revient à choisir la variable qui, à chaque étape, possède la plus forte corrélation partielle avec le modèle sélectionné à l'étape précédente ou dont l'effet sur l'augmentation de la statistique R^2 est maximal.

On peut facilement voir que cette méthode ne converge pas. La puissance de chaque test augmente avec la taille de l'échantillon nous assurant de sélectionner chacune des variables appartenant au "vrai" modèle. Cependant, dans α pourcent

des cas, une variable ne faisant pas partie du "vrai" modèle sera ajoutée. Par conséquent, la probabilité que toutes les variables à exclure le sont ne peut être nulle. Dès lors, la méthode ne converge pas.

Méthode de retrait par étapes

La méthode de retrait par étapes consiste à débuter avec le modèle complet et à enlever les variables une à la fois en testant, à tour de rôle, que chacun des coefficients est égal à zéro, sachant le modèle de l'étape précédente. On enlève la variable dont le test F du coefficient est le plus faible jusqu'à ce que la statistique soit supérieure à un seuil préalablement choisi.

La méthode de retrait par étapes est basée sur la même statistique que la méthode d'addition par étapes; elle ne convergera pas non plus. Toutes les variables appartenant au "vrai" modèle seront éventuellement sélectionnées, mais la probabilité que toutes les variables n'appartenant pas au "vrai" modèle ne soient pas sélectionnées n'est pas nulle.

Méthode de sélection pas-à-pas

Il existe également une troisième méthode, la sélection pas-à-pas. On débute de la même façon que pour la méthode d'addition par étapes, mais à chaque fois qu'une variable a été ajoutée, nous testons si on peut enlever une variable déjà présente dans le modèle. La sélection se termine lorsque l'on ne peut plus ajouter ou retirer une variable.

Tout comme les deux autres méthodes de sélection séquentielle, la méthode de sélection pas-à-pas ne converge pas. Encore une fois, toutes les variables appartenant au "vrai" modèle finiront par être choisies, mais un ajout erroné, qui survient avec une probabilité α, ne sera pas assurément retiré dans une étape ultérieure.

1.4. MÉTHODES DE SÉLECTION DU MEILLEUR SOUS-ENSEMBLE

Dans cette section nous allons exposer différentes méthodes de sélection du meilleur sous-ensemble. Nous introduirons brièvement le critère généralisé d'erreur de prédiction finale pour ensuite voir les méthodes de sélection du C_p de Mallows, la méthode BIC et celle de Ducharme.

Considérons tout d'abord le critère généralisé d'erreur de prédiction finale introduit par Shibata (1984)

$$C_A(\lambda) = SSE_A + \lambda p_A \hat{\sigma}^2_\Omega, \qquad (\lambda > 1) \tag{1.4.1}$$

où λ est un terme de pénalité pour surajustement et $\hat{\sigma}^2_\Omega$ est l'estimé de la variance calculé à partir du modèle complet estimé par (1.1.3). Le modèle de régression sélectionné, $\hat{A}$ de taille $p_{\hat{A}}$ sera

$$\hat{A} = argmin_{A \in \mathcal{A}}\, C_A(\lambda).$$

En d'autres termes, pour identifier le modèle qui minimise (1.4.1), nous identifions tout d'abord chacun des modèles de taille $p_A, p_A = 2, \ldots, p$, incluant une constante, qui minimise SSE_A. Nous ajustons ensuite chacun de ces estimés avec un facteur de pénalité fonction de la taille du modèle.

Plusieurs options ont été proposées pour le choix de λ. C'est ce choix qui déterminera la convergence ou non de la méthode de sélection.

Considérons tout d'abord $\lambda = 2$. On obtient alors le célèbre C_p de Mallows introduit par ce dernier en 1973.

Nishii (1984) démontre que le critère (1.4.1) respecte la condition 1.2.1.a pour tout choix de λ. Il démontre également que pour le critère du C_p de Mallows, la condition 1.2.1.b n'est pas respectée. La probabilité de sélectionner un modèle trop petit tendera rapidement vers 0. Par conséquent, le C_p de Mallows aura tendance à sélectionner des modèles trop grands.

Quel devrait être le choix de λ afin de s'assurer de la convergence de l'estimateur? Sous certaines conditions, il est possible de démontrer que si $\lambda_n \to \infty$ et que $\lambda_n/n \to 0$ alors $P\{C_{A_0}(\lambda_n) \leq C_A(\lambda_n),\ \forall A\} \to 1$. Par conséquent, la condition 1.2.1.b est respectée. En fait, lorsque λ est fixe (indépendant de n), le critère généralisé d'erreur de prédiction finale ne convergera pas.

Si nous considérons $\lambda_n = \log n$, on obtient une version asymptotiquement équivalente au BIC de Schwartz (1978). Nous utiliserons cependant le critère BIC

$$C_{A_{BIC}} = n \log SSE_A - \log n - p_A + p_A \log n.$$

Puisque $\log n \to \infty$ et que $(\log n)/n \to 0$ lorsque $n \to \infty$, ce critère est convergent.

Ducharme (1997) propose d'adapter le C_p de Mallows afin de respecter les deux conditions sur λ_n. Soit $p_{A_{CPM}}$ le nombre de paramètres du modèle A_{CPM} choisi par la méthode du C_p de Mallows. Ducharme a suggéré d'utiliser $\lambda_n = n^{d_n}$ où $d_n = 0{,}5 + 0{,}5[(p_A - p_{A_{CPM}})/p_A]$. Nous appelons cette méthode basée sur ce critère la méthode de Ducharme.

Dans le cas où λ_n est dépendant de n, la vitesse de convergence de la condition 1.2.1.b augmente (Nishii 1984). Pour une taille échantillonnale fixe, le critère BIC et la méthode de Ducharme auront tendance à sélectionner des modèles trop petits.

Chapitre 2

RÉÉCHANTILLONNAGE ET SOUS-ÉCHANTILLONNAGE

Peu importe que nous considérions le modèle 1.1.1 ou 1.1.2, nous serons non seulement intéressés à estimer le paramètre $\underset{\sim}{\beta}$, mais également à faire de l'inférence sur ce paramètre, plus particulièrement calculer des intervalles de confiance. Pour construire ces intervalles de façon exacte nous devons connaître la loi de $F(\underset{\sim}{x}, y)$ dans le cas du modèle 1.1.1 ou la loi de $F(\epsilon)$ dans le cas du modèle 1.1.2. En général, nous faisons l'hypothèse que la distribution des erreurs est normale ou que la loi conjointe de $\underset{\sim}{x}$ et y est multinormale. Cependant, même si nous connaissions la loi de $F(\underset{\sim}{x}, y)$ ou de $F(\epsilon)$, nous ne serions pas en mesure de calculer des intervalles de confiance exacts si nous sélectionnions un modèle avant d'estimer le paramètre $\underset{\sim}{\beta}$. Le rééchantillonnage est une méthode qui nous permet d'estimer la distribution inconnue d'un estimateur. A partir de cette approximation de la distribution, nous serons à même de calculer des intervalles de confiance.

Dans ce chapitre, nous verrons tout d'abord ce que nous entendons par rééchantillonnage et comment nous construisons les intervalles de confiance en utilisant cette méthode. Ensuite, nous verrons comment appliquer cette méthode à la sélection de modèles en régression linéaire.

2.1. PRÉSENTATION DU RÉÉCHANTILLONNAGE

Considérons le cas général où nous avons des observations $X = (X_1, X_2, ..., X_n)$ i.i.d. tirées d'une loi F inconnue. Nous sommes intéressés à construire un intervalle de confiance pour un paramètre $\theta(F)$. Nous sommes en mesure d'estimer F par sa distribution expérimentale $\hat{F}_n$ grâce à nos observations. Nous pouvons également estimer θ par $\hat{\theta} = \theta(\hat{F}_n)$. Pour construire l'intervalle de confiance nous devons connaître la loi centrée ou centrée et réduite de cet estimateur.

Soit $J_n(F)$, la distribution de $\sqrt{n}(\hat{\theta} - \theta(F))$ et $K_n(F)$ la distribution de $\sqrt{n}(\hat{\theta} - \theta(F))/\hat{\sigma}$ avec $\hat{\sigma}$ un estimateur de l'écart type de $\sqrt{n}\hat{\theta}$. Efron (1979) a introduit une méthode permettant d'estimer $J_n(F)$ et $K_n(F)$ (voir également Efron et Tibshirani, 1993).

Considérons tout d'abord le cas de la distribution centrée $J_n(F)$. Lorsque nous la calculons, nous utilisons $\hat{\theta}$ un estimateur de θ basé sur un échantillon i.i.d. de taille n de F. L'idée de Efron consiste à estimer $J_n(F)$ par $J_n(\hat{F}_n)$ en utilisant $\hat{\theta}^* = \theta(\hat{G}_n)$ où $\hat{G}_n$ est la distribution expérimentale d'un échantillon i.i.d. de taille n tiré à partir de la distribution $\hat{F}_n$. Nous obtenons ainsi $J_n(\hat{F}_n)$, la distribution de $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ comme estimateur de $J_n(F)$. Afin d'estimer $K_n(F)$, nous procéderons de la même façon en utilisant $K_n(\hat{F}_n)$, la distribution de $\sqrt{n}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$, où $\hat{\sigma}^*$ est l'estimé de l'écart type de $\sqrt{n}\hat{\theta}^*$, comme estimateur. Cette façon de procéder est appelée le rééchantillonnage ou bootstrap.

Idéalement, nous aimerions calculer les distributions exactes de $J_n(\hat{F}_n)$ et $K_n(\hat{F}_n)$. Pour ce faire, nous devrions sélectionner les n^n échantillons possibles pour lesquels nous calculerions $\hat{\theta}^*$ et $\hat{\sigma}^*$. Il serait cependant souvent trop long de calculer un estimateur $\hat{\theta}^*$ pour chacun des échantillons bootstrap possibles. Nous nous contenterons d'utiliser l'approximation de Monte Carlo en utilisant B échantillons

bootstrap. Les B estimés de $\hat{\theta}^*$ et de $\hat{\sigma}^*$ ainsi obtenus nous permettront de calculer les distributions expérimentales $\hat{J}_n(\hat{F}_n)$ et $\hat{K}_n(\hat{F}_n)$.

Intervalles de confiance bootstrap

Voyons maintenant comment construire les intervalles de confiance bootstrap. Nous considérerons trois sortes d'intervalles de confiance bootstrap: l'intervalle de confiance percentile, pivotal et bootstrap-t.

Soit $J_n(t,F) = P_F\{\sqrt{n}(\hat{\theta}-\theta) \leq t\}$. Un intervalle de confiance exact pour θ de niveau $1-\alpha$ est obtenu en solutionnant l'équation

$$P_F\{J_n^{-1}(\alpha/2,F) \leq \sqrt{n}(\hat{\theta}-\theta) \leq J_n^{-1}(1-\alpha/2,F)\} = 1-\alpha$$

où $J_n^{-1}(\delta,F) = inf\{t \mid J_n(t,F) \geq \delta\}$. L'intervalle de confiance exact, dans ce cas, est donc

$$[\hat{\theta} - J_n^{-1}(1-\alpha/2,F)/\sqrt{n}, \hat{\theta} - J_n^{-1}(\alpha/2,F)/\sqrt{n}].$$

Puisque la distribution F est inconnue, nous utilisons son estimateur $\hat{F}_n$ pour obtenir l'intervalle de confiance pivotal. Soit $J_n(t,\hat{F}_n) = P_{\hat{F}_n}\{\sqrt{n}(\hat{\theta}^*-\hat{\theta}) \leq t\}$.

Définition 2.1.1. *Nous appelons intervalle de confiance pivotal de niveau* $1-\alpha$ *l'intervalle de confiance pour* $\theta(F)$ *donné par*

$$[\hat{\theta} - J_n^{-1}(1-\alpha/2,\hat{F}_n)/\sqrt{n}, \hat{\theta} - J_n^{-1}(\alpha/2,\hat{F}_n)/\sqrt{n}].$$

Nous pouvons également considérer la distribution expérimentale non centrée de $\hat{\theta}^*$. Soit $G_n(t,\hat{F}_n) = P_{\hat{F}_n}\{\hat{\theta}^* \leq t\}$.

Définition 2.1.2. *Nous appelons intervalle de confiance percentile de niveau* $1-\alpha$ *l'intervalle de confiance pour* $\theta(F)$ *donné par*

$$[G_n^{-1}(\alpha/2,\hat{F}_n), G_n^{-1}(1-\alpha/2,\hat{F}_n)].$$

Nous constatons immédiatement que $G_n^{-1}(\alpha, \hat{F}_n) = \hat{\theta} + J_n^{-1}(\alpha, \hat{F}_n)/\sqrt{n}$. Nous pouvons alors définir de façon équivalente à la définition 2.1.2 l'intervalle de confiance percentile de niveau $1 - \alpha$ par

$$[\hat{\theta} + J_n^{-1}(\alpha/2, \hat{F}_n)/\sqrt{n}, \hat{\theta} + J_n^{-1}(1 - \alpha/2, \hat{F}_n)/\sqrt{n}].$$

Nous devons être prudent quant à l'utilisation de cette équivalence. Dans certains cas, en particulier le rééchantillonnage des résidus que nous verrons plus loin, certaines modifications doivent être apportées. Nous utiliserons la seconde définition dans la majorité des cas.

Finalement, l'intervalle de confiance bootstrap-t est obtenu en utilisant la distribution centrée et réduite de $\hat{\theta}$. Soit $K_n(t, \hat{F}_n) = P_{\hat{F}_n}\{\sqrt{n}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^* \leq t\}$ avec $\hat{\sigma}^*$, l'estimé de l'écart type de $\sqrt{n}\hat{\theta}^*$.

Définition 2.1.3. *Nous appelons intervalle de confiance bootstrap-t de niveau* $1 - \alpha$ *l'intervalle de confiance pour* $\theta(F)$ *donné par*

$$[\hat{\theta} - \hat{\sigma}K_n^{-1}(1 - \alpha/2, \hat{F}_n)/\sqrt{n}, \hat{\theta} - \hat{\sigma}K_n^{-1}(\alpha/2, \hat{F}_n)/\sqrt{n}].$$

2.2. PRÉSENTATION DU SOUS-ÉCHANTILLONNAGE

Il existe une autre méthode afin d'estimer les distributions $J_n(F)$ et $K_n(F)$. Celle-ci est basée sur la distribution expérimentale des $N_n = \binom{n}{b}$ sous-échantillons de taille b sélectionnés sans remise parmi les n éléments de l'échantillon X initial. Soient $\mathcal{S}$, l'ensemble de tous ces sous-échantillons, $\hat{G}_{n;b}$ la distribution expérimentale d'un des sous-échantillons, $\hat{\theta}^\circ = \theta(\hat{G}_{n;b})$, l'estimateur de θ basé sur cet échantillon et $\mathcal{I}\{\mathcal{L}\}$, la fonction indicatrice prenant la valeur 1 si la condition $\mathcal{L}$ est respectée et 0 sinon. Nous allons estimer $J_n(F)$ par

$$J_{n;b}(t) = N_n^{-1} \sum_{\mathcal{S}} \mathcal{I}\{\sqrt{b}(\hat{\theta}^\circ - \hat{\theta}) \leq t\} \tag{2.2.1}$$

et $K_n(F)$ par

$$K_{n;b}(t) = N_n^{-1} \sum_{\mathcal{S}} \mathcal{I}\{\sqrt{b}(\hat{\theta}^\circ - \hat{\theta})/\hat{\sigma}^\circ \leq t\} \tag{2.2.2}$$

où $\hat{\sigma}^\circ$ est l'estimateur de l'écart type de $\sqrt{b}\hat{\theta}^\circ$ basé sur un sous-échantillon. Pour simplifier la notation, nous avons sommé implicitement les variables indicatrices $\mathcal{I}$ pour chacune des valeurs prises par $\hat{\theta}^\circ$ dans $\mathcal{S}$.

Politis et Romano (1994) démontrent que si $J_n(F)$ et $K_n(F)$, les distributions de $\sqrt{n}(\hat{\theta} - \theta)$ et $\sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma}$, convergent respectivement vers les lois limites $J(F)$ et $K(F)$, alors les distributions expérimentales $J_{n;b}(t)$ et $K_{n;b}(t)$ convergent respectivement en probabilité vers $J(F)$ et $K(F)$ pourvu que $n \to \infty, b \to \infty$ et $b/n \to 0$ (voir également Politis, Romano et Wolf, 1999).

Idéalement, nous aimerions tirer tous les $\binom{n}{b}$ échantillons possibles, mais puisque ce nombre peut être très grand, nous nous contenterons de tirer B échantillons avec remise parmi l'ensemble $\mathcal{S}$ des échantillons de taille b. A partir des B estimés $\hat{\theta}^\circ$ nous pourrons calculer $\hat{J}_{n;b}(t)$ et $\hat{K}_{n;b}(t)$ les approximations des distributions expérimentales (2.2.1) et (2.2.2).

Intervalles de confiance par sous-échantillonnage

Pour construire les intervalles de confiance pivotal et percentile par la méthode du sous-échantillonnage, nous utiliserons les quantiles des distributions expérimentales (2.2.1) et (2.2.2). Certaines modifications doivent être apportées aux intervalles de confiance bootstrap lorsque nous utilisons le sous-échantillonnage. Puisque nous désirons obtenir un intervalle de confiance pour un échantillon de taille n, nous devons ajuster l'intervalle obtenu en multipliant les quantiles de (2.2.1) par $\sqrt{1/n}$ ou bien en multipliant la statistique $(\hat{\theta}^\circ - \hat{\theta})$ par $\sqrt{b/n}$. En

d'autres termes, nous rétrécissons l'intervalle de confiance obtenu avec b observations car nous savons que la longueur de l'intervalle de confiance est inversement proportionnelle à la racine carrée de la taille de l'échantillon. Les intervalles de confiance pivotal et percentile par sous-échantillonnage sont les suivants.

Définition 2.2.1. *Nous appelons intervalle de confiance par sous-échantillonnage pivotal de niveau* $1-\alpha$ *l'intervalle de confiance pour* $\theta(F)$ *donné par*

$$[\hat{\theta} - J_{n;b}^{-1}(1-\alpha/2)/\sqrt{n}, \hat{\theta} - J_{n;b}^{-1}(\alpha/2)/\sqrt{n}].$$

Définition 2.2.2. *Nous appelons intervalle de confiance par sous-échantillonnage percentile de niveau* $1-\alpha$ *l'intervalle de confiance pour* $\theta(F)$ *donné par*

$$[\hat{\theta} + J_{n;b}^{-1}(\alpha/2)/\sqrt{n}, \hat{\theta} + J_{n;b}^{-1}(1-\alpha/2)/\sqrt{n}].$$

De la même façon, nous devons multiplier les quantiles de (2.2.2) par $1/\sqrt{n}$ afin d'obtenir les intervalles de confiance par sous-échantillonnage bootstrap-t.

Définition 2.2.3. *Nous appelons intervalle de confiance par sous-échantillonnage bootstrap-t de niveau* $1-\alpha$ *l'intervalle de confiance pour* $\theta(F)$ *donné par*

$$[\hat{\theta} - \hat{\sigma}K_{n;b}^{-1}(1-\alpha/2)/\sqrt{n}, \hat{\theta} - \hat{\sigma}K_{n;b}^{-1}(\alpha/2)/\sqrt{n}].$$

2.3. Modèle de régression linéaire: rééchantillonnage et sous-échantillonnage

Dans la section précédente, nous avons vu le fonctionnement du rééchantillonnage pour le cas général. Dans cette section, nous verrons le cas particulier où θ est un élément du vecteur de paramètres $\underset{\sim}{\beta}$ et $\hat{\theta}$ est l'estimateur par la méthode des moindres carrés suite à la sélection d'un modèle.

Nous débuterons par construire, à titre comparatif, les intervalles de confiance classiques pour β_i, un élément de $\underset{\sim}{\beta}$, avec et sans sélection de modèles. Ensuite, nous verrons trois méthodes de rééchantillonnage: deux méthodes pour le modèle 1.1.1 et une méthode pour le modèle 1.1.2. Pour chacune de ces méthodes, nous verrons comment obtenir des estimations des distributions $J_n(F)$ et $K_n(F)$ et comment construire les différents intervalles de confiance bootstrap lorsqu'un modèle a été précédemment choisi par une méthode de sélection S_e.

2.3.1. Intervalles de confiance classiques

Supposons que $(\underset{\sim}{x}, y)$ suit une loi multinormale si nos observations proviennent du modèle 1.1.1 ou que ϵ suit une loi normale si nos observations proviennent du modèle 1.1.2. Supposons également que nous n'effectuons aucune sélection de modèle. Un intervalle de confiance exact pour β_i, $i = 0, \ldots, p-1$, au niveau de confiance $1-\alpha$ est donné par

$$\left[\hat{\beta}_i - \sqrt{\widehat{Var}(\hat{\beta}_i)}t_{n-p}^{1-\alpha/2}, \hat{\beta}_i - \sqrt{\widehat{Var}(\hat{\beta}_i)}t_{n-p}^{\alpha/2}\right], \tag{2.3.1}$$

où t_{n-p}^{δ} est le δ-ième quantile de la loi de Student à $n-p$ degrés de liberté, $\hat{\beta}_i$ est le $i^{\text{ième}}$ élément de $\underset{\sim}{\hat{\beta}}$ obtenu par (1.1.1) et $\widehat{Var}(\hat{\beta}_i) = [\widehat{Var}(\underset{\sim}{\hat{\beta}})]_{ii}$ est le $i^{\text{ième}}$ élément de la diagonale de la matrice $\widehat{Var}(\underset{\sim}{\hat{\beta}})$ obtenu par (1.1.2).

Si un modèle a été sélectionné par une méthode S_e, nous estimons $\underset{\sim}{\hat{\beta}}_{\hat{A}}$ par (1.1.4) et (1.1.8) et $\widehat{Var}(\underset{\sim}{\hat{\beta}}_{\hat{A}})$ par (1.1.5) et (1.1.9) en utilisant la matrice X_1 constituée des colonnes correspondant aux indices $\hat{A}$ des variables choisies. La construction des intervalles de confiance se fera en faisant l'hypothèse que le modèle sélectionné est toujours le vrai modèle. Nous utiliserons l'intervalle de confiance (2.3.1) de niveau $1-\alpha$ avec $p = p_{\hat{A}}$ pour les variables sélectionnées, $\underset{\sim}{\hat{\beta}}_i = \underset{\sim}{\hat{\beta}}_{i\hat{A}}$ et l'ensemble $\{0\}$ comme intervalle pour les variables non sélectionnées.

Notez que l'intervalle de confiance ne sera qu'approximatif malgré l'hypothèse de normalité car nous savons que l'hypothèse qui dit que le modèle sélectionné est toujours le vrai modèle est fausse.

2.3.2. Rééchantillonnage des paires d'observations

Supposons que nos observations proviennent du modèle 1.1.1. Nous cherchons à estimer, pour $i = 0, \ldots, p-1$,

$$J_n(t, \beta_i, F, \hat{A}) = P_F\{\sqrt{n}(\hat{\beta}_{i\hat{A}} - \beta_i) \leq t\}, \tag{2.3.2}$$

la distribution centrée de chacun des éléments du vecteur $\hat{\underset{\sim}{\beta}}_{\hat{A}}$, l'estimation du vecteur de paramètres obtenus après la sélection d'un modèle par une méthode S_e, pour être en mesure de construire les intervalles de confiance bootstrap. Nous voulons donc estimer (2.3.2), pour $i = 0, \ldots, p-1$, par

$$J_n(t, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A}) = P_{\hat{F}_n}\{\sqrt{n}(\hat{\beta}^*_{i\hat{A}^*} - \hat{\beta}_{i\hat{A}}) \leq t\}, \tag{2.3.3}$$

avec $\hat{A}^*$, le modèle sélectionné à partir des données bootstrap, et $\hat{\beta}^*_{i\hat{A}^*}$, le $i^{\text{ème}}$ élément du vecteur bootstrap $\hat{\underset{\sim}{\beta}}^*_{\hat{A}^*}$ obtenu suite à la sélection de ce modèle.

Pour ce faire, nous procédons de la façon suivante. Soit $Z = (X, \underset{\sim}{y})$ la matrice formée de la matrice des observations X et du vecteur $\underset{\sim}{y}$. Soit $\underset{\sim}{z}_i$, la $i^{\text{ème}}$ ligne de cette matrice. Nous sélectionnons tout d'abord un modèle $\hat{A}$ avec une méthode de sélection S_e basé sur la matrice Z, un échantillon i.i.d. de taille n de $F(\underset{\sim}{z})$. Nous pouvons par la suite estimer le "vrai" $\underset{\sim}{\beta}$ par $\hat{\underset{\sim}{\beta}}_{\hat{A}}$ obtenu avec (1.1.4) et (1.1.8) basé sur ce même échantillon.

Lorsque nous estimons $\hat{\underset{\sim}{\beta}}^*_{\hat{A}^*}$ nous utilisons B échantillons, chacun étant formé de n lignes de la matrice Z choisies avec remise, c'est-à-dire d'un échantillon i.i.d. de taille n de $\hat{F}_n(\underset{\sim}{z})$. Nous formerons ainsi la matrice $Z^* = (X^*, \underset{\sim}{y}^*)$ où X^* est la matrice dont les n lignes et $\underset{\sim}{y}^*$ le vecteur dont les n élements correspondent aux n

lignes choisies de la matrice Z. Ceci constituera une approximation de (2.3.3) car la distribution exacte basée sur les n^n échantillons serait trop longue à calculer. A partir de ces échantillons, nous sélectionnons un modèle $\hat{A}^*$. Ensuite, nous évaluons $\underset{\sim}{\hat{\beta}}^*_{\hat{A}^*}$ par (1.1.4) et (1.1.8). Nos B estimés nous permettront de construire la distribution expérimentale de (2.3.3).

Nous pouvons dès lors construire les intervalles de confiance pivotal et percentile en utilisant les quantiles $\alpha/2$ et $1-\alpha/2$ de $J_n(t, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A})$ que nous dénoterons $J_n^{-1}(\delta, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A})$ pour le quantile δ.

Définition 2.3.1. *Nous appelons intervalle de confiance pivotal de niveau* $1-\alpha$ *par la méthode de rééchantillonnage des paires l'intervalle de confiance pour* β_i, $i = 0, \ldots, p-1$, *donné par*

$$[\hat{\beta}_{i\hat{A}} - J_n^{-1}(1-\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A})/\sqrt{n}, \hat{\beta}_{i\hat{A}} - J_n^{-1}(\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A})/\sqrt{n}]$$

suite à la sélection du modèle $\hat{A}$ *par une méthode* S_e, *en utilisant la distribution (2.3.3).*

Définition 2.3.2. *Nous appelons intervalle de confiance percentile de niveau* $1-\alpha$ *par la méthode de rééchantillonnage des paires l'intervalle de confiance pour* β_i, $i = 0, \ldots, p-1$, *donné par*

$$[\hat{\beta}_{i\hat{A}} + J_n^{-1}(\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A})/\sqrt{n}, \hat{\beta}_{i\hat{A}} + J_n^{-1}(1-\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A})/\sqrt{n}]$$

suite à la sélection du modèle $\hat{A}$ *par une méthode* S_e, *en utilisant la distribution (2.3.3).*

Comme nous l'avions vu dans la section 2.1, nous pourrions également utiliser la distribution centrée et réduite de chacun des éléments de $\underset{\sim}{\hat{\beta}}_{\hat{A}}$. Appelons, pour $i = 0, \ldots, p-1$,

$$K_n(t, \beta_i, F, \hat{A}) = P_F\{\sqrt{n}(\hat{\beta}_{i\hat{A}} - \beta_i)/\hat{\sigma}_i \leq t\},$$

cette distribution et sa version bootstrap, pour $i = 0, \ldots, p-1$,

$$K_n(t, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A}) = P_{\hat{F}_n}\{\sqrt{n}(\hat{\beta}^*_{i\hat{A}^*} - \hat{\beta}_{i\hat{A}})/\hat{\sigma}^*_i \leq t\}.$$

Ici, il y a la difficulté supplémentaire d'estimer $\hat{\sigma}^*_i$, un estimé de l'écart type de $\sqrt{n}\hat{\beta}^*_i$. Si nous utilisons l'estimateur de la méthode des moindres carrés défini en (1.1.5) et (1.1.9) nous risquons d'obtenir une variance nulle et par conséquent un intervalle de confiance de longueur infinie. Ce problème survient lorsque la méthode de sélection utilisée lors du rééchantillonnage ne choisit pas une variable qui avait été préalablement choisie.

Pour remédier à la situation, nous utiliserons uniquement la variabilité due aux erreurs et nous omettrons celle due aux variables sélectionnées dans le modèle. Définissons pour $i = 0, \ldots, p-1$,

$$L_n(t, \beta_i, F, \hat{A}) = P_F\{\sqrt{n}(\hat{\beta}_{i\hat{A}} - \beta_i)/\hat{\sigma}_{\hat{A}} \leq t\}, \tag{2.3.4}$$

$$\text{avec } (\hat{\sigma}_{\hat{A}})^2 = \frac{(\underset{\sim}{y} - X\underset{\sim}{\hat{\beta}}_{\hat{A}})'(\underset{\sim}{y} - X\underset{\sim}{\hat{\beta}}_{\hat{A}})}{n - p_{\hat{A}}} \tag{2.3.5}$$

et sa version bootstrap pour $i = 0, \ldots, p-1$,

$$L_n(t, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A}) = P_{\hat{F}_n}\{\sqrt{n}(\hat{\beta}^*_{i\hat{A}^*} - \hat{\beta}_{i\hat{A}})/\hat{\sigma}^*_{\hat{A}^*} \leq t\}, \tag{2.3.6}$$

$$\text{avec } (\hat{\sigma}^*_{\hat{A}^*})^2 = \frac{(\underset{\sim}{y}^* - X^*\underset{\sim}{\hat{\beta}}^*_{\hat{A}^*})'(\underset{\sim}{y}^* - X^*\underset{\sim}{\hat{\beta}}^*_{\hat{A}^*})}{n - p_{\hat{A}^*}}. \tag{2.3.7}$$

En prenant les quantiles de cette distribution, nous obtenons l'intervalle de confiance bootstrap-t-MSE.

Définition 2.3.3. *Nous appelons intervalle de confiance bootstrap-t-MSE de niveau* $1 - \alpha$ *par le rééchantillonnage des paires l'intervalle de confiance pour* β_i, $i = 0, \ldots, p-1$, *donné par*

$$[\hat{\beta}_{i\hat{A}} - \hat{\sigma}_{\hat{A}} L_n^{-1}(1 - \alpha/2, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A})/\sqrt{n}, \hat{\beta}_{i\hat{A}} - \hat{\sigma}_{\hat{A}} L_n^{-1}(\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{F}_n, \hat{A})/\sqrt{n}].$$

suite à la sélection du modèle $\hat{A}$ par une méthode S_e, en utilisant la distribution (2.3.6).

2.3.3. Rééchantillonnage des résidus

Un des avantages de la méthode précédente est que la sélection du modèle et l'estimation du vecteur de paramètres se font en une seule étape. Le rééchantillonnage des résidus requiert deux étapes d'estimation.

Supposons que les observations proviennent du modèle 1.1.2. Non seulement la distribution des erreurs $F(\epsilon)$ est-elle inconnue, mais les erreurs elles-mêmes le sont. Nous ne pouvons donc estimer la distribution des erreurs avant de connaître celles-ci.

Nous allons utiliser la distribution expérimentale des résidus, $\hat{F}_n(\hat{\epsilon})$, afin d'estimer $F(\epsilon)$. D'où proviennent ces résidus? Nous devons les calculer en estimant le modèle une première fois. Nous pouvons utiliser le modèle complet afin d'estimer $\underset{\sim}{\beta}$. Cependant, puisque nous tentons d'estimer le vrai modèle, nous pouvons également l'estimer en utilisant une méthode de sélection S_b, différente ou pas de la méthode S_e que nous désirons utiliser afin de construire les intervalles de confiance bootstrap.

Soit $\tilde{A}$, le modèle estimé par la méthode de sélection S_b, et $\underset{\sim}{\hat{\beta}}_{\tilde{A}}$, l'estimé du vecteur de paramètres sous ce modèle, tous deux basés sur un échantillon d'erreurs i.i.d. de taille n de $F(\epsilon)$. La première étape de la méthode du rééchantillonnage des résidus consiste à calculer $\underset{\sim}{\hat{\epsilon}}_{\tilde{A}} = \underset{\sim}{y} - X\underset{\sim}{\hat{\beta}}_{\tilde{A}}$, les résidus du modèle obtenu suite à la sélection de celui-ci par une méthode S_b. Dans le cas où nous utilisons le modèle complet, nous pourrons conserver cette notation en considérant $\tilde{A} = \Omega$.

Nous devons nous assurer que les résidus ainsi obtenus soient centrés à 0 (Freedman, 1981). Puisque dans notre cas la constante du modèle de régression est toujours incluse, nous n'aurons pas à centrer les résidus explicitement.

Nous voulons estimer (2.3.2), pour $i = 0, \ldots, p-1$, par

$$J_n(t, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A}) = P_{\hat{F}_{\tilde{A}}}\{\sqrt{n}(\hat{\beta}^*_{i\hat{A}^*} - \hat{\beta}_{i\tilde{A}}) \leq t\}, \tag{2.3.8}$$

la distribution centrée de chaque élément de $\underset{\sim}{\hat{\beta}}^*_{\hat{A}^*}$, mais centrée par rapport à $\underset{\sim}{\hat{\beta}}_{\tilde{A}}$ et non pas $\underset{\sim}{\hat{\beta}}_{\hat{A}}$. Pour ce faire, nous sélectionnerons $\hat{A}^*$, le modèle issu d'une méthode de sélection S_e, et évaluerons $\underset{\sim}{\hat{\beta}}^*_{\hat{A}^*}$ basé sur un échantillon i.i.d. de taille n de $\hat{F}_{\tilde{A}} = \hat{F}_n(\hat{\epsilon}_{\tilde{A}})$ de la façon suivante. Tirons, avec remise, n éléments du vecteur $\underset{\sim}{\hat{\epsilon}}_{\tilde{A}}$. Nous obtenons ainsi le vecteur des erreurs bootstrap $\underset{\sim}{\epsilon}^*$. Ensuite, nous allons construire un modèle bootstrap avec

$$\underset{\sim}{y}^* = X\underset{\sim}{\hat{\beta}}_{\tilde{A}} + \underset{\sim}{\epsilon}^*. \tag{2.3.9}$$

A partir de ce modèle, nous sélectionnons $\hat{A}^*$ et évaluons $\underset{\sim}{\hat{\beta}}^*_{\hat{A}^*}$. Nous aimerions calculer les n^n échantillons possibles afin d'obtenir la distribution exacte de (2.3.8), mais par souci d'économie de temps, nous nous contenterons de B échantillons. La distribution expérimentale de (2.3.8) ainsi obtenue nous permettra de calculer les intervalles de confiance pivotal par rééchantillonnage des résidus.

Définition 2.3.4. *Nous appelons intervalle de confiance pivotal de niveau* $1-\alpha$ *par la méthode de rééchantillonnage des résidus issus du modèle* $\tilde{A}$ *sélectionné par la méthode* S_b*, l'intervalle de confiance pour* β_i*,* $i = 0, \ldots, p-1$ *donné par*

$$[\hat{\beta}_{i\hat{A}} - J_n^{-1}(1-\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}, \hat{\beta}_{i\hat{A}} - J_n^{-1}(\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}]$$

suite à la sélection du modèle $\hat{A}$ *par une méthode* S_e*, en utilisant la distribution (2.3.8).*

Afin de construire les intervalles de confiance percentile, nous utiliserons la distribution expérimentale non centrée de $\hat{\beta}^*_{i\hat{A}^*}$. Soit, pour $i = 0, \ldots, p-1$,

$$G_n(t, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A}) = P_{\hat{F}_{\tilde{A}}}\{\hat{\beta}^*_{i\hat{A}^*} \leq t\} \tag{2.3.10}$$

Définition 2.3.5. *Nous appelons intervalle de confiance percentile de niveau* $1-\alpha$ *par la méthode de rééchantillonnage des résidus issus du modèle* $\tilde{A}$ *sélectionné par la méthode* S_b*, l'intervalle de confiance pour* β_i*,* $i = 0, , \ldots, p-1$ *donné par*

$$[G_n^{-1}(\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A}), G_n^{-1}(1-\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})]$$

suite à la sélection du modèle $\hat{A}$ *par une méthode* S_e*, en utilisant la distribution (2.3.10).*

Pour obtenir les intervalles de confiance percentile en utilisant la distribution (2.3.8), nous devrions utiliser l'intervalle

$$[\hat{\beta}_{i\tilde{A}} + J_n^{-1}(\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}, \hat{\beta}_{i\tilde{A}} + J_n^{-1}(1-\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}]$$

centré en $\hat{\beta}_{i\tilde{A}}$ et non pas l'intervalle

$$[\hat{\beta}_{i\hat{A}} + J_n^{-1}(\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}, \hat{\beta}_{i\hat{A}} + J_n^{-1}(1-\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}]$$

centré en $\hat{\beta}_{i\hat{A}}$.

En effet, par rééchantillonnage des résidus, nous avons

$$G_n^{-1}(t, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A}) = \hat{\beta}_{i\tilde{A}} + J_n^{-1}(t, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}$$

et non pas

$$G_n^{-1}(t, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A}) = \hat{\beta}_{i\hat{A}} + J_n^{-1}(t, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}.$$

Finalement, nous pouvons calculer l'intervalle de confiance bootstrap-t-MSE, en utilisant, encore une fois, uniquement la partie de la variance due aux erreurs.

Soit, pour $i = 0, \ldots, p-1$,

$$L_n(t, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A}) \;=\; P_{\hat{F}_{\tilde{A}}}\{\sqrt{n}(\hat{\beta}^*_{i\hat{A}^*} - \hat{\beta}_{i\tilde{A}})/\hat{\sigma}^*_{\hat{A}^*} \leq t\}, \qquad (2.3.11)$$

l'estimateur, par la méthode de rééchantillonnage des résidus, de (2.3.4).

Définition 2.3.6. *Nous appelons intervalle de confiance bootstrap-t-MSE de niveau* $1-\alpha$ *par le rééchantillonnage des résidus issus du modèle* $\tilde{A}$ *sélectionné par la méthode* S_b*, l'intervalle de confiance pour* β_i*,* $i = 0, \ldots, p-1$ *donné par*

$$[\hat{\beta}_{i\hat{A}} - \hat{\sigma}_{\hat{A}} L_n^{-1}(1-\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}, \hat{\beta}_{i\hat{A}} - \hat{\sigma}_{\hat{A}} L_n^{-1}(\alpha/2, \hat{\beta}_{i\tilde{A}}, \hat{F}_{\tilde{A}}, \hat{A})/\sqrt{n}]$$

suite à la sélection du modèle $\hat{A}$ *par une méthode* S_e*, en utilisant la distribution (2.3.11).*

2.3.4. Sous-échantillonnage

Finalement, nous pouvons utiliser la méthode du sous-échantillonnage si nous supposons que les observations proviennent du modèle 1.1.1. Soit $\hat{A}^\circ$, le modèle sélectionné par la méthode S_e basé sur un sous-échantillon de taille b tiré sans remise parmi les lignes de la matrice Z définie dans la section 2.3.2. Soit $\hat{\beta}^\circ_{i\hat{A}^\circ}$, une composante de l'estimateur du vecteur de paramètres pour cet échantillon. Nous estimerons les quantiles de (2.3.2), pour $i = 0, \ldots, p-1$ par les quantiles de

$$J_{n;b}(t, \hat{\beta}_{i\hat{A}}, \hat{A}) = N_n^{-1} \sum_{\mathcal{S}} \mathcal{I}\{\sqrt{b}(\hat{\beta}^\circ_{i\hat{A}^\circ} - \hat{\beta}_{i\hat{A}}) \leq t\}, \qquad (2.3.12)$$

la distribution expérimentale centrée de $\hat{\beta}^\circ_{i\hat{A}^\circ}$. Pour simplifier la notation, nous avons implicitement sommé les variables indicatrices pour toutes les valeurs prises par $\hat{\beta}^\circ_{i\hat{A}^\circ}$ dans $\mathcal{S}$, l'ensemble des $\binom{n}{b}$ sous-échantillons possibles.

En pratique, nous utiliserons B échantillons de taille b car il serait souvent trop long de calculer $\hat{\beta}^\circ_{i\hat{A}^\circ}$ pour tous les échantillons possibles. Suite à l'ajustement

des quantiles de (2.3.12) par $\sqrt{1/n}$ ou de l'ajustement de la statistique $(\hat{\beta}^{\circ}_{i\hat{A}^{\circ}} - \hat{\beta}_{i\hat{A}})$ par $\sqrt{b/n}$ nous obtenons les intervalles de confiance par la méthode du sous-échantillonnage suivants.

Définition 2.3.7. *Nous appelons intervalle de confiance pivotal de niveau* $1-\alpha$ *par la méthode du sous-échantillonnage l'intervalle de confiance pour* β_i, $i = 0, \ldots, p-1$, *donné par*

$$[\hat{\beta}_{i\hat{A}} - J^{-1}_{n;b}(1-\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{A})/\sqrt{n}, \hat{\beta}_{i\hat{A}} - J^{-1}_{n;b}(\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{A})/\sqrt{n}]$$

suite à la sélection du modèle $\hat{A}$ *par une méthode* S_e, *en utilisant la distribution (2.3.12).*

Définition 2.3.8. *Nous appelons intervalle de confiance percentile de niveau* $1-\alpha$ *par la méthode du sous-échantillonnage l'intervalle de confiance pour* β_i, $i = 0, \ldots, p-1$, *donné par*

$$[\hat{\beta}_{i\hat{A}} + J^{-1}_{n;b}(\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{A})/\sqrt{n}, \hat{\beta}_{i\hat{A}} + J^{-1}_{n;b}(1-\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{A})/\sqrt{n}]$$

suite à la sélection du modèle $\hat{A}$ *par une méthode* S_e, *en utilisant la distribution (2.3.12).*

Pour construire l'intervalle de confiance bootstrap-t-MSE, nous utiliserons la distribution, pour $i = 0, \ldots, p-1$,

$$L_{n;b}(t, \hat{\beta}_{i\hat{A}}, \hat{A}) \;=\; N_n^{-1} \sum_{S} 1\{\sqrt{b}(\hat{\beta}^{\circ}_{i\hat{A}^{\circ}} - \hat{\beta}_{i\hat{A}})/\hat{\sigma}^{\circ}_{\hat{A}^{\circ}} \leq t\}, \qquad (2.3.13)$$

pour estimer (2.3.4). Encore une fois, cette distribution est basée sur des échantillons de taille b dont les éléments sont choisis sans remise, et nous n'utilisons que la partie de la variance due aux erreurs. Suite à l'ajustement des quantiles

par un facteur $\sqrt{1/n}$, nous obtenons l'intervalle de confiance bootstrap-t-MSE par la méthode du sous-échantillonnage.

Définition 2.3.9. *Nous appelons intervalle de confiance bootstrap-t-MSE de niveau* $1-\alpha$ *par la méthode du sous-échantillonnage l'intervalle de confiance pour* β_i, $i = 0, \ldots, p-1$, *donné par*

$$[\hat{\beta}_{i\hat{A}} - \hat{\sigma}_{\hat{A}} L_{n;b}^{-1}(1-\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{A})/\sqrt{n}, \hat{\beta}_{i\hat{A}} - \hat{\sigma}_{\hat{A}} L_{n;b}^{-1}(\alpha/2, \hat{\beta}_{i\hat{A}}, \hat{A})/\sqrt{n}]$$

suite à la sélection du modèle $\hat{A}$ *par une méthode* S_e, *en utilisant la distribution (2.3.13).*

2.4. ALGORITHMES

Nous disposons maintenant de tous les atouts nous permettant de construire les intervalles de confiance bootstrap et par sous-échantillonnage. Nous verrons en détail l'algorithme nous permettant de construire ces intervalles lorsque nous utilisons la méthode de rééchantillonnage des paires d'observations. Nous verrons ensuite les modifications à apporter à cet algorithme lorsque nous utilisons le rééchantillonnage des résidus et le sous-échantillonnage.

2.4.1. Rééchantillonnage des paires d'observations

I Choisir un modèle avec une méthode de sélection S_e, une des méthodes de sélection vue dans les sections 1.3 et 1.4. Suite à la sélection de ce modèle $\hat{A}$, utiliser la matrice $X_{\hat{A}}$, la matrice dont les colonnes correspondent aux indices $\hat{A}$, afin de calculer un estimateur de $\underset{\sim}{\beta}$, $\underset{\sim}{\hat{\beta}}_{\hat{A}}$ défini en (1.1.4) et (1.1.8), un estimateur de sa variance, $\widehat{Var}(\underset{\sim}{\hat{\beta}}_{\hat{A}})$ défini en (1.1.5) et (1.1.9), et la partie de la variance due aux erreurs, $\hat{\sigma}^2_{\hat{A}}$ défini en (2.3.5). Ces estimés nous serviront à calculer les divers intervalles de confiance plus loin.

II Pour $b = 1, ..., B$ (B grand)

Pour simplifier la notation, nous omettrons d'indiquer l'échantillon de référence pour chacune des 3 étapes suivantes. La matrice Z^* devrait se lire $Z^*(b)$.

a) Sélectionner, avec remise, un échantillon de taille n, parmi les lignes de $Z = (X, y)$. Nous obtenons $Z^* = (X^*, \underset{\sim}{y}^*)$, la matrice composée de X^*, les lignes de la matrice X correspondant aux lignes de Z^*, et $\underset{\sim}{y}^*$, les éléments du vecteur $\underset{\sim}{y}$ correspondant aux lignes de Z^*.

b) En utilisant la matrice X^* et le vecteur $\underset{\sim}{y}^*$, choisir le modèle $\hat{A}^*$ avec la méthode de sélection S_e de l'étape I.

c) Estimer $\underset{\sim}{\hat{\beta}}_{\hat{A}}$, à partir de la matrice $X^*_{\hat{A}^*}$ dont les colonnes correspondent aux indices du modèle $\hat{A}^*$, par

$$\underset{\sim}{\hat{\beta}}^*_{\hat{A}^*} = \begin{cases} (X^{*\prime}_{\hat{A}^*} X^*_{\hat{A}^*})^{-1} (X^{*\prime}_{\hat{A}^*} \underset{\sim}{y}^*) & \text{pour l'ensemble des variables sélectionnées;} \\ \underset{\sim}{0} & \text{pour l'ensemble des variables non sélectionnées.} \end{cases}$$

Estimer ensuite la variance des erreurs par

$$(\hat{\sigma}^*_{\hat{A}^*})^2 = \frac{(\underset{\sim}{y}^* - X^* \underset{\sim}{\hat{\beta}}^*_{\hat{A}^*})' (\underset{\sim}{y}^* - X^* \underset{\sim}{\hat{\beta}}^*_{\hat{A}^*})}{n - p_{\hat{A}^*}}$$

où $p_{\hat{A}^*}$ est le nombre de variables contenu dans le modèle sélectionné à partir des données bootstrap.

d) Pour $i = 0, ..., p - 1$, calculer

$$C_{ib} = (\hat{\beta}^*_{i\hat{A}^*}(b) - \hat{\beta}_{i\hat{A}})$$

ainsi que

$$V_{ib} = \frac{(\hat{\beta}^*_{i\hat{A}^*}(b) - \hat{\beta}_{i\hat{A}})}{\hat{\sigma}^*_{\hat{A}^*}(b)}$$

où $\hat{\beta}^*_{i\hat{A}^*}(b)$ et $\hat{\sigma}^*_{\hat{A}^*}(b)$ sont les estimés bootstrap en utilisant le $b^{\text{ième}}$ échantillon.

III Ordonner les C_{ib} et les V_{ib} de façon à obtenir les statistiques d'ordre $C_{i(1)} \leq C_{i(2)} \leq ... \leq C_{i(B)}$ et $V_{i(1)} \leq V_{i(2)} \leq ... \leq V_{i(B)}$.

IV a) Construire un intervalle de confiance bootstrap pivotal de niveau de confiance $1-\alpha$ pour chaque composante β_i $(i = 0, \ldots, p-1)$ en calculant

$$[\hat{\beta}_{i\hat{A}} - C_{i((1-\alpha/2)B)}, \hat{\beta}_{i\hat{A}} - C_{i((\alpha/2)B)}].$$

b) Construire un intervalle de confiance bootstrap percentile de niveau de confiance $1-\alpha$ pour chaque composante β_i $(i = 0, \ldots, p-1)$ en calculant

$$[\hat{\beta}_{i\hat{A}} + C_{i((\alpha/2)B)}, \hat{\beta}_{i\hat{A}} + C_{i((1-\alpha/2)B)}].$$

c) Construire un intervalle de confiance bootstrap-t-MSE de niveau de confiance $1-\alpha$ pour chaque composante β_i $(i = 0, \ldots, p-1)$ en calculant

$$[\hat{\beta}_{i\hat{A}} - \hat{\sigma}_{\hat{A}} V_{i((1-\alpha/2)B)}, \hat{\beta}_{i\hat{A}} - \hat{\sigma}_{\hat{A}} V_{i((\alpha/2)B)}].$$

2.4.2. Rééchantillonnage des résidus

Ajouter l'étape I′ entre l'étape I et l'étape II de la section 2.4.1.

I′ a) Choisir un modèle avec une méthode de sélection S_b afin d'obtenir un modèle $\tilde{A}$ ($\tilde{A} = \Omega$ si aucune sélection n'est effectuée et le modèle complet est utilisé)

b) Calculer $\hat{\underset{\sim}{\beta}}_{\tilde{A}}$ tel que défini en (1.1.4) et (1.1.8) et le vecteur des résidus centrés

$$\hat{\underset{\sim}{\epsilon}}_{\tilde{A}} = \underset{\sim}{y} - X\hat{\underset{\sim}{\beta}}_{\tilde{A}}.$$

Modifier l'étape II de la section 2.4.1 de la façon suivante:

II Pour $b = 1, \ldots, B$ (B grand)

Pour simplifier la notation, nous omettrons d'indiquer l'échantillon de référence pour chacune des 3 étapes suivantes. Le vecteur $\underset{\sim}{\epsilon}^*$ devrait se lire $\underset{\sim}{\epsilon}^*(b)$.

a) Sélectionner, avec remise, un échantillon de taille n parmi les éléments du vecteur $\hat{\underset{\sim}{\epsilon}}_{\tilde{A}}$. Nous obtenons le vecteur des erreurs bootstrap $\underset{\sim}{\epsilon}^*$. Calculer ensuite

$$\underset{\sim}{y}^* = X\hat{\underset{\sim}{\beta}}_{\tilde{A}} + \underset{\sim}{\epsilon}^*$$

qui constitue le modèle bootstrap.

b) En utilisant la matrice X et le vecteur $\underset{\sim}{y}^*$, choisir un modèle $\hat{A}^*$ avec la méthode de sélection S_e de l'étape I.

c) Estimer $\hat{\underset{\sim}{\beta}}_{\tilde{A}}$, à partir de la matrice $X^*_{\hat{A}^*}$ dont les colonnes correspondent aux indices du modèle $\hat{A}^*$, par

$$\hat{\underset{\sim}{\beta}}^*_{\hat{A}^*} = \begin{cases} (X'_{\hat{A}^*}X_{\hat{A}^*})^{-1}(X'_{\hat{A}^*}\underset{\sim}{y}^*) & \text{pour l'ensemble des variables sélectionnées;} \\ \underline{0} & \text{pour l'ensemble des variables non sélectionnées.} \end{cases}$$

Estimer ensuite la variance des erreurs par

$$(\hat{\sigma}^*_{\hat{A}^*})^2 = \frac{(\underset{\sim}{y}^* - X\hat{\underset{\sim}{\beta}}^*_{\hat{A}^*})'(\underset{\sim}{y}^* - X\hat{\underset{\sim}{\beta}}^*_{\hat{A}^*}(b))}{n - p_{\hat{A}^*}}$$

où $p_{\hat{A}^*}$ est le nombre de variables contenu dans le modèle sélectionné.

d) Pour $i = 0, ..., p-1$, calculer

$$\begin{aligned} C_{ib} &= (\hat{\beta}^*_{i\hat{A}^*}(b) - \hat{\beta}_{i\tilde{A}}), \\ D_{ib} &= \hat{\beta}^*_{i\hat{A}^*}(b) \end{aligned}$$

ainsi que

$$V_{ib} = \frac{(\hat{\beta}^*_{i\hat{A}^*}(b) - \hat{\beta}_{i\tilde{A}})}{\hat{\sigma}^*_{\hat{A}^*}(b)}$$

où $\hat{\beta}^*_{i\hat{A}^*}(b)$ et $\hat{\sigma}^*_{\hat{A}^*}(b)$ sont les estimés bootstrap en utilisant le $b^{\text{ième}}$ échantillon.

Modifier l'étape III de la section 2.4.1 de la façon suivante:

III Ordonner les C_{ib}, D_{ib} et les V_{ib} de façon à obtenir les statistiques d'ordre $C_{i(1)} \leq C_{i(2)} \leq ... \leq C_{i(B)}$, $D_{i(1)} \leq D_{i(2)} \leq ... \leq D_{i(B)}$ et $V_{i(1)} \leq V_{i(2)} \leq ... \leq V_{i(B)}$.

Modifier l'étape IVb) de la section 2.4.1 de la façon suivante:

IV b) Construire un intervalle de confiance bootstrap percentile de niveau de confiance $1-\alpha$ pour chaque composante β_i $(i = 0, \ldots, p-1)$ en calculant

$$[D_{i((\alpha/2)B)}, D_{i((1-\alpha/2)B)}].$$

2.4.3. Sous-échantillonnage

Modifier l'étape II de la façon suivante:

II Pour $b = 1, ..., B$ (B grand)

Pour simplifier la notation, nous omettrons d'indiquer l'échantillon de référence pour chacune des 3 étapes suivantes. La matrice Z° devrait se lire $Z^\circ(b)$.

a) Sélectionner, sans remise, un échantillon de taille b, $b < n$ et $b > p$ parmi les lignes de $Z = (X, y)$. Nous obtenons $Z^\circ = (X^\circ, \underset{\sim}{y}^\circ)$, la matrice composée de X°, les lignes de la matrice X correspondant aux lignes de Z°, et $\underset{\sim}{y}^\circ$, les éléments du vecteur $\underset{\sim}{y}$ correspondant aux lignes de Z°.

b) En utilisant la matrice X° et le vecteur $\underset{\sim}{y}^\circ$, choisir le modèle $\hat{A}^\circ$ avec la méthode de sélection S_e de l'étape I.

c) Estimer $\underset{\sim}{\hat{\beta}}_{\hat{A}}$, à partir de la matrice $X^\circ_{\hat{A}^\circ}$ dont les colonnes correspondent aux indices du modèle $\hat{A}^\circ$, par

$$\underset{\sim}{\hat{\beta}}^\circ_{\hat{A}^\circ} = \begin{cases} (X^{\circ\prime}_{\hat{A}^\circ} X^\circ_{\hat{A}^\circ})^{-1} (X^{\circ\prime}_{\hat{A}^\circ} \underset{\sim}{y}^\circ) & \text{pour l'ensemble des variables sélectionnées;} \\ \underset{\sim}{0} & \text{pour l'ensemble des variables non sélectionnées.} \end{cases}$$

Estimer ensuite la variance des erreurs par

$$(\hat{\sigma}^\circ_{\hat{A}^\circ})^2 = \frac{(\underset{\sim}{y}^\circ - X^\circ \underset{\sim}{\hat{\beta}}^\circ_{\hat{A}^\circ})'(\underset{\sim}{y}^\circ - X^\circ \underset{\sim}{\hat{\beta}}^\circ_{\hat{A}^\circ})}{b - p_{\hat{A}^\circ}}$$

où $p_{\hat{A}^\circ}$ est le nombre de variables contenu dans le modèle sélectionné à partir des données du sous-échantillon.

d) Pour $i = 0, \ldots, p-1$ calculer

$$C_{ib} = \sqrt{b/n}(\hat{\beta}^{\circ}_{i\hat{A}^{\circ}}(b) - \hat{\beta}_{i\hat{A}})$$

ainsi que

$$V_{ib} = \sqrt{\frac{b}{n}}\frac{(\hat{\beta}^{\circ}_{i\hat{A}^{\circ}}(b) - \hat{\beta}_{i\hat{A}})}{\hat{\sigma}^{\circ}_{\hat{A}^{\circ}}(b)}$$

où $\hat{\beta}^{\circ}_{i\hat{A}^{\circ}}(b)$ et $\hat{\sigma}^{\circ}_{\hat{A}^{\circ}}(b)$ sont les estimés bootstrap en utilisant le $b^{\text{ième}}$ échantillon.

Chapitre 3

SIMULATIONS

Dans le chapitre précédent, nous avons regardé comment appliquer le rééchantillonnage des paires d'observations, des résidus et le sous-échantillonnage à l'inférence suite à la sélection d'un modèle. Dans ce chapitre, nous débuterons par exposer le plan de nos simulations. Avant de regarder les résultats que nous avons obtenus, nous reviendrons sur les conclusions obtenues par Carignan (1996) sur le rééchantillonnage des résidus. Nous examinerons ensuite les résultats de nos simulations pour le rééchantillonnage des résidus dans la deuxième section, pour le rééchantillonnage des paires d'observations dans la section 3.3 et pour le sous-échantillonnage dans la dernière section. Nous commenterons et comparerons les résultats de nos simulations pour les trois méthodes.

3.1. PLAN DES SIMULATIONS

Nous avons simulé un modèle de régression linéaire tel que décrit par le modèle 1.1.5. Nous avons utilisé 3 méthodes de rééchantillonnage, le rééchantillonnage des paires d'observations (la méthode paires), le rééchantillonnage des résidus (la méthode résidus) et le sous-échantillonnage, afin de calculer des intervalles de confiance sur les coefficients du vecteur $\underset{\sim}{\beta}$.

Pour chacune de ces méthodes nous avons effectué des simulations où nous devions déterminer la matrice de design X, le vecteur des coefficients β, la distribution des erreurs ϵ, le nombre de répétitions de chacune des simulations, le nombre d'échantillons bootstrap générés, le niveau de confiance voulu et la méthode de sélection à évaluer (S_e). De plus, pour la méthode résidus nous devions déterminer la méthode de sélection utilisée, s'il y en a une, afin de déterminer le modèle duquel sera généré les observations bootstrap (S_b). Finalement, nous devions déterminer la taille du sous-échantillon lorsque nous utilisions la méthode du sous-échantillonnage.

Présentons d'abord les paramètres qui sont demeurés constants tout au long des simulations et ce pour chacune des méthodes de rééchantillonnage. Le vecteur des coefficients $\underset{\sim}{\beta}$ est de longueur 8 et fixé à (1,0; 2,0; 1,5; 1,5; 0,5; 0; 0; 0). Ce vecteur est le même que celui utilisé par Zhang (1992) et Carignan (1996) et comprend le coefficient 1,0 correspondant à la constante du modèle de régression. Le vecteur d'erreurs $\underset{\sim}{\epsilon}$ a été généré à partir d'une distribution normale de moyenne 0 et de variance σ^2. Chaque simulation comporte 500 répétitions et nous générons 1000 échantillons bootstrap afin d'obtenir les quantiles bootstrap. La couverture prescrite des différents intervalles de confiance bilatéraux calculés est de 95% et ceux-ci sont construits de façon symétrique.

Présentons maintenant les paramètres qui ont varié lors des simulations. Nous avons utilisé 2 matrices de design X. La première, X_{50}, est la même que celle utilisée par Carignan (1996). Cette matrice fut construite comme celle de Hurvich et Tsai (1990) et Zhang (1992) à l'exception de la colonne de 1. Elle comporte une colonne de 1 et 7 colonnes dont chacune des composantes est une réalisation de variables aléatoires i.i.d. N(0,1). Le nombre d'observations de chacune des variables est de $n = 50$. Cette matrice a été utilisée pour chacune des méthodes

de rééchantillonnage. La seconde matrice, X_{1000}, contient $n = 1000$ observations et est construite de la même façon que X_{50}. Elle a également été utilisée pour chacune des méthodes de rééchantillonnage. Notez que ces deux matrices ont été générées une seule fois et sont demeurées fixes tout au long de la simulation.

Selon la matrice de design X utilisée nous avons employé différentes valeurs de σ^2 pour la variance des erreurs de notre modèle. Pour la matrice X_{50}, nous avons employé $\sigma = 0{,}1$ et $\sigma = 10$; pour la matrice X_{1000}, $\sigma = 0{,}447, 4{,}47$ et $44{,}7$. Nous justifierons plus loin le choix de ces valeurs.

Nous avons considéré 5 méthodes de sélection de modèle à évaluer (S_e): le critère BIC, le retrait par étapes, le C_p de Mallows, la méthode de Ducharme et l'addition par étapes. De plus, pour la méthode résidus, nous devions choisir une méthode S_b. Ces méthodes sont les mêmes que pour la méthode S_e auxquelles nous avons ajouté la possibilité de n'utiliser aucune méthode, c'est-à-dire de considérer le modèle complet. Toutes les combinaisons ont été considérées. Pour chacune des sélections, la constante était automatiquement incluse dans le modèle. La sélection s'effectuait sur les 7 autres variables. Finalement, pour la méthode du sous-échantillonnage, nous devions fixer la taille du sous-échantillon (b). Nous avons utilisé $b = 25$ lorsque nous utilisions X_{50} et $b = 30, 40, 50, 75, 100, 150$ et 200 lors de l'utilisation de la matrice X_{1000}.

Caractéristiques étudiées

Pour chacune des simulations nous avons évalué la qualité du choix du modèle en calculant les proportions associées au nombre de fois où le modèle sélectionné est le "vrai" modèle, un modèle biaisé et un modèle trop grand (voir les définitions 1.2.1 à 1.2.3). Nous avons calculé le pourcentage de couverture de l'intervalle de confiance bilatéral classique et bootstrap (pivotal, percentile, bootstrap-t-MSE)

suite à la sélection d'un modèle ainsi que la proportion des fois où la vraie valeur du coefficient se retrouve à l'extérieur, soit à gauche ou à droite de l'intervalle de confiance. Nous avons aussi estimé la moyenne et l'écart type de la longueur de ces intervalles. Pour certaines simulations, nous avons de plus calculé le pourcentage d'inclusion de chacun des coefficients du vecteur $\underset{\sim}{\beta}$ dans le modèle sélectionné et observé leur taille. Nous avons également observé un certain nombre d'intervalles de confiance pour différentes simulations. Finalement, nous avons étudié la distribution des estimateurs des coefficients de $\underset{\sim}{\beta}$ pour une répétition (1000 échantillons bootstrap) particulière.

Langage utilisé

Carignan (1996) a utilisé le logiciel Splus afin de simuler la sélection de modèle en utilisant le rééchantillonnage des résidus. Le désavantage de ce logiciel est le temps de calcul. Ce dernier n'est pas linéaire en fonction du nombre de répétitions et il est fortement croissant. Afin d'améliorer l'efficacité des programmes, nous avons utilisé le langage Fortran 77 dont le temps de calcul est linéaire en fonction du nombre de répétitions. La librairie NAG version mark 16 de sous-routines Fortran a servi à sélectionner le modèle en utilisant les méthodes séquentielles, à estimer le vecteur des coefficients $\underset{\sim}{\beta}$ et à générer les échantillons sans remise de la méthode du sous-échantillonnage. Nous avons employé l'algorithme de Smith (1991) pour effectuer la sélection de modèle en utilisant les méthodes du meilleur sous-ensemble. L'algorithme de Marsaglia et Tsang (1984) a été utilisé pour générer les N(0,1) et celui de L'Ecuyer (1995) pour les Uniforme[0,1].

3.2. RÉÉCHANTILLONNAGE DES RÉSIDUS

Dans cette section, nous commencerons par résumer les résultats obtenus par Carignan (1996) sur la qualité de couverture des intervalles de confiance classique et bootstrap suite à la sélection d'un modèle en régression linéaire multiple. Ensuite nous verrons les résultats de nos simulations pour cette méthode de rééchantillonnage. Nous comparerons les résultats des intervalles de confiance bootstrap et classiques, étudierons l'importance de la convergence de la méthode de sélection à évaluer, S_e, et l'importance de la méthode de sélection utilisée pour établir le modèle bootstrap, S_b.

3.2.1. Résultats antérieurs

Carignan (1996) a également étudié la qualité de couverture des intervalles de confiance bootstrap et classique en utilisant le rééchantillonnage des résidus. Il voulait, entre autres, déterminer si l'on devait utiliser la même méthode de sélection pour générer les observations bootstrap (c'est-à-dire $S_e = S_b$), l'importance de la convergence de la méthode de sélection S_b et l'impact d'utiliser le modèle complet pour générer les observations bootstrap. Nous exposerons maintenant ses conclusions.

Intervalles de confiance classiques

Débutons tout d'abord par une justification du choix du σ utilisé dans les simulations. A l'aide du logiciel Splus, Carignan a simulé 500 répétitions de la sélection de modèle faites à partir de la même matrice X_{50} et de 500 vecteurs d'observations $\underset{\sim}{y}$ différents. Il a utilisé la méthode de sélection S_e du C_p de Mallows pour construire les intervalles de confiance classiques. Afin de déterminer la valeur de σ à utiliser par la suite, il a examiné la qualité du modèle sélectionné selon la

valeur de σ et le pourcentage de couverture des intervalles de confiance selon le rapport signal-bruit. Nous définirons, de la même façon que Carignan, le rapport signal-bruit pour la i-ième variable comme suit:

$$R_i = \frac{signal}{bruit} = \frac{E(\hat{\beta}_i)}{\sqrt{Var(\hat{\beta}_i)}} = \frac{\beta_i}{\sigma\sqrt{(X'X)^{-1}_{ii}}}.$$

Les tableaux 3.2.1 et 3.2.2 reprennent ces résultats.

TABLEAU 3.2.1. *Qualité de la sélection de modèle pour différentes valeurs du bruit pour 500 sélections effectuées à l'aide de la méthode S_e CPM.*

Matrice de design: X_{50}, S_e CPM			
	Vrai modèle	Modèle trop grand	Modèle biaisé
$\sigma = 0{,}1$	0,564	0,426	0,010
$\sigma = 1$	0,560	0,420	0,020
$\sigma = 2$	0,384	0,288	0,328
$\sigma = 3$	0,234	0,178	0,588
$\sigma = 5$	0,070	0,050	0,880
$\sigma = 10$	0,004	0,008	0,988

Suite à l'analyse de ces tableaux, Carignan a choisi d'utiliser des valeurs de σ de 0,1 et 10 car ces deux valeurs représentent chacune un cas extrême. La valeur de $\sigma = 0{,}1$ lui donne des pourcentages de couverture près de la valeur prescrite de 95% et une très faible proportion de modèles biaisés. Puisque l'approximation de Monte Carlo a été utilisée, les pourcentages de couverture ne sont pas exacts. Si nous supposons que la "vraie" probabilité de couverture de chacun des intervalles est de $p = 95\%$, nous sommes en présence d'une Binomiale(500, p) dont l'écart type est, en pourcentage, $\sqrt{500 \times 0{,}95 \times 0{,}05}/500 \approx 1\%$. Nous considérerons que

TABLEAU 3.2.2. *Couverture des intervalles de confiance classiques pour différentes valeurs du bruit lorsque la méthode de sélection S_e CPM est employée.*

Matrice de design: X_{50}, S_e CPM									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
$\sigma = 0{,}1$									
rapport signal-bruit		62,9	149,8	110,7	76,9	35,9	0	0	0
Classique	Uni. gauche 2,5%	3,0	3,0	4,2	3,6	2,0	3,4	2,2	3,2
	Uni. droite 2,5%	2,4	3,8	5,2	6,4	1,8	2,6	4,0	2,8
	Bilatéral 95%	94,6	93,2	90,6	90,0	96,2	94,0	93,8	94,0
$\sigma = 1$									
rapport signal-bruit		6,29	14,98	11,07	7,69	3,59	0	0	0
Classique	Uni. gauche 2,5%	3,0	3,2	4,4	3,8	2,4	3,2	2,4	3,4
	Uni. droite 2,5%	2,4	3,6	4,8	6,2	2,4	2,6	4,0	2,8
	Bilatéral 95%	94,6	93,2	90,8	90,0	95,2	94,2	93,6	93,8
$\sigma = 2$									
rapport signal-bruit		3,15	7,49	5,54	3,85	1,80	0	0	0
Classique	Uni. gauche 2,5%	3,0	2,8	4,8	3,0	2,2	3,2	2,4	4,4
	Uni. droite 2,5%	2,6	3,8	5,0	6,8	31,6	2,6	4,0	2,6
	Bilatéral 95%	94,4	93,4	90,2	90,2	66,2	94,2	93,6	93,0
$\sigma = 3$									
rapport signal-bruit		2,10	4,99	3,69	2,56	1,20	0	0	0
Classique	Uni. gauche 2,5%	2,2	3,8	3,4	2,4	2,4	2,8	2,4	4,2
	Uni. droite 2,5%	3,2	3,4	3,8	11,0	56,8	1,8	2,2	3,2
	Bilatéral 95%	94,6	92,8	92,8	86,6	40,8	95,4	95,4	92,6
$\sigma = 5$									
rapport signal-bruit		1,26	3,00	2,21	1,54	0,72	0	0	0
Classique	Uni. gauche 2,5%	3,6	4,4	3,4	3,0	3,2	4,8	3,8	6,8
	Uni. droite 2,5%	4,4	6,4	17,4	46,0	77,2	3,0	3,8	2,2
	Bilatéral 95%	92,0	89,2	79,2	51,0	19,6	92,2	92,4	91,0
$\sigma = 10$									
rapport signal-bruit		0,63	1,50	1,11	0,77	0,36	0	0	0
Classique	Uni. gauche 2,5%	3,2	4,6	4,6	3,4	2,0	5,6	4,0	4,8
	Uni. droite 2,5%	3,2	39,4	56,2	75,0	82,4	2,8	3,6	3,0
	Bilatéral 95%	93,6	56,0	39,2	21,6	15,6	91,6	92,4	92,2

le pourcentage de couverture, $\hat{p}$, est près de la valeur prescrite lorsque le test

de niveau 95% ne rejette pas l'hypothèse que $p = 95\%$, c'est-à-dire lorsque $\hat{p} \in [93\%, 97\%]$. Par conséquent, nous considérerons tout résultat en dehors de cet intervalle comme étant significativement différent de 95%. Pour les simulations utilisant la valeur de $\sigma = 10$, les pourcentages de couverture sont très différents de la valeur prescrite pour les coefficients non nuls, à l'exception de β_0, les intervalles de confiance sont très asymétriques et une très forte proportion des modèles sont biaisés. Puisque l'on désire comparer les intervalles de confiance bootstrap aux intervalles de confiance classiques, nous choisirons, tout comme Carignan, une valeur de σ où l'intervalle de confiance classique performe très bien, $\sigma = 0{,}1$ et une autre, $\sigma = 10$, où il performe moins bien. Nous qualifierons de rapport signal-bruit élevé le premier de ces cas et de rapport signal-bruit faible, le second.

Les résultats exposés dans le tableau 3.2.2 nous permettent également de subdiviser les coefficients en 3 catégories. Les coefficients non nuls ou clairement non nuls (β_0 à β_3) sont associés à un rapport signal-bruit au moins 75% plus élevé que celui du coefficient presque nul (β_4). Les coefficients nuls (β_5, β_6 et β_7) sont associés à un rapport signal-bruit nul.

Rapport signal-bruit élevé

Pour un rapport signal-bruit élevé, Carignan a simulé 500 répétitions de la sélection de modèle faites sur la même matrice X_{50} et 500 vecteurs $\underset{\sim}{y}$ différents en utilisant le rééchantillonnage des résidus. Les quantiles des distributions J_n et L_n, définis en (2.3.8) et (2.3.11), pour le calcul des intervalles de confiance bootstrap, ont été obtenus en sélectionnant 1000 modèles avec la méthode S_e sur les 1000 jeux de données bootstrap obtenus à partir de la méthode S_b. Les combinaisons suivantes de S_b et S_e ont été étudiées par Carignan.

S_e: Ducharme S_b: Ducharme, Retrait

S_e: Mallows S_b: Ducharme, Mallows

S_e: Retrait S_b: Ducharme, Retrait

S_e: Addition S_b: Ducharme

Suite à ses simulations, Carignan conclut que l'effet de la méthode S_b est négligeable. L'utilisation de la méthode S_e Ducharme lors de la sélection de modèle entraîne des taux de sélection pratiquement nuls pour les coefficients égaux à zéro et la sélection du vrai modèle dans plus de 90% des cas. La conséquence de cette sélection est le pourcentage de couverture de tout près de 100% pour les coefficients nuls lorsque cette méthode est utilisée. Les trois autres méthodes sélectionnent le vrai modèle dans une proportion variant de 40% à 60%.

Du côté des pourcentages de couverture, Carignan conclut que parmi les 3 intervalles de confiance bootstrap, l'intervalle de confiance bootstrap-t-MSE performe le mieux car le pourcentage de couverture de cet intervalle est plus près de la valeur prescrite que ne l'est le pourcentage de couverture des intervalles de confiance pivotal et percentile. Cependant, la longueur des intervalles de confiance bootstrap-t-MSE pour les coefficients nuls $(\beta_5, \beta_6, \beta_7)$ est plus grande que celle des intervalles de confiance classiques pour les mêmes coefficients. Par conséquent la méthode classique est préférée au rééchantillonnage des résidus pour un rapport signal-bruit élevé.

Toutes ces simulations ont été reprises par notre étude et nous en verrons les résultats détaillés plus loin.

Rapport signal-bruit faible

Carignan simule ensuite, de la même façon, la sélection de modèle pour un rapport signal-bruit faible et les combinaisons de S_e et de S_b suivantes:

S_e: Ducharme S_b: Ducharme, Mallows, Retrait

S_e: Mallows S_b: Ducharme, Mallows, Retrait, Addition, Aucun

S_e: Retrait S_b: Ducharme, Mallows, Retrait, Addition, Aucun

S_e: Addition S_b: Addition

Pour chacune de ces combinaisons, la proportion de modèles biaisés est supérieure à 95%. Nous avons obtenu des proportions de modèles biaisés également supérieures à 95% pour toutes les autres combinaisons S_b et S_e.

Encore une fois, Carignan conclut que l'intervalle de confiance bootstrap-t-MSE est l'intervalle de confiance qui donne les meilleurs résultats parmi les 3 intervalles de confiance bootstrap.

De façon générale, les intervalles de confiance classiques ont des pourcentages de couverture très inférieurs à la valeur prescrite de 95% pour un rapport signal-bruit faible. Suite à ses simulations, Carignan conclut que la méthode de rééchantillonnage des résidus fait mieux que la méthode d'inférence classique. Parmi les combinaisons testées, seules les combinaisons qui utilisent S_e Ducharme ou celles qui utilisent le modèle complet pour construire le jeu de données bootstrap ont des pourcentages de couverture majoritairement en deça de la valeur prescrite pour l'intervalle de confiance bootstrap-t-MSE.

Il a observé que les modèles sélectionnés par la méthode S_e Ducharme sont de petites tailles et que ce fait jumelé à la proportion de modèles biaisés explique la piètre performance de cette méthode par rapport aux autres.

Bref, Carignan conclut que bien que la méthode de rééchantillonnage des résidus fasse mieux pour la couverture des intervalles de confiance que l'inférence classique, les résultats seront beaucoup moins bons pour la méthode de sélection S_e Ducharme. La méthode de sélection S_b semble avoir peu d'influence sur la couverture des intervalles de confiance si ce n'est qu'il est préférable d'en utiliser

une plutôt que d'utiliser le modèle complet pour construire les jeux de données bootstrap.

Nous reprendrons plus loin les combinaisons utilisées par Carignan et exposerons les résultats de façon plus détaillée.

3.2.2. Expérience de simulations

Nous exposerons maintenant en détail les résultats de nos simulations. Nous avons planifié une expérience avec 4 facteurs. Les facteurs sont: l'écart type à 2 niveaux, 0,1 et 10; la méthode de sélection à évaluer S_e à 5 niveaux, BIC, retrait par étapes (BWD), C_p de Mallows (CPM), Ducharme (DUC) et addition par étapes (FWD); la méthode de sélection S_b à 6 niveaux, les 5 premiers étant les mêmes que pour le facteur S_e et le sixième étant l'utilisation du modèle complet (NON); le type d'intervalles de confiance à 4 niveaux, classique, pivotal, percentile et bootstrap-t-MSE. Le plan d'expérience est factoriel donc toutes les combinaisons des méthodes S_e et S_b ont été utilisées. La variable de réponse est le vecteur du pourcentage de couverture du vecteur $\underset{\sim}{\beta}$. Le test de la trace de Pillai-Bartlett a été utilisé pour évaluer les différents effets.

Une première analyse de variance multivariée nous permet de différencier deux groupes selon l'écart type au niveau 5%. On peut noter aussi un effet principal significatif pour la méthode S_b et pour le type d'intervalle au niveau 5%. Les interactions doubles significatives sont les interactions incluant le facteur écart type.

Puisque nous avons noté une interaction entre l'écart type et les autres facteurs, nous avons refait deux autres analyses de variance multivariée, une pour chacun des niveaux du facteur écart type. Pour chacun des niveaux, nous concluons à un effet des méthodes S_e et S_b, ainsi qu'à un effet du facteur du type

d'intervalle de confiance. Aucune interaction double n'est significative au niveau 5%. Il est intéressant de noter que malgré que la méthode S_e n'avait pas un effet significatif globalement, son effet est significatif pour chacun des sous-groupes écart type; l'effet d'un sous-groupe annule l'effet de l'autre.

Afin d'étudier plus à fond les différents contrastes, nous procéderons de manière graphique et par l'observation directe des résultats car la diminution des degrés de liberté des différentes comparaisons ne nous permettrait pas de tirer de conclusions.

Nous avons subdivisé pour un écart type donné les résultats en 6 catégories. La méthode S_e est divisée en deux catégories: les méthodes convergentes (BIC et DUC) et les méthodes non convergentes (BWD, CPM, FWD). La méthode S_b est subdivisée en trois catégories: les méthodes convergentes, non convergentes et l'absence de sélection de modèle (NON). Les 6 catégories sont donc formées par les combinaisons de ces catégories pour les méthodes S_b et S_e.

Pour un écart type de 0,1, la figure 3.2.1 nous montre les 4 types d'intervalle de confiance étudiés selon les 6 catégories mentionnées. Chaque point d'un graphique représente la moyenne des pourcentages de couverture pour un intervalle de confiance donné et les méthodes de sélection S_e et S_b de cette catégorie. Par exemple, le graphique S_e convergente et S_b convergente contient les moyennes des combinaisons S_e et S_b BIC et DUC. On peut voir également l'intervalle [93%, 97%] indiquant la zone pour laquelle les intervalles de confiance n'offrent pas une couverture différente de la couverture prescrite de 95%. Pour toutes les catégories, l'intervalle de confiance classique n'est pas différent de la valeur prescrite. Les intervalles de confiance bootstrap semblent bons même si la méthode de sélection S_e n'est pas convergente excepté pour les intervalles de confiance pivotal

et bootstrap-t-MSE lorsqu'aucune méthode de sélection n'est employée en combinaison avec une méthode S_e convergente.

Cette différence peut être expliquée comme suit. Carignan avait déjà noté que pour la méthode de sélection de Ducharme, la proportion de sélection du vrai modèle était très élevée. Lorsque le rééchantillonnage des résidus est employé, la sélection du modèle au niveau bootstrap est le résultat de deux étapes. A la première étape, la méthode S_b sélectionne un modèle à partir des données originales. Nous estimons alors $\tilde{\underset{\sim}{\beta}}$ et construisons des observations bootstrap où le "vrai" modèle est $X\tilde{\underset{\sim}{\beta}}$. Ensuite, nous sélectionnons un second modèle à partir de ces données bootstrap. Nous obtenons alors $\hat{\underset{\sim}{\beta}}^*$.

En général, si la valeur de β_i est près de 0, son estimé sera également petit et cette variable ne sera pas incluse dans le modèle. Si la variable est choisie alors $\hat{\beta}_i$ ne sera pas près de 0.

Au niveau bootstrap, le "vrai" modèle est celui donné par $\tilde{\underset{\sim}{\beta}}$. Si ce modèle a été obtenu à partir d'une méthode de sélection S_b convergente, il s'agit du vrai modèle dans au moins 90% des cas. S'il a été obtenu à partir d'une méthode de sélection S_b non convergente, cette proportion tombe à 60%. Si nous utilisons le modèle complet afin de construire les observations bootstrap, nous n'avons jamais le vrai modèle. Cependant, la "vraie" valeur (au niveau bootstrap) des coefficients dont la vraie valeur est 0 (β_5 à β_7) est habituellement petite.

A la seconde étape de sélection, la proportion de vrais modèles $\hat{\underset{\sim}{\beta}}^*$ diminue par rapport aux modèles $\tilde{\underset{\sim}{\beta}}$ de la première étape. Dans le pire des cas, lorsque le modèle complet est utilisé, nous obtenons néanmoins au moins 60% de vrais modèles sélectionnés lorsqu'une méthode S_e convergente est utilisée. Par conséquent, une

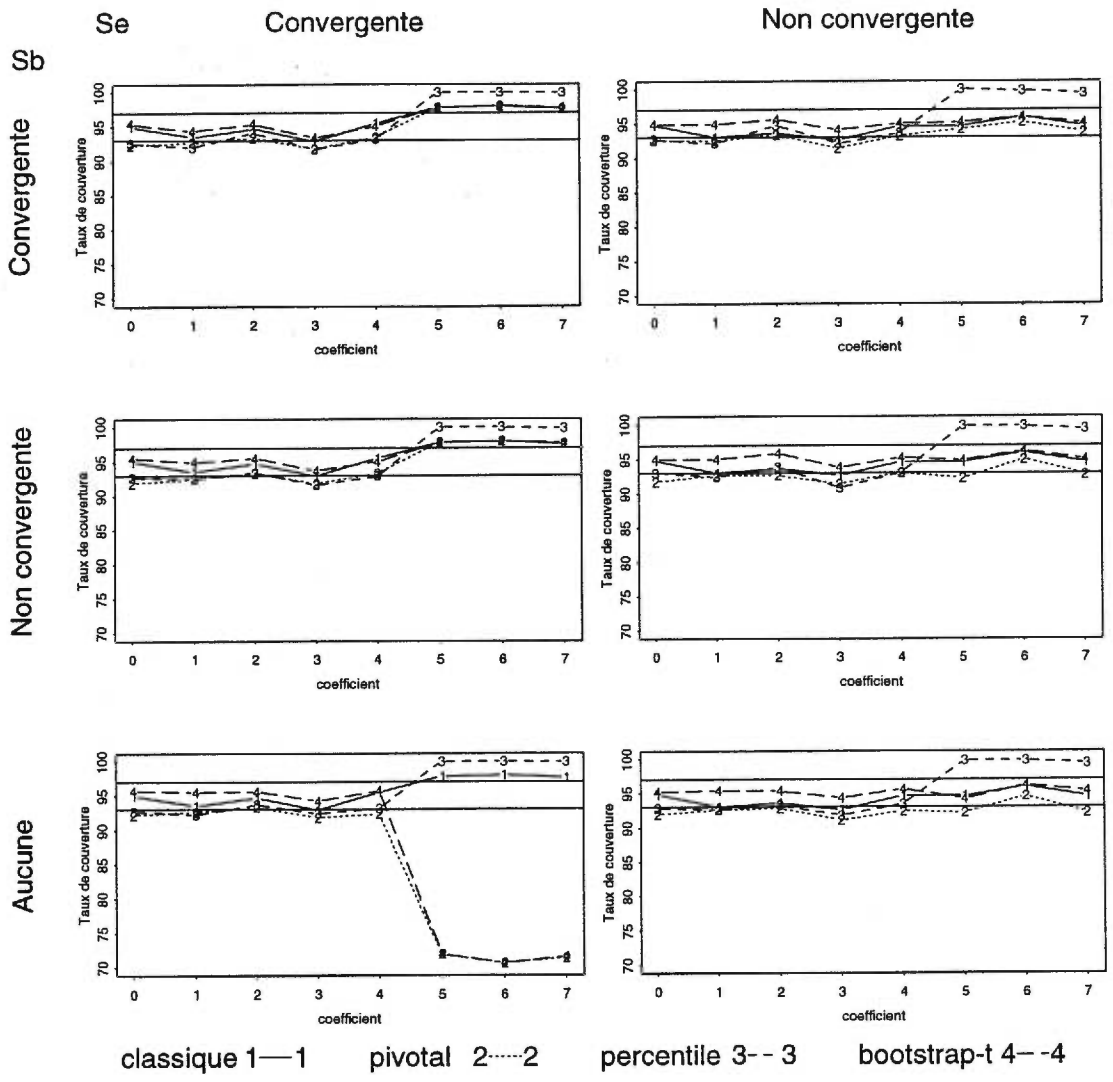

FIGURE 3.2.1. *Moyennes des pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit élevé pour les méthodes de sélection S_e (colonnes) convergentes (BIC, DUC) et non convergentes (BWD, CPM, FWD), en combinaison avec S_b (lignes) convergentes, non convergentes et sans sélection de modèle lorsque la méthode du rééchantillonnage des résidus est employée.*

grande proportion des $\hat{\beta}_i^*$ sont différents de 0 pour $i = 1, \ldots, 4$ et $\hat{\beta}_i^*$ est égal à 0 pour $i = 5, 6, 7$.

Lorsque nous construisons les intervalles de confiance percentile, nous basons nos intervalles sur la distribution des $\hat{\beta}_i^*$. Donc, même si le modèle complet est utilisé, nous inclurons très souvent 0 dans les intervalles de confiance percentile des coefficients β_5, β_6 et β_7.

Les intervalles de confiance pivotal et bootstrap-t-MSE utilisent la distribution de $(\hat{\underset{\sim}{\beta}}^* - \tilde{\underset{\sim}{\beta}})$. En particulier, l'intervalle de confiance pivotal, pour β_i, est de la forme

$$\left[\hat{\beta}_i - (\hat{\beta}_i^* - \tilde{\beta}_i)_{(975)}, \hat{\beta}_i - (\hat{\beta}_i^* - \tilde{\beta}_i)_{(25)}\right],$$

où $(\hat{\beta}_i^* - \tilde{\beta}_i)_{(\delta)}$ représente la δ-ième statistique d'ordre parmi les 1000 estimations de $(\hat{\beta}_i^* - \tilde{\beta}_i)$. Or, nous avons vu que pour une méthode de sélection S_e convergente nous obtenons dans plusieurs cas $\hat{\beta}_i^* = 0$ pour $i = 5, 6, 7$ même si le modèle complet a été utilisé afin de générer les observations bootstrap. Dans ce cas, nous savons que $\tilde{\beta}_i \neq 0$, mais $\hat{\beta}_i = 0$ très souvent pour un S_e convergent. Si suffisamment de $\hat{\beta}_i^*$ sont nuls, nous obtenons deux quantiles bootstrap nuls et l'intervalle de confiance pivotal devient alors uniquement l'ensemble $\{\tilde{\beta}_i\}$ qui ne contient pas la valeur 0. Le même problème survient lorsque nous construisons l'intervalle de confiance bootstrap-t-MSE. C'est pour cette raison que les pourcentages de couverture des intervalles de confiance pivotal et bootstrap-t-MSE des coefficients nuls avec l'utilisation du modèle complet et une méthode de sélection S_e convergente sont inférieurs aux pourcentages de couverture obtenus lorsque nous utilisons une méthode de sélection S_b.

Regardons plus en détail les pourcentages de couverture des intervalles de confiance pour les méthodes S_e BIC et CPM en combinaison avec les méthodes S_b BIC, CPM et NON pour un écart type de 0,1. Nous notons dans les tableaux 3.2.3 et 3.2.4 que l'intervalle de confiance bootstrap dont le pourcentage de couverture

est le plus près de la valeur prescrite de 95% est l'intervalle de confiance bootstrap-t-MSE excepté pour la combinaison S_b NON et S_e BIC. Dans ce cas, l'intervalle de confiance percentile fait mieux. Sur ce point nous confirmons les conclusions de Carignan.

Si nous regardons la longueur moyenne des intervalles de confiance dans le tableau 3.2.5, nous remarquons, tout comme Carignan, que pour les coefficients nuls, β_5, β_6 et β_7, les intervalles de confiance bootstrap sont toujours plus larges que les intervalles de confiance classiques. Pour la méthode de sélection S_e BIC, la combinaison avec la méthode S_b BIC entraîne des intervalles de confiance bootstrap 3 à 4 fois plus larges que les intervalles de confiance classiques. En combinaison avec la méthode S_b CPM, les intervalles de confiance bootstrap sont 6 à 8 fois plus larges et si aucune sélection n'est effectuée à la première étape, les intervalles de confiance bootstrap sont au moins 20 fois plus larges. Il n'y a pas de différence pour les coefficients différents de zéro.

En regardant de plus près cette dernière combinaison, S_b NON et S_e BIC, nous remarquons que non seulement le taux de couverture des coefficients nuls est-il éloigné de la couverture prescrite, mais les longueurs de ces intervalles sont au moins 3 fois plus grandes que les longueurs des intervalles de confiance bootstrap utilisant une méthode de sélection S_b. Pour une méthode de sélection S_e non convergente sans sélection de modèle à la première étape, cette différence n'existe pas même si toutes les combinaisons utilisant la méthode de sélection S_e CPM entraînent des intervalles de confiance bootstrap 5 à 6 fois plus larges que les intervalles de confiance classiques.

Cette progression dans la longueur des intervalles de confiance bootstrap est directement reliée à la proportion de vrais modèles sélectionnés. Lorsque cette proportion est élevée, davantage d'estimés de β_5, β_6 et β_7 sont nuls et l'intervalle

TABLEAU 3.2.3. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit élevé pour la méthode de sélection S_e BIC en combinaison avec S_b BIC, CPM et NON lorsque la méthode du rééchantillonnage des résidus est employée.*

Matrice de design X_{50}, $\sigma = 0{,}10$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC									
Classique	Uni. gauche 2,5%	2,40	2,60	3,20	4,00	1,20	1,80	1,60	2,20
	Uni. droite 2,5%	2,60	4,00	2,40	3,40	3,60	2,00	1,60	1,60
	Bilatéral 95%	95,00	93,40	94,40	92,60	95,20	96,20	96,80	96,20
S_b:BIC									
Pivotal	Uni. gauche 2,5%	4,00	2,60	3,60	5,20	2,20	2,80	2,40	2,60
	Uni. droite 2,5%	3,60	5,00	3,20	3,60	4,60	3,00	2,40	3,60
	Bilatéral 95%	92,40	92,40	93,20	91,20	93,20	94,20	95,20	93,80
Percentile	Uni. gauche 2,5%	3,60	3,00	3,40	4,80	2,40	1,20	1,20	1,40
	Uni. droite 2,5%	3,60	4,60	3,20	4,00	4,40	1,80	1,00	1,80
	Bilatéral 95%	92,80	92,40	93,40	91,20	93,20	97,00	97,80	96,80
Bootstrap-t	Uni. gauche 2,5%	2,60	2,20	2,60	4,00	1,60	2,20	2,00	2,40
	Uni. droite 2,5%	2,60	2,80	1,80	2,00	3,40	2,60	2,00	2,60
	Bilatéral 95%	94,80	95,00	95,60	94,00	95,00	95,20	96,00	95,00
S_b:CPM									
Pivotal	Uni. gauche 2,5%	4,40	3,00	3,60	4,80	2,00	1,80	1,60	2,20
	Uni. droite 2,5%	3,60	4,60	3,20	3,80	4,80	2,00	1,60	1,60
	Bilatéral 95%	92,00	92,40	93,20	91,40	93,20	96,20	96,80	96,20
Percentile	Uni. gauche 2,5%	3,80	3,00	4,00	4,80	2,20	0,00	0,00	0,20
	Uni. droite 2,5%	3,40	4,40	2,80	3,80	4,80	0,00	0,00	0,00
	Bilatéral 95%	92,80	92,60	93,20	91,40	93,00	100,00	100,00	99,80
Bootstrap-t	Uni. gauche 2,5%	2,20	2,00	2,60	4,00	1,80	1,80	1,60	2,20
	Uni. droite 2,5%	2,40	3,20	2,20	2,40	3,20	2,00	1,60	1,40
	Bilatéral 95%	95,40	94,80	95,20	93,60	95,00	96,20	96,80	96,40
S_b:NON									
Pivotal	Uni. gauche 2,5%	4,00	2,60	3,80	4,80	2,40	6,60	8,60	6,60
	Uni. droite 2,5%	3,80	4,80	3,20	3,80	5,20	8,20	5,60	6,60
	Bilatéral 95%	92,20	92,60	93,00	91,40	92,40	85,20	85,80	86,80
Percentile	Uni. gauche 2,5%	3,20	2,80	3,80	4,20	2,20	0,00	0,00	0,20
	Uni. droite 2,5%	3,80	4,80	2,20	3,40	4,60	0,00	0,00	0,00
	Bilatéral 95%	93,00	92,40	94,00	92,40	93,20	100,00	100,00	99,80
Bootstrap-t	Uni. gauche 2,5%	1,60	1,60	2,80	2,80	1,60	6,60	8,60	6,60
	Uni. droite 2,5%	2,60	3,00	1,80	2,80	3,00	8,20	5,60	6,40
	Bilatéral 95%	95,80	95,40	95,40	94,40	95,40	85,20	85,80	87,00

de confiance est plus court. Par exemple, pour la méthode de sélection S_e BIC, la

TABLEAU 3.2.4. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit élevé pour la méthode de sélection S_e CPM en combinaison avec S_b BIC, CPM et NON lorsque la méthode du rééchantillonnage des résidus est employée.*

Matrice de design X_{50}, $\sigma = 0{,}10$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e CPM									
Classique	Uni. gauche 2,5%	2,40	2,60	3,60	4,40	1,40	2,40	1,80	2,40
	Uni. droite 2,5%	2,80	4,40	2,80	3,00	4,00	3,00	2,20	3,00
	Bilatéral 95%	94,80	93,00	93,60	92,60	94,60	94,60	96,00	94,60
S_b:BIC									
Pivotal	Uni. gauche 2,5%	4,00	2,60	3,60	5,20	2,20	2,80	2,40	2,60
	Uni. droite 2,5%	3,60	5,00	3,20	3,60	4,60	3,00	2,40	3,60
	Bilatéral 95%	92,40	92,40	93,20	91,20	93,20	94,20	95,20	93,80
Percentile	Uni. gauche 2,5%	3,80	3,60	3,20	4,60	2,20	0,00	0,00	0,60
	Uni. droite 2,5%	3,60	4,40	2,40	3,80	4,40	0,00	0,20	0,00
	Bilatéral 95%	92,60	92,00	94,40	91,60	93,40	100,00	99,80	99,40
Bootstrap-t	Uni. gauche 2,5%	2,60	2,20	2,60	4,00	1,60	2,20	2,00	2,40
	Uni. droite 2,5%	2,60	2,80	1,80	2,00	3,40	2,60	2,00	2,60
	Bilatéral 95%	94,80	95,00	95,60	94,00	95,00	95,20	96,00	95,00
S_b:CPM									
Pivotal	Uni. gauche 2,5%	4,60	2,60	3,80	5,00	2,20	3,80	2,60	3,00
	Uni. droite 2,5%	3,80	5,00	3,60	3,60	4,80	4,00	2,40	4,20
	Bilatéral 95%	91,60	92,40	92,60	91,40	93,00	92,20	95,00	92,80
Percentile	Uni. gauche 2,5%	3,60	3,00	3,60	5,00	2,40	0,20	0,00	0,60
	Uni. droite 2,5%	3,40	4,60	2,60	4,20	4,60	0,00	0,20	0,00
	Bilatéral 95%	93,00	92,40	93,80	90,80	93,00	99,80	99,80	99,40
Bootstrap-t	Uni. gauche 2,5%	2,60	2,20	2,60	4,20	1,40	2,40	1,80	2,20
	Uni. droite 2,5%	2,40	2,80	1,60	2,00	3,40	2,80	2,00	2,80
	Bilatéral 95%	95,00	95,00	95,80	93,80	95,20	94,80	96,20	95,00
S_b:NON									
Pivotal	Uni. gauche 2,5%	4,20	2,60	4,00	4,60	2,40	3,80	2,80	3,40
	Uni. droite 2,5%	4,00	4,80	3,20	4,20	5,20	4,00	2,60	4,40
	Bilatéral 95%	91,80	92,60	92,80	91,20	92,40	92,20	94,60	92,20
percentile	Uni. gauche 2,5%	3,20	2,60	4,40	4,20	2,20	0,20	0,00	0,60
	Uni. droite 2,5%	4,00	4,80	2,40	4,00	4,40	0,00	0,20	0,00
	Bilatéral 95%	92,80	92,60	93,20	91,80	93,40	99,80	99,80	99,40
Bootstrap-t	Uni. gauche 2,5%	2,20	1,80	2,80	3,40	1,40	2,60	1,80	2,20
	Uni. droite 2,5%	2,60	2,80	2,00	2,20	3,00	3,20	2,00	2,40
	Bilatéral 95%	95,20	95,40	95,20	94,40	95,60	94,20	96,20	95,40

proportion de vrais modèles sélectionnés parmi les 500 000 modèles bootstrap est

TABLEAU 3.2.5. *Moyennes et écarts type des longueurs des intervalles de confiance pour une rapport signal-bruit élevé pour les méthodes de sélection S_e BIC et CPM en combinaison avec S_b BIC, CPM et NON lorsque la méthode du rééchantillonnage des résidus est employée.*

Matrice de design X_{50}, $\sigma = 0{,}10$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC									
Classique	Moyenne	0,0612	0,0502	0,0494	0,0721	0,0536	0,0023	0,0017	0,0024
	Écart type	0,0065	0,0053	0,0052	0,0076	0,0057	0,0116	0,0093	0,0120
S_b:BIC									
Pivotal	Moyenne	0,0572	0,0471	0,0469	0,0683	0,0502	0,0080	0,0064	0,0079
& Percentile	Écart type	0,0064	0,0052	0,0051	0,0076	0,0056	0,0161	0,0140	0,0173
Bootstrap-t	Moyenne	0,0626	0,0515	0,0514	0,0747	0,0547	0,0097	0,0077	0,0095
	Écart type	0,0069	0,0057	0,0056	0,0083	0,0061	0,0192	0,0167	0,0208
S_b:CPM									
Pivotal	Moyenne	0,0569	0,0471	0,0473	0,0685	0,0500	0,0146	0,0113	0,0145
& Percentile	Écart type	0,0063	0,0052	0,0056	0,0077	0,0056	0,0223	0,0192	0,0232
Bootstrap-t	Moyenne	0,0629	0,0519	0,0522	0,0757	0,0550	0,0178	0,0138	0,0176
	Écart type	0,0070	0,0058	0,0063	0,0087	0,0062	0,0270	0,0232	0,0281
S_b:NON									
Pivotal	Moyenne	0,0567	0,0472	0,0479	0,0690	0,0497	0,0382	0,0333	0,0402
& Percentile	Écart type	0,0063	0,0053	0,0056	0,0077	0,0056	0,0165	0,0145	0,0161
Bootstrap-t	Moyenne	0,0633	0,0526	0,0534	0,0771	0,0552	0,0458	0,0400	0,0480
	Écart type	0,0070	0,0059	0,0063	0,0086	0,0062	0,0197	0,0172	0,0192
S_e CPM									
Classique	Moyenne	0,0610	0,0502	0,0496	0,0723	0,0535	0,0102	0,0081	0,0103
	Écart type	0,0065	0,0053	0,0054	0,0077	0,0057	0,0233	0,0197	0,0237
S_b:BIC									
Pivotal	Moyenne	0,0578	0,0480	0,0485	0,0699	0,0507	0,0598	0,0528	0,0621
& Percentile	Écart type	0,0065	0,0054	0,0053	0,0078	0,0056	0,0065	0,0059	0,0071
Bootstrap-t	Moyenne	0,0634	0,0526	0,0531	0,0767	0,0554	0,0652	0,0575	0,0677
	Écart type	0,0071	0,0059	0,0058	0,0085	0,0061	0,0073	0,0065	0,0078
S_b:CPM									
Pivotal	Moyenne	0,0574	0,0477	0,0484	0,0696	0,0503	0,0589	0,0519	0,0611
& Percentile	Écart type	0,0064	0,0054	0,0055	0,0077	0,0056	0,0066	0,0061	0,0072
Bootstrap-t	Moyenne	0,0636	0,0528	0,0535	0,0772	0,0555	0,0655	0,0576	0,0679
	Écart type	0,0071	0,0059	0,0061	0,0085	0,0061	0,0074	0,0066	0,0078
S_b:NON									
Pivotal	Moyenne	0,0571	0,0478	0,0487	0,0699	0,0500	0,0523	0,0458	0,0537
& Percentile	Écart type	0,0064	0,0054	0,0055	0,0078	0,0056	0,0099	0,0085	0,0103
Bootstrap-t	Moyenne	0,0641	0,0535	0,0544	0,0784	0,0559	0,0606	0,0530	0,0623
	Écart type	0,0072	0,0060	0,0061	0,0087	0,0062	0,0101	0,0087	0,0106

de 83,13% pour la combinaison avec la méthode de sélection S_b BIC, 72,70% pour la combinaison avec S_b CPM et 61,53% pour la combinaison avec S_b NON. Le tableau 3.2.6 montre la qualité des 500 000 modèles bootstrap sélectionnés avec toutes les combinaisons de méthodes de sélection S_b et S_e possibles. Certaines combinaisons avaient été étudiées par Carignan, mais nous les avons toutes reprises.

Etudions maintenant les taux de couverture des intervalles de confiance en utilisant un rapport signal-bruit faible. La figure 3.2.2 illustre les différents taux de couverture des intervalles de confiance selon les 6 catégories établies précédemment. Nous notons immédiatement la piètre performance des intervalles de confiance classiques pour chacune des 6 catégories, en particulier pour les coefficients non nuls $(\beta_1, \beta_2, \beta_3, \beta_4)$, excluant la constante, β_0. Pour une méthode de sélection S_e convergente, il n'y a pas de différence entre les intervalles de confiance bootstrap et les intervalles classiques lorsqu'on la combine avec une méthode de sélection S_b convergente. Une amélioration du taux de couverture est observée lorsqu'une méthode S_b non convergente est utilisée. Si nous utilisons le modèle complet afin de construire les observations bootstrap, le taux de couverture des intervalles de confiance bootstrap est encore meilleur pour les coefficients non nuls, mais le taux de couverture des coefficients nuls diminue sous la valeur prescrite de 95% contrairement au cas où une méthode de sélection S_b est utilisée.

Pour une méthode de sélection S_e non convergente, les taux de couverture observés des intervalles de confiance bootstrap ne sont pas différents de la valeur prescrite lorsque les observations bootstrap ont été obtenues suite à la sélection d'un modèle par une méthode S_b. Lorsque le modèle complet a été utilisé, les taux de couverture de certains coefficients ne sont pas différents de 95% selon que

TABLEAU 3.2.6. *Qualité de la sélection de modèle pour 500 000 sélections bootstrap effectuées par rééchantillonnage des résidus.*

Matrice de design X_{50}, $\sigma = 0{,}10$				
S_e	S_b	Vrai modèle	Modèle trop grand	Modèle biaisé
BIC	BIC	0,831	0,169	0,000
	BWD	0,724	0,276	0,000
	CPM	0,727	0,273	0,000
	DUC	0,876	0,124	0,000
	FWD	0,726	0,274	0,000
	NON	0,615	0,385	0,000
BWD	BIC	0,528	0,472	0,000
	BWD	0,417	0,583	0,000
	CPM	0,420	0,580	0,000
	DUC	0,566	0,434	0,000
	FWD	0,419	0,581	0,000
	NON	0,278	0,722	0,000
CPM	BIC	0,538	0,462	0,000
	BWD	0,425	0,575	0,000
	CPM	0,429	0,571	0,000
	DUC	0,577	0,423	0,000
	FWD	0,428	0,572	0,000
	NON	0,285	0,715	0,000
DUC	BIC	0,917	0,083	0,000
	BWD	0,845	0,155	0,000
	CPM	0,846	0,154	0,000
	DUC	0,952	0,048	0,000
	FWD	0,846	0,154	0,000
	NON	0,781	0,219	0,000
FWD	BIC	0,530	0,470	0,000
	BWD	0,418	0,582	0,000
	CPM	0,422	0,578	0,000
	DUC	0,568	0,432	0,000
	FWD	0,421	0,579	0,000
	NON	0,280	0,720	0,000

l'intervalle de confiance est bootstrap-t-MSE, pivotal ou percentile. L'intervalle de confiance bootstrap-t-MSE performe bien pour tous les coefficients saufs les coefficients β_1 et β_2. Les pourcentages de couverture des intervalles de confiance

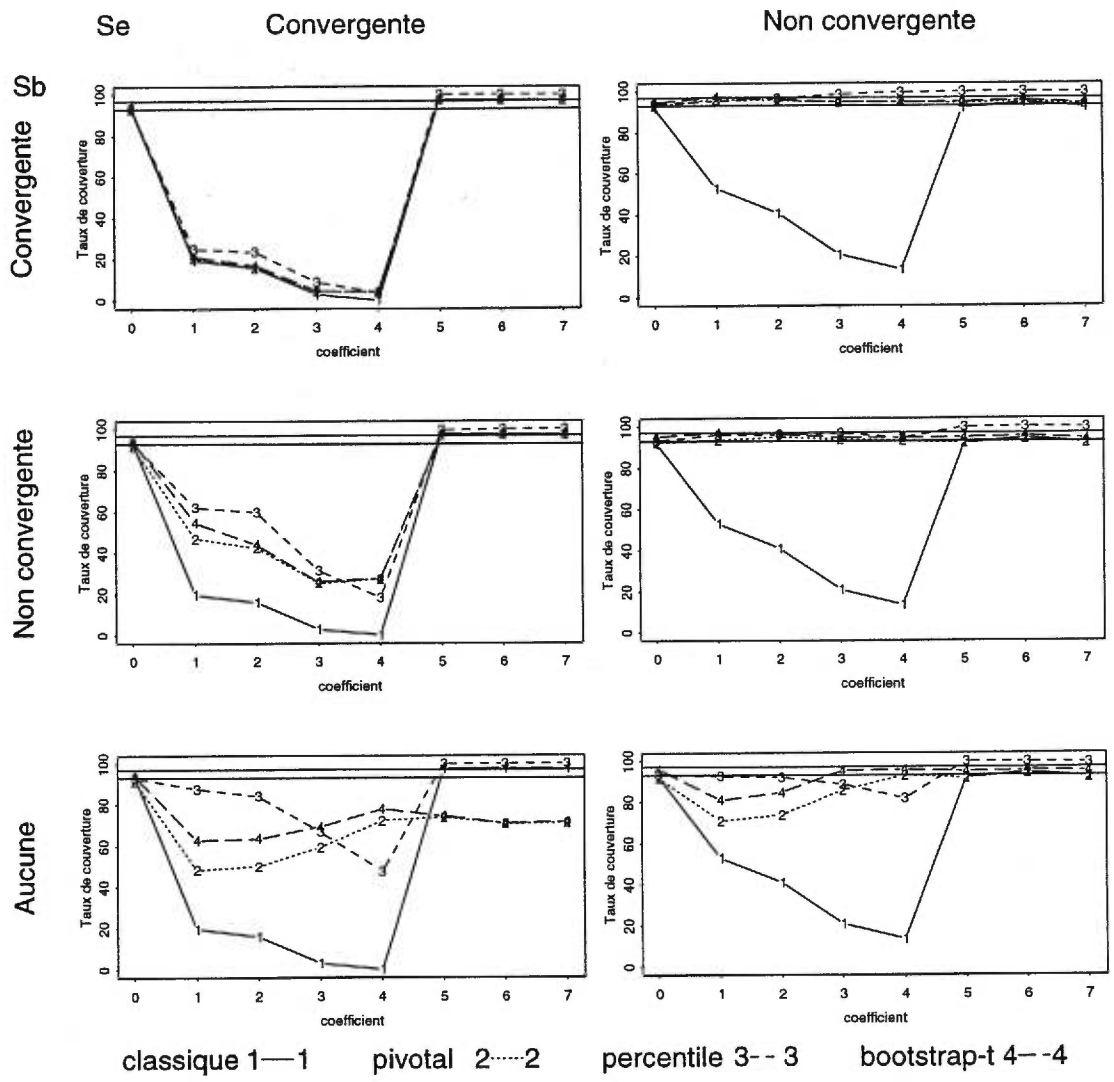

FIGURE 3.2.2. *Moyennes des pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit faible pour les méthodes de sélection S_e (colonnes) convergentes (BIC, DUC) et non convergentes (BWD, CPM, FWD), en combinaison avec S_b (lignes) convergentes, non convergentes et sans sélection de modèle lorsque la méthode du rééchantillonnage des résidus est employée.*

pivotal sont légèrement inférieurs aux pourcentages des intervalles bootstrap-t-MSE. En raison de ce plus faible pourcentage, nous devons ajouter le coefficient

β_3 à la liste des coefficients dont le pourcentage de couverture diffère de 95%. L'intervalle de confiance percentile performe bien pour β_1, β_2 et les coefficients nuls, mais les pourcentages de couverture sont sous la valeur prescrite pour β_3 et β_4.

Nous remarquons lors de l'observation détaillée des taux de couverture dans le tableau 3.2.7 que pour une méthode S_e convergente, l'intervalle de confiance percentile est le meilleur pour tous les coefficients excepté le coefficient β_4. Dans ce cas, l'intervalle de confiance bootstrap-t-MSE lui est supérieur. Pour tous les coefficients cependant, les pourcentages de couverture sont très en-dessous de la valeur prescrite.

Le tableau 3.2.8 nous montre qu'il y a peu de différences entre les intervalles de confiance pivotal, percentile et bootstrap-t-MSE. Ils sont tous bons lorsque jumelés avec une méthode de sélection S_b. Nous pouvons donner un léger avantage à l'intervalle de confiance bootstrap-t-MSE. Lorsqu'aucune sélection n'est effetuée à la première étape, l'intervalle de confiance percentile est le meilleur pour tous les coefficients sauf β_3 et β_4. Dans ces 2 cas, l'intervalle de confiance bootstrap-t-MSE est le meilleur.

Puisque les pourcentages de couverture des intervalles de confiance bootstrap sont très supérieurs aux pourcentages de couverture des intervalles de confiance classiques, il est normal que les intervalles de confiance bootstrap soient plus larges que les intervalles de confiance classiques, trop optimistes. Le tableau 3.2.9 nous montre les longueurs moyennes et les écarts type des intervalles de confiance pour les méthodes de sélection S_e BIC et CPM en combinaison avec les méthodes de sélection S_b BIC, CPM et NON lorsque le raport signal-bruit est faible. Plus l'intervalle de confiance des coefficients non nuls est large, plus le pourcentage de couverture est élevé. Ce phénomène est particulièrement visible pour le coefficient

TABLEAU 3.2.7. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit faible pour la méthode de sélection S_e BIC en combinaison avec S_b BIC, CPM et NON lorsque la méthode du rééchantillonnage des résidus est employée.*

Matrice de design X_{50}, $\sigma = 10$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC									
Classique	Uni. gauche 2,5%	3,80	3,40	4,20	3,80	1,60	3,40	2,40	3,80
	Uni. droite 2,5%	2,80	71,80	75,40	92,20	98,20	0,40	1,20	0,20
	Bilatéral 95%	93,40	24,80	20,40	4,00	0,20	96,20	96,40	96,00
S_b:BIC									
Pivotal	Uni. gauche 2,5%	3,80	3,80	4,80	4,20	1,80	3,40	2,40	3,60
	Uni. droite 2,5%	3,60	70,00	74,60	89,00	87,80	0,40	1,20	0,20
	Bilatéral 95%	92,60	26,20	20,60	6,80	10,40	96,20	96,40	96,20
Percentile	Uni. gauche 2,5%	3,60	2,00	0,40	0,00	0,00	0,00	0,00	0,00
	Uni. droite 2,5%	2,80	63,40	66,40	86,40	92,60	0,00	0,00	0,00
	Bilatéral 95%	93,60	34,60	33,20	13,60	7,40	100,00	100,00	100,00
Bootstrap-t	Uni. gauche 2,5%	3,20	2,80	3,40	3,60	1,60	3,40	2,40	3,60
	Uni. droite 2,5%	2,60	69,80	74,60	89,00	87,80	0,20	1,20	0,20
	Bilatéral 95%	94,20	27,40	22,00	7,40	10,60	96,40	96,40	96,20
S_b:CPM									
Pivotal	Uni. gauche 2,5%	4,40	3,80	4,20	4,20	1,60	3,40	2,40	3,60
	Uni. droite 2,5%	3,80	47,40	53,20	67,60	64,20	0,40	1,20	0,20
	Bilatéral 95%	91,80	48,80	42,60	28,20	34,20	96,20	96,40	96,20
Percentile	Uni. gauche 2,5%	3,40	1,40	0,00	0,00	0,00	0,20	0,00	0,00
	Uni. droite 2,5%	2,80	30,80	33,40	62,60	77,60	0,00	0,00	0,00
	Bilatéral 95%	93,80	67,80	66,60	37,40	22,40	99,80	100,00	100,00
Bootstrap-t	Uni. gauche 2,5%	3,00	2,60	2,80	3,60	1,40	2,80	2,40	3,40
	Uni. droite 2,5%	2,60	41,40	52,80	67,60	64,00	0,20	1,20	0,20
	Bilatéral 95%	94,40	56,00	44,40	28,80	34,60	97,00	96,40	96,40
S_b:NON									
Pivotal	Uni. gauche 2,5%	4,20	2,80	4,60	4,20	2,80	5,60	5,20	6,40
	Uni. droite 2,5%	4,20	47,00	42,20	31,40	15,60	12,40	14,00	12,40
	Bilatéral 95%	91,60	50,20	53,20	64,40	81,60	82,00	80,80	81,20
Percentile	Uni. gauche 2,5%	3,20	1,60	0,20	0,00	0,00	0,00	0,00	0,00
	Uni. droite 2,5%	3,40	8,60	12,00	26,60	45,20	0,00	0,00	0,00
	Bilatéral 95%	93,40	89,80	87,80	73,40	54,80	100,00	100,00	100,00
Bootstrap-t	Uni. gauche 2,5%	2,40	2,40	2,60	3,40	1,60	4,40	5,00	5,40
	Uni. droite 2,5%	3,00	33,20	30,80	22,40	12,00	12,40	14,00	12,40
	Bilatéral 95%	94,60	64,40	66,60	74,20	86,40	83,20	81,00	82,20

β_3 lorsque la méthode de sélection S_e BIC est employée. Pour les combinaisons

TABLEAU 3.2.8. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit faible pour la méthode de sélection S_e CPM en combinaison avec S_b BIC, CPM et NON lorsque la méthode du rééchantillonnage des résidus est employée.*

Matrice de design X_{50}, $\sigma = 10$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e CPM									
Classique	Uni. gauche 2,5%	4,20	3,00	4,40	3,80	1,80	5,00	3,40	4,80
	Uni. droite 2,5%	3,80	44,40	55,80	75,80	85,00	3,20	2,20	2,80
	Bilatéral 95%	92,00	52,60	39,80	20,40	13,20	91,80	94,40	92,40
S_b:BIC									
Pivotal	Uni. gauche 2,5%	4,00	3,00	4,00	3,80	1,60	3,60	3,00	4,00
	Uni. droite 2,5%	3,60	1,40	0,60	1,40	3,60	2,40	1,80	2,40
	Bilatéral 95%	92,40	95,60	95,40	94,80	94,80	94,00	95,20	93,60
Percentile	Uni. gauche 2,5%	3,80	3,40	3,20	1,20	0,00	0,20	0,00	0,20
	Uni. droite 2,5%	2,40	1,00	0,20	0,40	0,80	0,00	0,00	0,00
	Bilatéral 95%	93,80	95,60	96,60	98,40	99,20	99,80	100,00	99,80
Bootstrap-t	Uni. gauche 2,5%	2,60	2,60	3,40	3,60	1,60	2,80	2,60	3,80
	Uni. droite 2,5%	3,00	0,40	0,60	1,40	3,80	1,80	1,80	2,00
	Bilatéral 95%	94,40	97,00	96,00	95,00	94,60	95,40	95,60	94,20
S_b:CPM									
Pivotal	Uni. gauche 2,5%	4,00	3,40	4,40	4,60	2,60	5,20	3,40	4,80
	Uni. droite 2,5%	3,60	2,80	0,60	1,40	4,60	3,20	2,40	3,20
	Bilatéral 95%	92,40	93,80	95,00	94,00	92,80	91,60	94,20	92,00
Percentile	Uni. gauche 2,5%	3,80	2,80	3,40	1,60	0,00	0,20	0,00	0,20
	Uni. droite 2,5%	3,00	1,40	0,60	1,40	5,20	0,00	0,00	0,00
	Bilatéral 95%	93,20	95,80	96,00	97,00	94,80	99,80	100,00	99,80
Bootstrap-t	Uni. gauche 2,5%	2,20	2,40	3,40	3,80	1,60	3,40	2,80	3,40
	Uni. droite 2,5%	2,40	1,00	0,60	1,40	3,80	2,20	1,80	2,40
	Bilatéral 95%	95,40	96,60	96,00	94,80	94,60	94,40	95,40	94,20
S_b:NON									
Pivotal	Uni. gauche 2,5%	4,40	3,00	4,40	4,60	2,80	4,80	3,40	4,20
	Uni. droite 2,5%	3,80	26,60	23,20	9,80	4,40	3,80	2,40	3,60
	Bilatéral 95%	91,80	70,40	72,40	85,60	92,80	91,40	94,20	92,20
Percentile	Uni. gauche 2,5%	2,40	2,80	3,80	2,20	0,00	0,20	0,00	0,60
	Uni. droite 2,5%	4,00	4,80	4,60	9,80	18,20	0,00	0,20	0,00
	Bilatéral 95%	93,60	92,40	91,60	88,00	81,80	99,80	99,80	99,40
Bootstrap-t	Uni. gauche 2,5%	1,60	1,80	2,80	3,00	1,60	2,20	2,80	2,80
	Uni. droite 2,5%	2,60	17,80	12,80	1,80	2,80	2,80	1,80	2,20
	Bilatéral 95%	95,80	80,40	84,40	95,20	95,60	95,00	95,40	95,00

avec S_b BIC, CPM et NON, l'intervalle de confiance bootstrap est respectivement

1,75, 4,5 et 8,8 fois plus large que l'intervalle de confiance classique. Le pourcentage de couverture pour l'intervalle de confiance percentile pour ce coefficient passe de 13,60% à 37,40% à 73,40%.

Nos simulations nous ont permis de confirmer la plupart des conclusions de Carignan. Si nous désirons utiliser une méthode de sélection S_e convergente, il est préférable d'utiliser le rééchantillonnage des résidus et l'intervalle de confiance percentile avec une pré-sélection bootstrap par une méthode de sélection S_b. En effet, puisqu'en général nous ne connaissons pas la valeur du rapport signal-bruit, les pourcentages de couverture obtenus par cette méthode sont toujours supérieurs ou égaux aux pourcentages de couverture obtenus par inférence classique. Nous concluons, comme Carignan, que les intervalles de confiance classiques sont plus courts lorsque le rapport signal-bruit est élevé, mais les pourcentages de couverture sont très faibles lorsque le rapport signal-bruit est faible. Nous ajoutons cependant aux conclusions de Carignan que l'utilisation du modèle complet pour construire les observations bootstrap est particulièrement néfaste lorsque les intervalles de confiance pivotal et bootstrap-t-MSE sont utilisés et ce, même pour un rapport signal-bruit élevé.

Si nous utilisons une méthode de sélection S_e non convergente, il est encore préférable d'utiliser le rééchantillonnage des résidus avec une pré-sélection bootstrap par une méthode de sélection S_b. L'intervalle de confiance bootstrap-t-MSE est légèrement supérieur aux deux autres intervalles de confiance bootstrap, mais tous les trois peuvent être utilisés. Pour un rapport signal-bruit élevé, les intervalles de confiance classiques sont plus courts que les intervalles de confiance bootstrap, mais les pourcentages de couverture des intervalles de confiance bootstrap sont très supérieurs aux pourcentages de couverture des intervalles de confiance classiques lorsque le rapport signal-bruit est faible. Certaines précautions sont de

TABLEAU 3.2.9. *Moyennes et écarts type des longueurs des intervalles de confiance pour une rapport signal-bruit faible pour les méthodes de sélection S_e BIC et CPM en combinaison avec S_b BIC, CPM et NON lorsque la méthode du rééchantillonnage des résidus est employée.*

Matrice de design X_{50}, $\sigma = 10$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC									
Classique	Moyenne	5,7340	1,3279	1,1770	0,5716	0,1307	0,2121	0,1898	0,2303
	Écart type	0,6041	2,1374	2,0667	1,9220	0,8031	1,0756	0,9914	1,1341
S_b:BIC									
Pivotal	Moyenne	5,6949	1,8598	1,6551	1,0034	0,5531	1,1000	0,5438	0,7177
& Percentile	Écart type	0,6455	2,4993	2,4691	2,4605	1,2643	1,8184	1,3833	1,6335
Bootstrap-t	Moyenne	6,0263	2,0536	1,8338	1,1193	0,6304	1,2169	0,6200	0,8156
	Écart type	0,6703	2,7350	2,7164	2,7110	1,4323	2,0093	1,5678	1,8332
S_b:CPM									
Pivotal	Moyenne	5,6488	3,2880	2,9893	2,5770	1,6018	2,4227	1,5350	2,0507
& Percentile	Écart type	0,6621	2,2961	2,3119	3,1246	1,9622	2,1839	1,9735	2,3002
Bootstrap-t	Moyenne	6,1738	3,8035	3,4374	3,0222	1,9074	2,8048	1,8062	2,4069
	Écart type	0,7340	2,6239	2,6461	3,6453	2,3461	2,5695	2,3233	2,7152
S_b:NON									
Pivotal	Moyenne	5,6600	4,2441	4,0961	5,0277	3,3096	3,9278	3,3112	3,9215
& Percentile	Écart type	0,6528	1,3290	1,4318	1,9941	1,3885	1,5220	1,5316	1,6920
Bootstrap-t	Moyenne	6,3076	5,0130	4,8350	6,0389	3,9943	4,7160	3,9690	4,7246
	Écart type	0,7380	1,5083	1,6185	2,3392	1,6660	1,8130	1,7931	1,9809
S_e CPM									
Classique	Moyenne	5,6871	2,6408	2,1360	1,7875	0,9857	1,0865	0,8203	1,0610
	Écart type	0,6107	2,3934	2,3977	3,0722	2,0279	2,2630	1,9405	2,2966
S_b:BIC									
Pivotal	Moyenne	5,8946	5,1366	5,1790	7,3601	5,2333	6,1007	5,4716	6,2512
& Percentile	Écart type	0,6745	0,6426	0,6408	0,8686	0,5773	0,6983	0,6116	0,6981
Bootstrap-t	Moyenne	6,1934	5,3880	5,4492	7,7134	5,4765	6,3962	5,7194	6,5484
	Écart type	0,6918	0,6791	0,6992	0,9002	0,6089	0,7397	0,6541	0,7342
S_b:CPM									
Pivotal	Moyenne	5,7594	4,9636	5,0062	7,1142	5,0831	5,9333	5,2754	6,0889
& Percentile	Écart type	0,6463	0,6417	0,6630	0,8737	0,5906	0,7183	0,6237	0,7094
Bootstrap-t	Moyenne	6,2789	5,4831	5,5232	7,8017	5,5609	6,4942	5,7543	6,6694
	Écart type	0,7026	0,6738	0,7074	0,9060	0,6279	0,7464	0,6674	0,7616
S_b:NON									
Pivotal	Moyenne	5,7375	4,6085	4,5605	6,3288	4,5511	5,2635	4,6025	5,4481
& Percentile	Écart type	0,6474	0,8446	0,8975	1,1771	0,8188	0,9907	0,8706	1,0134
Bootstrap-t	Moyenne	6,4148	5,3204	5,2673	7,3389	5,2471	6,0773	5,3138	6,2836
	Écart type	0,7293	0,8674	0,9411	1,1965	0,8318	1,0109	0,8908	1,0464

mises si aucune méthode de sélection pré-bootstrap n'est appliquée car pour un rapport signal-bruit faible, cette méthode est la seule à donner des pourcentages de couverture sous la valeur prescrite.

3.3. RÉÉCHANTILLONNAGE DES PAIRES D'OBSERVATIONS

Dans cette section, nous verrons tout d'abord le comportement des intervalles de confiance lorsque le rééchantillonnage des paires d'observations est utilisé. Nous verrons en détail les résultats pour un rapport signal-bruit élevé et faible pour être en mesure, par la suite, de les comparer avec les résultats obtenus lors de l'utilisation du rééchantillonnage des résidus. Cette comparaison sera graphique car pour l'étude des différents contrastes, nous possédons trop peu d'observations. La diminution des degrés de liberté ne nous permettrait pas de tirer des conclusions.

3.3.1. Rapport signal-bruit élevé.

Nous utiliserons, encore une fois, un écart type de 0,1 et de 10 avec le même vecteur $\underset{\sim}{\beta}$ qu'à la section précédente afin de comparer les résultats de chacune des méthodes de rééchantillonnage. Nous avons effectué la sélection de modèle pour chacune des 5 méthodes S_e exposées précédemment et ce sur la même matrice X_{50}.

Regardons les résultats pour la qualité du modèle choisi exposés dans le tableau 3.3.1. Les lignes "Originales" représentent les proportions de modèles biaisés, de vrais modèles et de modèles trop grands des 500 modèles sélectionnés à partir de la matrice X originale et du vecteur $\underset{\sim}{y}$ généré à chacune des répétitions. Les lignes "Bootstrap" représentent ces mêmes proportions pour les 500 000 modèles

bootstrap sélectionnés à partir des matrices X^* et des vecteurs $\underset{\sim}{y}^*$ choisis par rééchantillonnage des paires.

La proportion de vrais modèles et de modèles trop grands est similaire entre les méthodes de sélection S_e convergentes BIC et DUC. Une similitude est également observé entre les méthodes de sélection S_e non convergentes, BWD, CPM et FWD. Par la suite, nous exposerons en détail seulement les résultats pour les méthodes S_e BIC et S_e CPM, représentatives des méthodes convergentes et non convergentes.

TABLEAU 3.3.1. *Qualité de la sélection de modèle pour un rapport signal-bruit élevé effectuée sur les données originales ou bootstrap par rééchantillonnage des paires d'observations.*

Matrice de design X_{50}, $\sigma = 0{,}1$				
Sélection	Données	Vrai modèle	Modèle trop grand	Modèle biaisé
BIC	Originales	0,900	0,100	0,000
	Bootstrap	0,633	0,367	0,000
BWD	Originales	0,596	0,404	0,000
	Bootstrap	0,296	0,704	0,000
CPM	Originales	0,606	0,394	0,000
	Bootstrap	0,305	0,695	0,000
DUC	Originales	0,978	0,022	0,000
	Bootstrap	0,792	0,208	0,000
FWD	Originales	0,602	0,398	0,000
	Bootstrap	0,303	0,697	0,000

Nous pouvons observer dans le tableau 3.3.2 que l'inférence classique donne des pourcentages de couverture pour les intervalles de confiance qui ne sont pas différents de la valeur prescrite de 95%.

TABLEAU 3.3.2. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit élevé pour les méthodes de sélection S_e BIC et CPM lorsque la méthode du rééchantillonnage des paires d'observations est employée.*

Matrice de design X_{50}, $\sigma = 0{,}1$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC									
Classique	Uni. gauche 2,5%	2,40	2,60	3,20	4,00	1,20	1,80	1,60	2,20
	Uni. droite 2,5%	2,60	4,00	2,40	3,40	3,60	2,00	1,60	1,60
	Bilatéral 95%	95,00	93,40	94,40	92,60	95,20	96,20	96,80	96,20
Pivotal	Uni. gauche 2,5%	3,40	3,20	2,40	4,20	2,00	1,40	1,40	2,20
	Uni. droite 2,5%	2,00	3,00	2,40	2,80	3,40	1,60	1,40	1,60
	Bilatéral 95%	94,60	93,80	95,20	93,00	94,60	97,00	97,20	96,20
Percentile	Uni. gauche 2,5%	2,20	2,40	2,80	2,80	2,00	0,00	0,00	0,00
	Uni. droite 2,5%	2,60	2,60	2,00	2,60	3,00	0,00	0,00	0,00
	Bilatéral 95%	95,20	95,00	95,20	94,60	95,00	100,00	100,00	100,00
Bootstrap-t	Uni. gauche 2,5%	2,00	1,80	1,40	2,20	0,80	0,40	1,20	1,20
	Uni. droite 2,5%	0,80	2,40	1,20	1,80	1,60	1,00	1,40	1,20
	Bilatéral 95%	97,20	95,80	97,40	96,00	97,60	98,60	97,40	97,60
S_e CPM									
Classique	Uni. gauche 2,5%	2,40	2,60	3,60	4,40	1,40	2,40	1,80	2,40
	Uni. droite 2,5%	2,80	4,40	2,80	3,00	4,00	3,00	2,20	3,00
	Bilatéral 95%	94,80	93,00	93,60	92,60	94,60	94,60	96,00	94,60
Pivotal	Uni. gauche 2,5%	3,40	3,00	2,60	4,20	1,80	2,80	1,60	2,40
	Uni. droite 2,5%	2,40	3,00	2,20	2,80	3,20	1,40	1,80	2,60
	Bilatéral 95%	94,20	94,00	95,20	93,00	95,00	95,80	96,60	95,00
Percentile	Uni. gauche 2,5%	1,60	2,20	2,80	3,20	1,60	0,00	0,00	0,20
	Uni. droite 2,5%	2,80	2,40	2,00	2,80	2,80	0,00	0,00	0,00
	Bilatéral 95%	95,60	95,40	95,20	94,00	95,60	100,00	100,00	99,80
Bootstrap-t	Uni. gauche 2,5%	2,00	2,20	0,80	2,20	1,00	1,00	1,00	0,80
	Uni. droite 2,5%	1,20	2,00	1,80	1,60	1,80	1,00	1,40	1,40
	Bilatéral 95%	96,80	95,80	97,40	96,20	97,20	98,00	97,60	97,80

Nous notons également que les 3 intervalles de confiance bootstrap fonctionnent très bien. Pour certains coefficients cependant, le pourcentage de couverture est trop élevé nous indiquant que l'intervalle de confiance est trop conservateur. C'est le cas notamment pour certains coefficients non nuls de l'intervalle de confiance bootstrap-t-MSE. Pour les coefficients nuls, un pourcentage de

couverture près de 100% n'est pas nécessairement l'indicateur d'un intervalle de confiance trop conservateur, mais plutôt un indicateur de la qualité de la sélection de modèle.

Regardons de plus près la longueur des intervalles de confiance pour un rapport signal-bruit élevé. Nous notons dans le tableau 3.3.3 que bien qu'il n'y ait pas de différence entre les longueurs moyennes des intervalles de confiance classiques et bootstrap pour les coefficients non nuls, les longueurs des intervalles de confiance bootstrap sont au moins 5 fois plus larges que les intervalles de confiance classiques pour les coefficients nuls dans le cas de la méthode S_e CPM et au moins 20 fois plus larges dans le cas de la méthode S_e BIC. Puisque les pourcentages de couverture des intervalles de confiance classiques étaient déjà près de la valeur prescrite, nous concluons qu'il est préférable d'utiliser l'inférence classique au rééchantillonnage des paires d'observations pour un rapport signal-bruit élevé.

Toujours dans le tableau 3.3.3, nous pouvons comparer les longueurs moyennes des intervalles de confiance bootstrap entre eux. Comme on doit s'y attendre, l'intervalle de confiance bootstrap-t-MSE est plus large que les deux autres. Puisque les pourcentages de couverture sont similaires et qu'un pourcentage de couverture de 100% pour un coefficient nul n'est pas nécessairement une indication d'un intervalle de confiance trop conservateur, nous concluons que parmi les intervalles de confiance bootstrap, le meilleur semble être l'intervalle de confiance percentile mais la différence avec les 2 autres intervalles de confiance bootstrap reste très faible.

3.3.2. Rapport signal-bruit faible.

Analysons maintenant les résultats obtenus en utilisant un écart type de 10. Comme l'avait noté Carignan, il est beaucoup plus difficile de sélectionner le

TABLEAU 3.3.3. *Moyennes et écarts type des longueurs des intervalles de confiance pour un rapport signal-bruit élevé pour les méthodes de sélection S_e BIC et CPM lorsque la méthode du rééchantillonnage des paires d'observations est employée.*

Matrice de design X_{50}, $\sigma = 0{,}1$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC									
Classique	Moyenne	0,0612	0,0502	0,0494	0,0721	0,0536	0,0023	0,0017	0,0024
	Écart type	0,0065	0,0053	0,0052	0,0076	0,0057	0,0116	0,0093	0,0120
Pivotal	Moyenne	0,0629	0,0535	0,0538	0,0780	0,0568	0,0440	0,0390	0,0456
& Percentile	Écart type	0,0077	0,0089	0,0078	0,0122	0,0094	0,0180	0,0162	0,0199
Bootstrap-t	Moyenne	0,0716	0,0610	0,0615	0,0889	0,0645	0,0515	0,0458	0,0532
	Écart type	0,0091	0,0110	0,0096	0,0152	0,0114	0,0215	0,0192	0,0234
S_e CPM									
Classique	Moyenne	0,0610	0,0502	0,0496	0,0723	0,0535	0,0102	0,0081	0,0103
	Écart type	0,0065	0,0053	0,0054	0,0077	0,0057	0,0233	0,0197	0,0237
Pivotal	Moyenne	0,0638	0,0545	0,0553	0,0798	0,0578	0,0586	0,0526	0,0622
& Percentile	Écart type	0,0078	0,0091	0,0080	0,0129	0,0095	0,0135	0,0118	0,0138
Bootstrap-t	Moyenne	0,0728	0,0624	0,0631	0,0913	0,0660	0,0678	0,0609	0,0720
	Écart type	0,0091	0,0111	0,0097	0,0161	0,0115	0,0157	0,0138	0,0162

vrai modèle lorsque le rapport signal-bruit est faible. En effet, la proportion de modèles biaisés parmi les modèles choisis à partir des jeux de données originaux ou des jeux de données bootstrap est d'au moins 95% pour chacune des méthodes de sélection S_e.

Puisque la majorité des modèles sont biaisés, les pourcentages de couverture des intervalles de confiance sont inférieurs à la valeur prescrite de 95%. A titre d'exemple, nous allons regarder de plus près les résultats obtenus pour les méthodes de sélection S_e BIC et CPM respectivement représentatives des méthodes convergentes et non convergentes. Les résultats sont exposés dans le tableau 3.3.4.

Nous pouvons immédiatement séparer les pourcentages de couverture des intervalles de confiance selon que le vrai coefficient est ou n'est pas nul. Pour les coefficients nuls, les pourcentages de couverture des intervalles de confiance

TABLEAU 3.3.4. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit faible pour les méthodes de sélection S_e BIC et CPM lorsque la méthode du rééchantillonnage des paires d'observations est employée.*

Matrice de design X_{50}, σ =10,00									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC									
Classique	Uni. gauche 2,5%	3,80	3,40	4,20	3,80	1,60	3,40	2,40	3,80
	Uni. droite 2,5%	2,80	71,80	75,40	92,20	98,20	0,40	1,20	0,20
	Bilatéral 95%	93,40	24,80	20,40	4,00	0,20	96,20	96,40	96,00
Pivotal	Uni. gauche 2,5%	3,20	3,80	4,40	3,20	1,00	3,00	2,20	3,60
	Uni. droite 2,5%	2,20	64,80	65,80	71,60	62,00	0,20	1,20	0,20
	Bilatéral 95%	94,60	31,40	29,80	25,20	37,00	96,80	96,60	96,20
Percentile	Uni. gauche 2,5%	1,80	0,60	0,20	0,00	0,00	0,00	0,00	0,00
	Uni. droite 2,5%	2,40	8,80	10,80	25,80	41,20	0,00	0,00	0,00
	Bilatéral 95%	95,80	90,60	89,00	74,20	58,80	100,00	100,00	100,00
Bootstrap-t	Uni. gauche 2,5%	2,80	3,00	2,80	2,00	0,60	2,20	2,20	3,00
	Uni. droite 2,5%	1,40	64,60	65,80	71,60	62,00	0,20	0,80	0,20
	Bilatéral 95%	95,80	32,40	31,40	26,40	37,40	97,60	97,00	96,80
S_e CPM									
Classique	Uni. gauche 2,5%	4,20	3,00	4,40	3,80	1,80	5,00	3,40	4,80
	Uni. droite 2,5%	3,80	44,40	55,80	75,80	85,00	3,20	2,20	2,80
	Bilatéral 95%	92,00	52,60	39,80	20,40	13,20	91,80	94,40	92,40
Pivotal	Uni. gauche 2,5%	3,00	3,80	4,00	4,20	2,20	4,80	4,00	3,40
	Uni. droite 2,5%	2,80	27,00	27,00	28,60	24,20	1,60	2,00	1,80
	Bilatéral 95%	94,20	69,20	69,00	67,20	73,60	93,60	94,00	94,80
Percentile	Uni. gauche 2,5%	1,80	1,60	1,40	0,20	0,00	0,00	0,00	0,00
	Uni. droite 2,5%	3,20	2,60	3,60	9,20	17,20	0,00	0,00	0,00
	Bilatéral 95%	95,00	95,80	95,00	90,60	82,80	100,00	100,00	100,00
Bootstrap-t	Uni. gauche 2,5%	1,80	2,80	2,00	2,60	1,20	2,80	2,40	2,20
	Uni. droite 2,5%	1,80	24,00	25,80	28,60	22,20	0,80	0,60	1,20
	Bilatéral 95%	96,40	73,20	72,20	68,80	76,60	96,40	97,00	96,60

classiques et bootstrap ne sont pas différents de 95% sauf pour l'intervalle de confiance percentile. Cependant, puisque la longueur moyenne de ces intervalles est inférieure ou égale à celle des autres intervalles de confiance bootstrap, un pourcentage de couverture de 100% ne constitue pas un problème.

Pour les coefficients non nuls ou presque nul, à l'exception de la constante β_0, la vraie valeur du coefficient est fréquemment à droite de l'intervalle de confiance, c'est-à-dire que l'estimé du coefficient sous-estime la vraie valeur. Les pourcentages de couverture des intervalles de confiance classiques bilatéraux sont très inférieurs à 95%. Les méthodes convergentes ont également des pourcentages de couverture inférieurs aux méthodes non convergentes.

Pour toutes les méthodes, les intervalles de confiance bootstrap performent mieux que les intervalles de confiance classiques. Parmi les intervalles de confiance bootstrap, la meilleure performance est celle de l'intervalle de confiance percentile et ce, pour toutes les méthodes de sélection. En particulier pour la méthode de sélection S_e CPM, 3 des 5 coefficients non nuls ont des pourcentages de couverture de l'intervalle de confiance percentile qui ne sont pas différents de la valeur prescrite de 95% contrairement à un seul coefficient, la constante, pour l'intervalle de confiance bootstrap-t-MSE. Nous pouvons également constater que pour la méthode de sélection S_e BIC, l'intervalle de confiance percentile couvre de 21% à 58% plus souvent la vraie valeur des coefficients β_1 à β_4 que l'intervalle de confiance bootstrap-t-MSE. Pour la méthode S_e CPM, la supériorité des pourcentages de couverture de l'intervalle de confiance percentile varie entre 6% et 22%.

Nous concluons que pour un rapport signal-bruit faible, il est préférable d'utiliser la méthode du rééchantillonnage des paires d'observations à la méthode classique pour construire les intervalles de confiance et ce pour toutes les méthodes de sélection. Bien que certains coefficients non nuls ou presque nul ont un pourcentage de couverture inférieur à la valeur prescrite de 95%, l'intervalle de confiance percentile reste le meilleur choix parmi les intervalles de confiance bootstrap.

3.3.3. Supériorité de l'intervalle de confiance percentile par rééchantillonnage des paires d'observations

Nous avons noté une différence importante entre les pourcentages de couverture des intervalles de confiance percentile et pivotal lorsque le rééchantillonnage des paires est employé. Tentons d'expliquer ce phénomène en regardant de plus près les intervalles de confiance.

La figure 3.3.1 illustre les 25 premiers intervalles de confiance pivotal, percentile et bootstrap-t-MSE pour le coefficient β_4 lorsque nous utilisons le rééchantillonnage des paires d'observations avec un écart type de 10 et en sélectionnant le modèle avec la méthode S_e BIC. La vraie valeur du coefficient β_4 est indiquée par une ligne verticale. Clairement, les intervalles de confiance percentile couvrent davantage la vraie valeur que les autres intervalles. En effet, le pourcentage de couverture pour le coefficient β_4 est de 58,80% pour l'intervalle de confiance percentile et de 37,00% pour l'intervalle de confiance pivotal.

La variable X_4 est incluse dans le modèle original uniquement 13 fois sur les 500 modèles sélectionnés. La différence se situe lorsque la variable X_4 n'est pas sélectionnée. Ces 487 intervalles de confiance percentile sont répartis de la façon suivante: 50 sont l'ensemble $\{0\}$, 256 ont une borne inférieure nulle, 152 une borne supérieure nulle et pour 29 de ces intervalles aucune des 2 bornes n'est nulle. Puisque $\hat{\beta}_4 = 0$, l'intervalle de confiance pivotal est l'image miroir par rapport à 0 de l'intervalle de confiance percentile. En effet, lorsque $\hat{\beta}_4 = 0$, l'intervalle de confiance percentile devient

$$\left[\hat{\beta}^*_{4(25)}, \hat{\beta}^*_{4(975)}\right]$$

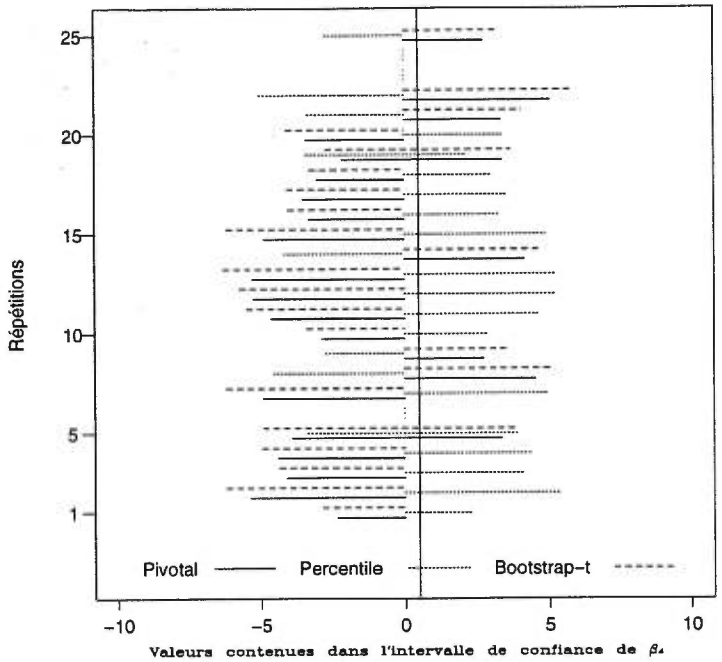

FIGURE 3.3.1. *25 premiers intervalles de confiance pivotal, percentile et bootstrap-t-MSE de β_4 pour un rapport signal-bruit faible, en utilisant la méthode de sélection S_e BIC lorsque le rééchantillonnage des paires d'observations est employé.*

et l'intervalle de confiance pivotal devient

$$\left[-\hat{\beta}^*_{4(975)}, -\hat{\beta}^*_{4(25)}\right]$$

lorsque nous utilisons 1000 répétitions bootstrap avec $\hat{\beta}^*_{4(\delta)}$ la δ-ième statistique d'ordre des 1000 coefficients estimés. Par conséquent, 152 intervalles de confiance pivotal ont une borne inférieure nulle et 256 ont une borne supérieure nulle.

Une borne est nulle lorsque la variable n'est pas suffisamment sélectionnée par rééchantillonnage des paires et que le poids de la distribution de $\hat{\beta}^*_4$ est essentiellement à gauche ($\hat{\beta}^*_{4(975)} = 0$) ou à droite ($\hat{\beta}^*_{4(25)} = 0$) de zéro. Notons que lorsque la variable est sélectionnée, son estimation est éloignée de 0 car l'hypothèse d'égalité

à 0 a essentiellement été rejetée lors de la sélection. La figure 3.3.2 nous montre une distribution type de $\hat{\beta}_4^*$ pour une répétition par rééchantillonnage des paires.

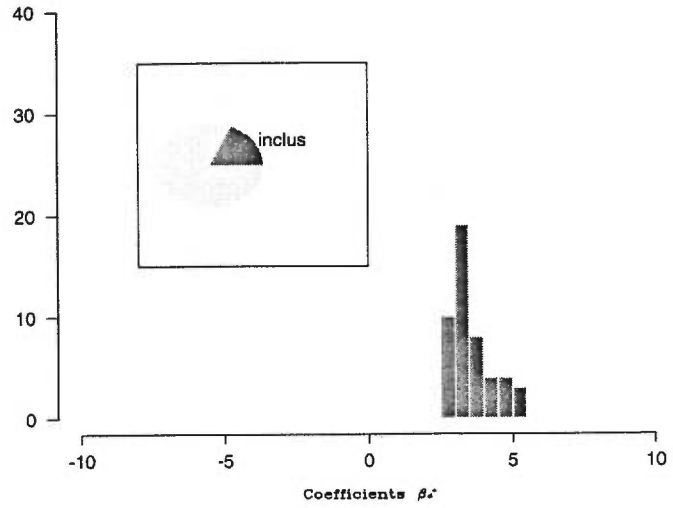

FIGURE 3.3.2. *Distribution des 1000 estimés bootstrap $\hat{\beta}_4^*$ de la $27^{ième}$ répétition pour un rapport-signal bruit faible lorsque le rééchantillonnage des paires d'observations avec la méthode de sélection S_e BIC est employé. La proportion de la variable X_4 incluse dans le modèle est indiquée par le graphique circulaire. L'histogramme montre la distribution de ces coefficients inclus dans le modèle ($\hat{\beta}_4^* \neq 0$).*

Pourquoi la distribution est-elle plus souvent à droite de zéro? La vraie valeur de ce coefficient est égale à 0,5 et nous rééchantillonnons des observations construites à partir de ce coefficient. L'estimé de $\hat{\beta}_4$ devrait être davantage positif. Par conséquent, l'intervalle de confiance percentile couvrira davantage la vraie valeur de β_4 que l'intervalle de confiance pivotal.

3.3.4. Comparaison entre le rééchantillonnage des paires d'observations et des résidus.

Débutons notre comparaison des 2 méthodes de rééchantillonnage vues jusqu'à présent en mettant en parallèle les résultats obtenus sur la qualité de la sélection de modèle. En observant de plus près les tableaux 3.2.6 et 3.3.1, nous sommes en mesure d'affirmer que pour un rapport signal-bruit élevé, le rééchantillonnage des résidus sélectionne davantage le vrai modèle que la méthode du réechantillonnage des paires d'observations. En effet, pour une méthode de sélection S_e donnée, seule l'utilisation du modèle complet afin de générer les observations bootstrap (S_b NON) par rééchantillonnage des résidus, donne des proportions de vrais modèles bootstrap similaires à celles données par rééchantillonnage des paires d'observations. Toutes les autres combinaisons S_b et S_e entraînent des proportions de vrais modèles plus élevées. Pour un rapport signal-bruit faible, peu importe la méthode de rééchantillonnage utilisée, la proportion de modèles biaisés reste supérieure à 95%.

Comparons maintenant les pourcentages de couverture des différents intervalles de confiance. Pour ce faire, observons la figure 3.3.3 qui illustre les pourcentages de couverture des intervalles de confiance bootstrap par rééchantillonnage des paires et des résidus et un rapport signal-bruit élevé et faible. Pour chaque graphique, la lettre "p" indique le pourcentage de couverture pour les intervalles de confiance bootstrap obtenus à partir de la méthode du rééchantillonnage des paires d'observations, la lettre "c" ceux obtenus à partir du rééchantillonnage des résidus en utilisant la méthode de sélection S_b CPM et la lettre "b" ceux obtenus en utilisant la méthode de sélection S_b BIC. Pour chaque graphique, les 3 premières colonnes de lettres sont les pourcentages de couverture pour les intervalles de confiance du coefficient β_3, les 3 suivantes du coefficient β_4 et les 3 dernières du

coefficient β_5 afin d'illustrer les résultats pour un coefficient clairement non nul, presque nul et nul. Pour chaque coefficient, la première des 3 colonnes de lettres représente l'intervalle de confiance pivotal, la seconde l'intervalle de confiance percentile et la troisième l'intervalle de confiance bootstrap-t-MSE.

Par exemple, pour le graphique intitulé "S_e BIC, sigma = 10", l'ensemble des 9 lettres disposées en 3 colonnes au-dessus du coefficient β_4 représentent les pourcentages de couverture des intervalles de confiance obtenus par rééchantillonnage des paires et des résidus. En particulier, la lettre "c" de la seconde colonne au-dessus du coefficient β_4 représente le pourcentage de couverture de l'intervalle de confiance percentile obtenu par rééchantillonnage des résidus en utilisant la combinaison S_b CPM et S_e BIC afin de sélectionner le modèle.

Pour un rapport signal-bruit élevé, il n'y a pas de différence importante entre les pourcentages de couverture des méthodes de rééchantillonnage, des types d'intervalles de confiance bootstrap et des coefficients. Cependant, comme nous le montre le tableau 3.3.5, les intervalles de confiance par rééchantillonnage des résidus sont plus courts que les intervalles de confiance par rééchantillonnage des paires d'observations, particulièrement pour les coefficients nuls des méthodes de sélection S_e convergentes.

Pour toutes ces raisons, nous jugeons préférable d'utiliser le rééchantillonnage des résidus au rééchantillonnage des paires d'observations parmi les méthodes de rééchantillonnage. Ce choix est particulièrement important s'il s'agit de méthodes de sélection S_e convergentes car la différence entre les longueurs moyennes des intervalles de confiance est dans ce cas plus importante. Il faut cependant souligner que les deux méthodes de rééchantillonnage sont bonnes.

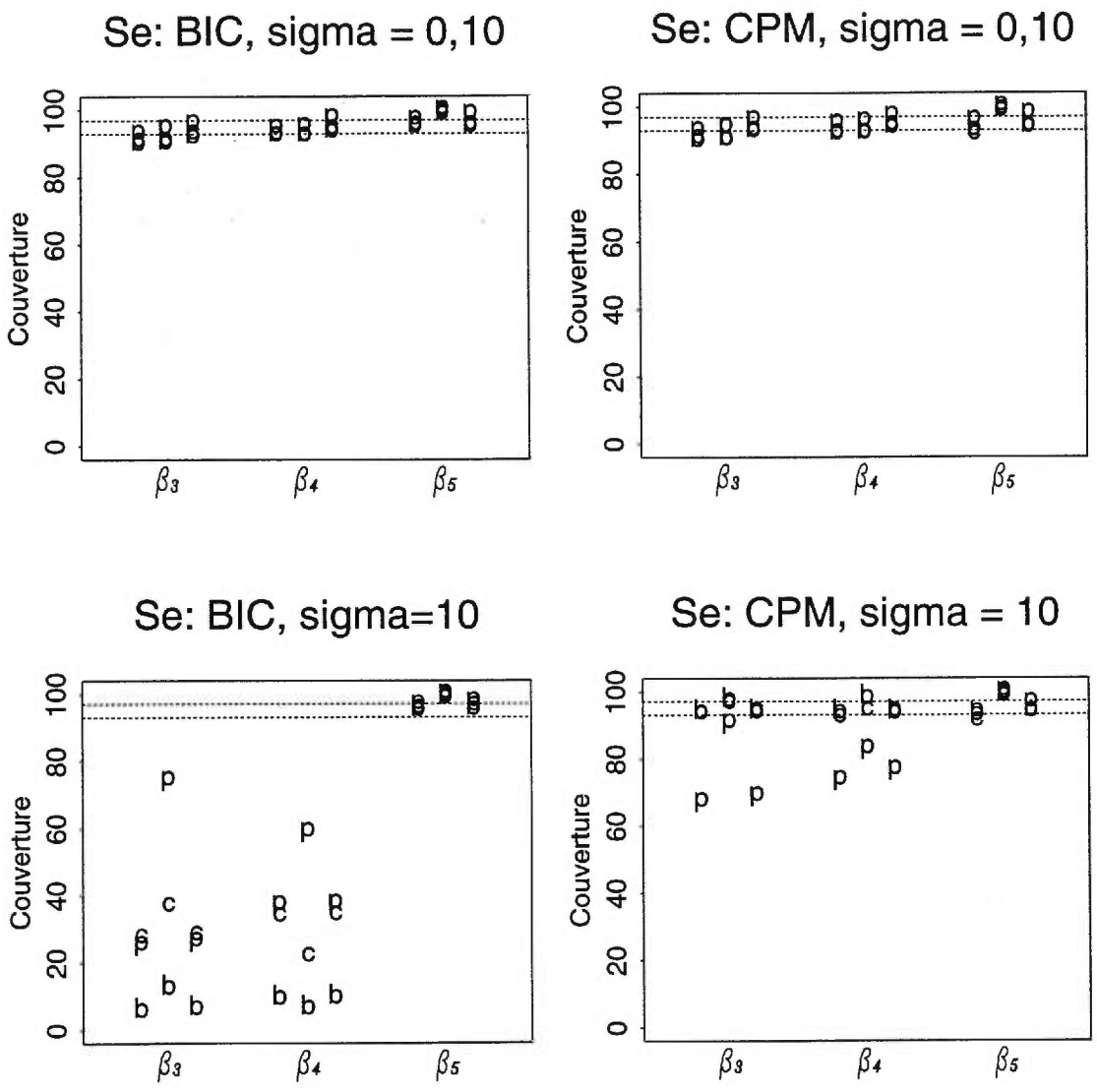

FIGURE 3.3.3. *Pourcentages de couverture des intervalles de confiance bootstrap des coefficients β_3, β_4 et β_5 pour les méthodes de sélection S_e BIC et CPM lorsque les méthodes du rééchantillonnage des paires, "p", et du rééchantillonnage des résidus avec les méthodes de sélection S_b BIC, "b", et S_b CPM, "c", sont employées. Pour chaque coefficient, la première des 3 colonnes est le pourcentage de couverture de l'intervalle de confiance pivotal, la seconde de l'intervalle de confiance percentile et la troisième de l'intervalle de confiance bootstrap-t-MSE.*

TABLEAU 3.3.5. *Rapports entre les longueurs moyennes des intervalles de confiance percentile par rééchantillonnage des paires d'observations et des résidus pour un rapport signal-bruit élevé.*

Matrice de design X_{50}, $\sigma = 0{,}10$								
Coefficients	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes	1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC								
Paires et Résidus S_b BIC	1,10	1,14	1,15	1,14	1,13	5,50	6,09	5,77
Paires et Résidus S_b CPM	1,11	1,14	1,14	1,14	1,14	3,01	3,45	3,14
S_e CPM								
Paires et Résidus S_b BIC	1,10	1,14	1,14	1,14	1,14	0,98	1,00	1,00
Paires et Résidus S_b CPM	1,11	1,14	1,14	1,15	1,15	0,99	1,01	1,02

Il en va autrement pour un rapport signal-bruit faible. Tout d'abord par la méthode de rééchantillonnage des paires d'observations, l'intervalle de confiance percentile couvre davantage la vraie valeur du coefficient que les autres intervalles de confiance bootstrap. Par rééchantillonnage des résidus, le meilleur intervalle de confiance bootstrap est l'intervalle de confiance bootstrap-t-MSE. De plus, pour une méthode de sélection S_e convergente, BIC en l'occurence, le pourcentage de couverture pour les coefficients non nuls est supérieur par la méthode paires par rapport à la méthode résidus. En effet, les lettres "p" pour les coefficients β_3 et β_4 du graphique intitulé "S_e BIC, sigma = 10" de la figure 3.3.3 sont situées au-dessus des lettres "b" et "c" pour une même colonne. Pour une méthode de sélection S_e non convergente, CPM, nous avons le résultat contraire; le pourcentage de couverture des coefficients non nuls est supérieur par la méthode résidus par rapport à la méthode paires.

Le tableau 3.3.6 nous montre les rapports des longueurs moyennes entre les intervalles de confiance par rééchantillonnage des paires d'observations et des résidus. Pour une méthode S_e convergente, la différence entre les longueurs moyennes est importante, mais les pourcentages de couverture des intervalles de confiance

par rééchantillonnage des paires d'observations sont très supérieurs à ceux par rééchantillonnage des résidus. Ces derniers sont tout simplement trop courts. Pour une méthode de sélection S_e non convergente, il n'y a pas de différence entre les longueurs moyennes des intervalles de confiance par rééchantillonnage des paires et des résidus.

TABLEAU 3.3.6. *Rapports entre les longueurs moyennes des intervalles de confiance percentile par rééchantillonnage des paires d'observations et des résidus pour un rapport signal-bruit faible.*

Matrice de design X_{50}, $\sigma = 10{,}00$								
Coefficients	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes	1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC								
Paires et Résidus S_b BIC	1,11	2,46	2,67	5,48	6,80	4,01	7,06	6,31
Paires et Résidus S_b CPM	1,11	1,39	1,48	2,13	2,34	1,82	2,50	2,21
S_e CPM								
Paires et Résidus S_b BIC	1,09	1,00	0,99	0,98	1,00	0,97	0,96	1,01
Paires et Résidus S_b CPM	1,11	1,03	1,02	1,02	1,03	1,00	1,00	1,03

A quoi devons-nous ces différences? Pourquoi le rééchantillonnage des paires donne-t-il de meilleurs résultats que le rééchantillonnage des résidus pour une méthode de sélection convergente avec un rapport signal-bruit faible, mais pas pour une méthode de sélection non convergente? Observons de plus près les intervalles de confiance afin de répondre à ces questions.

Attardons-nous tout d'abord à la méthode de sélection S_e convergente BIC. Pour mieux comprendre le passage d'une situation où les 2 méthodes de rééchantillonnage performent de la même façon ($\sigma = 0, 10$) à une situation où l'intervalle de confiance percentile par rééchantillonnage des paires d'observations est supérieur aux intervalles de confiance bootstrap par rééchantillonnage des résidus ($\sigma = 10$), nous avons simulé certains cas intermédiaires.

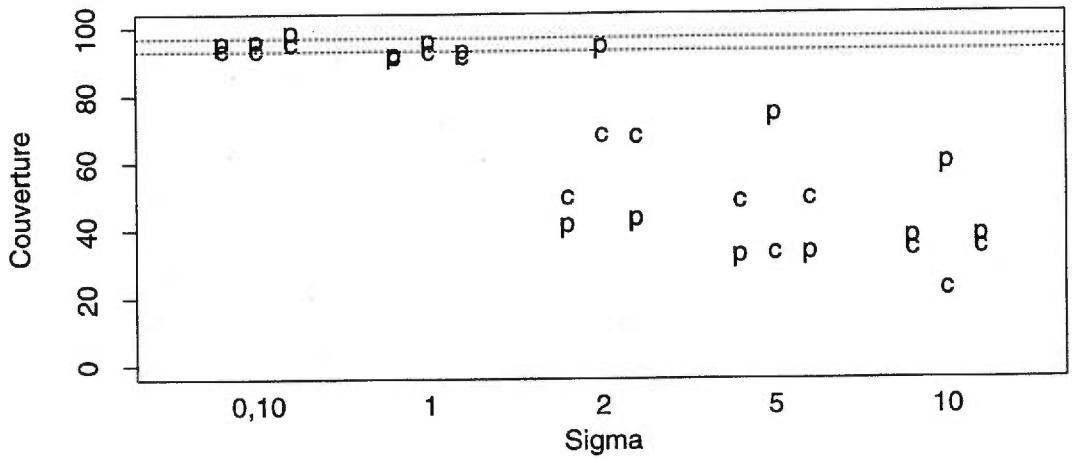

FIGURE 3.3.4. *Pourcentages de couverture des intervalles de confiance bootstrap du coefficient β_4 pour la méthode S_e BIC lorsque le rééchantillonnage des paires "p" et la combinaison S_b CPM, S_e BIC lorsque le rééchantillonnage des résidus, "c", sont employés avec différentes valeurs de bruit. Pour chaque écart type, la première des 3 colonnes est le pourcentage de couverture de l'intervalle de confiance pivotal, la seconde de l'intervalle de confiance percentile et la troisième de l'intervalle de confiance bootstrap-t-MSE.*

La figure 3.3.4 nous montre les pourcentages de couverture des différents intervalles de confiance bootstrap pour le coefficient β_4 et différentes valeurs de bruit. Les trois premières colonnes au-dessus de 0,10 sont les pourcentages de couverture pour un écart type de 0,10, les trois suivantes un écart type de 1 et ainsi de suite jusqu'aux trois dernières qui indiquent un écart type de 10. Comme précédemment, la lettre "p" indique le pourcentage de couverture des intervalles de confiance obtenus par rééchantillonnage des paires d'observations et la lettre "c" par rééchantillonnage des résidus en utilisant la méthode de sélection S_b CPM. Pour un écart type donné, la première des trois colonnes est l'intervalle

de confiance pivotal, la seconde l'intervalle de confiance percentile et la troisième l'intervalle de confiance bootstrap-t-MSE.

A mesure que l'écart type augmente, l'écart entre le pourcentage de couverture de l'intervalle de confiance percentile par rééchantillonnage des paires et les autres intervalles de confiance s'accentue. Le même phénomène se produit pour les autres coefficients non nuls, mais il débute pour un éacrt type plus élevé. On peut également noter un fait qui n'était pas apparent avant de considérer ces situations intermédiaires. Les intervalles de confiance pivotal et bootstrap-t-MSE obtenus par rééchantillonnage des résidus couvrent davantage la vraie valeur du coefficient β_4 que leurs pendants obtenus par rééchantillonnage des paires pour des écart types intermédiaires. Nous n'avions pu noter précédemment qu'une similitude entre les pourcentages de couverture de ces intervalles pour un rapport signal-bruit élevé et faible.

Les distributions des estimés $\hat{\beta}_4^*$ pour une répétition particulière explique ce phénomène. Débutons par regarder parmi les 500 intervalles de confiance obtenus par rééchantillonnage des paires d'observations et des résidus le nombre d'intervalles qui sont l'ensemble $\{0\}$, le nombre de quantiles bootstrap supérieurs nuls $(\hat{\beta}^*_{4(975)})$ et le nombre de quantiles bootstrap inférieurs nuls $(\hat{\beta}^*_{4(25)})$ pour chacune des valeurs de bruit. Ces statistiques sont exposées dans le tableau 3.3.7.

Notons que les ensembles $\{0\}$ incluent les intervalles de confiance percentile dont l'une ou l'autre des bornes est nulle. A mesure que le bruit augmente, la fréquence des intervalles de confiance de la forme $\{0\}$ augmentent drastiquement par rééchantillonnage des résidus. Ces ensembles $\{0\}$ sont obtenus lorsque pas plus de 2,5% des estimés $\hat{\beta}_4^*$ pour une répétition sont différents de 0 et ce à gauche et à droite de la distribution de $\hat{\beta}_4^*$. Cette augmentation explique en grande partie

TABLEAU 3.3.7. *Fréquences des intervalles de confiance* $\{0\}$, *dont la borne supérieure est nulle* ($\hat{\beta}^*_{4(975)} = 0$) *ou la borne inférieure est nulle* ($\hat{\beta}^*_{4(25)} = 0$) *parmi les 500 intervalles de confiance percentile pour différentes valeurs de bruit lorsque la méthode de sélection* S_e *BIC est employée par rééchantillonnage des paires d'observations ou des résidus.*

Matrice de design X_{50} S_e BIC, Coefficient β_4			
		Paires	Résidus (S_b CPM)
$\sigma = 0,10$	$\{0\}$	0	0
	$\hat{\beta}^*_{4(975)} = 0$	0	0
	$\hat{\beta}^*_{4(25)} = 0$	0	0
$\sigma = 1$	$\{0\}$	0	4
	$\hat{\beta}^*_{4(975)} = 0$	0	4
	$\hat{\beta}^*_{4(25)} = 0$	377	285
$\sigma = 2$	$\{0\}$	10	118
	$\hat{\beta}^*_{4(975)} = 0$	25	140
	$\hat{\beta}^*_{4(25)} = 0$	476	465
$\sigma = 5$	$\{0\}$	29	210
	$\hat{\beta}^*_{4(975)} = 0$	133	335
	$\hat{\beta}^*_{4(25)} = 0$	365	369
$\sigma = 10$	$\{0\}$	50	276
	$\hat{\beta}^*_{4(975)} = 0$	206	388
	$\hat{\beta}^*_{4(25)} = 0$	315	388

la différence entre les pourcentages de couverture des intervalles de confiance percentile par rééchantillonnage des paires d'observations et des résidus.

Nous pouvons également noter qu'un grand nombre d'intervalles de confiance dont l'une ou l'autre des bornes est nulle indique une distribution des $\hat{\beta}^*_4$ pour une répétition dont le poids est essentiellement à gauche ou à droite de 0. Ce phénomène expliquait déjà pourquoi l'intervalle de confiance percentile par rééchantillonnage des paires d'observations était supérieur à l'intervalle de confiance pivotal par rééchantillonnage des paires d'observations (section 3.3.3).

La distribution, pour une répétition, de $\hat{\beta}_4^*$ est différente lorsque nous employons le rééchantillonnage des résidus. La distribution de $\hat{\beta}_4^*$ pour une répétition obtenue par rééchantillonnage des paires est généralement centrée sur la vraie valeur de β_4. Par rééchantillonnage des résidus, la vraie valeur de ce coefficient pour les données bootstrap est $\tilde{\beta}_4$. La distribution se trouve alors centrée sur cette valeur. Lorsque le rapport signal-bruit est faible, nous obtenons de nombreux cas où $\tilde{\beta}_4 = 0$. La distribution de $\hat{\beta}_4^*$ est alors centrée, de façon erronée, à 0. Lorsque le rapport signal-bruit est plus élevé, particulièrement si la méthode de sélection S_b ne converge pas, nous obtenons davantage de $\tilde{\beta}_4$ différents de 0. Ces cas expliquent le meilleur pourcentage de couverture de l'intervalle de confiance pivotal par rééchantillonnage des résidus sur son pendant obtenu par rééchantillonnage des paires.

La figure 3.3.5 nous montre les 1000 coefficients bootstrap $\hat{\beta}_4^*$ estimés lors d'une répétition pour chacune des deux méthodes de rééchantillonnage. Nous voyons bien que par rééchantillonnage des paires tout le poids des coefficients non nuls est à droite de la distribution. Par rééchantillonnage des résidus, la distribution de $\hat{\beta}_4^*$ est davantage symétrique par rapport à 0, mais la proportion d'inclusion de la variable ne permet pas d'avoir un quantile différent de 0.

Pour une méthode de sélection S_e non convergente, CPM, les distributions de $\hat{\beta}_4^*$ ont la même forme que pour une méthode de sélection convergente, mais la sélection de la variable y est supérieure. La figure 3.3.6 montre la distribution pour une répétition de $\hat{\beta}_4^*$ par rééchantillonnage des paires et des résidus.

La diminution du nombre de variables non sélectionnées ne permet pas d'obtenir un quantile nul, particulièrement pour le rééchantillonnage des résidus. En

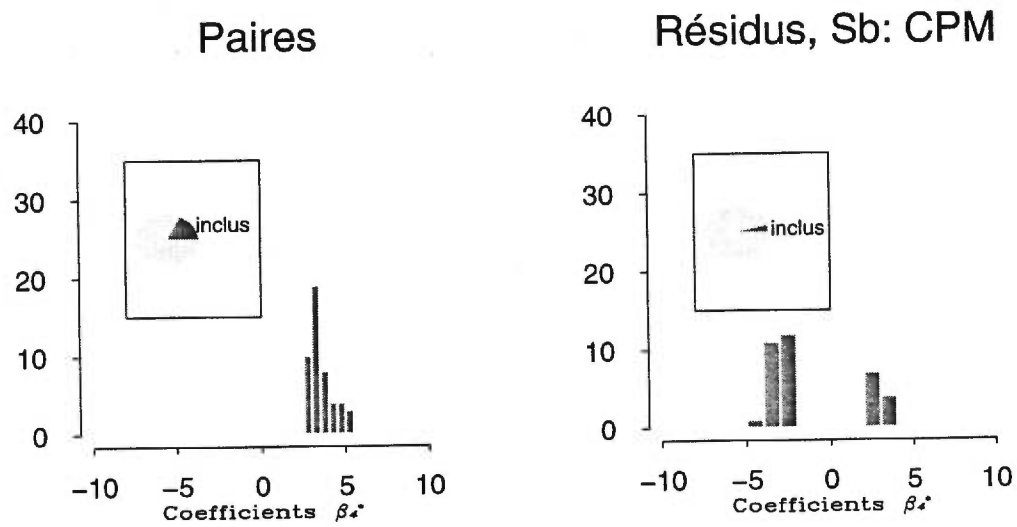

FIGURE 3.3.5. *Distribution des 1000 estimés bootstrap $\hat{\beta}_4^*$ de la $27^{ième}$ répétition pour un rapport-signal bruit faible lorsque le rééchantillonnage des paires d'observations avec la méthode de sélection S_e BIC (à gauche) ou le rééchantillonnage des résidus avec la combinaison S_b CPM et S_e BIC (à droite) sont employés. La proportion de la variable X_4 incluse dans le modèle est indiquée par le graphique circulaire. L'histogramme montre la distribution de ces coefficients inclus dans le modèle ($\hat{\beta}_4^* \neq 0$).*

effet, même si $\tilde{\beta}_4$ est nulle, la méthode de sélection S_e CPM sélectionne suffisamment de variables pour obtenir une distribution dont les quantiles 0,025 et 0,975 sont différents de 0. Parmi les 500 intervalles de confiance percentile, nous ne

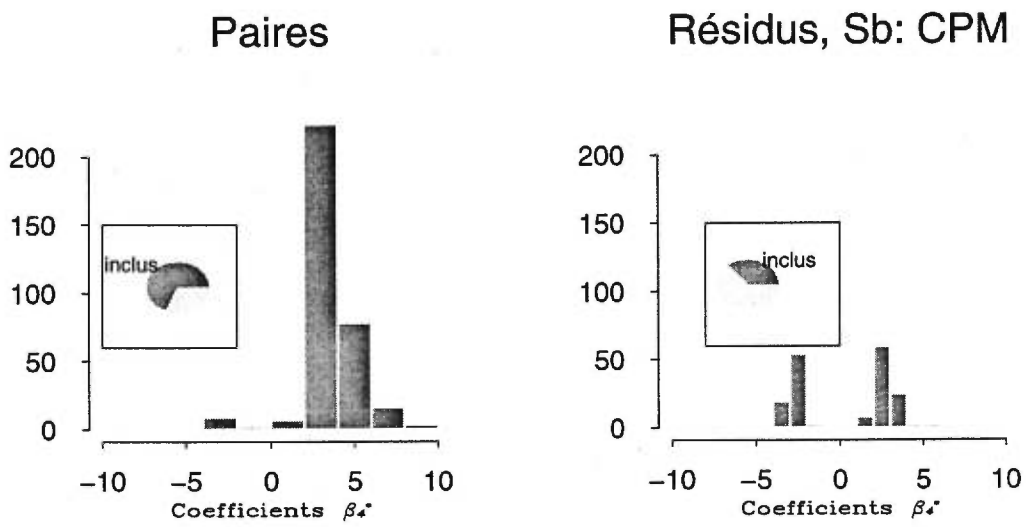

FIGURE 3.3.6. *Distribution des 1000 estimés bootstrap $\hat{\beta}_3^*$ de la seconde répétition pour un rapport-signal bruit faible lorsque le rééchantillonnage des paires d'observations avec la méthode de sélection S_e CPM (à gauche) ou le rééchantillonnage des résidus avec la combinaison S_b CPM et S_e CPM (à droite) sont employés. La proportion de la variable X_4 incluse dans le modèle est indiquée en foncé dans le graphique circulaire. L'histogramme montre la distribution de ces coefficients inclus dans le modèle ($\hat{\beta}_4^* \neq 0$).*

pouvons noter aucun ensemble $\{0\}$ et ce, pour chacune des méthodes. En raison de la forme des distributions, le rééchantillonnage des paires entraîne davantage de bornes nulles. On peut compter 173 bornes inférieures nulles et 86 bornes

supérieures nulles par rééchantillonnage des paires contre seulement 71 bornes inférieures nulles et 26 bornes supérieures nulles par rééchantillonnage des résidus lorsque l'écart type est de 10. Nous avons également observé que pour chaque intervalle qui ne couvrait pas la vraie valeur de β_4, la borne supérieure dans le cas de l'intervalle de confiance percentile était nulle. La forme de la distribution de $\hat{\beta}_4^*$ nous permet donc d'expliquer la différence dans les pourcentages de couverture pour un rapport signal-bruit faible.

3.3.5. Conclusion

En résumé, pour une méthode de sélection S_e convergente, il est préférable d'utiliser le rééchantillonnage des paires d'observations et l'intervalle de confiance percentile. Bien que pour un rapport signal-bruit élevé les méthodes classiques et du rééchantillonnage des résidus donnent des intervalles de confiance plus courts que ceux par rééchantillonnage des paires d'observations, les pourcentages de couverture sont les mêmes et ce dernier fait beaucoup mieux pour un rapport signal-bruit plus faible. La forme de la distribution des estimateurs centrée sur la vraie valeur de $\underset{\sim}{\beta}$ jumelée au faible pourcentage d'inclusion des variables pour une méthode convergente, explique cette différence. La forme de la distribution explique également la supériorité de l'intervalle de confiance percentile sur les 2 autres intervalles de confiance par rééchantillonnage des paires d'observations.

La forme de la distribution des estimateurs de $\underset{\sim}{\beta}$ pour une répétition est également la cause de la différence entre les 2 méthodes de rééchantillonnage pour une méthode de sélection non convergente. En effet, pour une méthode de sélection S_e non convergente, il est préférable d'utiliser le rééchantillonnage des résidus au rééchantillonnage des paires d'observations ou à la méthode classique. Cette

dernière donne des intervalles de confiance plus courts que les 2 méthodes de ré-échantillonnage, mais les pourcentages de couverture pour un rapport signal-bruit faible sont très faibles. Le pourcentage d'inclusion des variables permet d'obtenir deux quantiles différents de 0 par rééchantillonnage des résidus, mais un seul par rééchantillonnage des paires d'observations. L'intervalle de confiance bootstrap choisi a moins d'impact sur les pourcentages de couverture par rééchantillonnage des résidus. Nous privilégierons cependant l'intervalle de confiance bootstrap-t-MSE comme nous l'avons vu à la section 3.2.

3.4. SOUS-ÉCHANTILLONNAGE

Dans les sections précédentes, nous avons effectué de l'inférence à partir de deux méthodes de rééchantillonnage. Pour utiliser ces méthodes, nous avons besoin d'échantillons bootstrap de taille n. Si nous avons un très grand nombre d'observations, ces techniques peuvent demander un temps de calcul important pour être appliquées. Nous verrons maintenant les résultats de nos simulations qui utilisent le sous-échantillonnage, une technique qui utilise des échantillons de taille plus modeste que le nombre d'observations de notre jeu de données.

Afin de pouvoir comparer les résultats de cette méthode à ceux des deux autres méthodes de rééchantillonnage, nous avons débuté par simuler la sélection de modèle en utilisant la matrice de design X_{50} et des écarts type de 0,1 et 10 comme pour les 2 autres méthodes. Nous avons utilisé un sous-échantillon de taille $b = 25$. Les pourcentages de couverture des intervalles de confiance obtenus suite à cette sélection étaient inférieurs à la valeur prescrite de 95% pour un rapport signal-bruit élevé contrairement à l'inférence classique et aux deux autres méthodes de rééchantillonnage. Pour un rapport signal-bruit faible, les

pourcentages de couverture, bien que supérieurs à ceux de l'inférence classique, étaient inférieurs à ceux obtenus par rééchantillonnage.

Cette contre-performance du sous-échantillonnage peut être expliquée par le fait que le ratio de la taille du sous-échantillon sur le nombre d'observations (1/2) est élevé, contrairement à ce que prescrit la théorie de la section 2.1, c'est-à-dire un ratio qui tend vers 0 lorsque n tend vers l'infini. De plus, il est essentiel que la taille du sous-échantillon soit supérieure au nombre de variables utilisées. Puisque nous travaillions avec une matrice comportant 50 observations et 8 variables, nous avions peu de latitude pour choisir la taille du sous-échantillon.

Si nous possédons une matrice de design comportant un plus grand nombre d'observations, disons 1000, il serait plus intéressant d'appliquer le sous-échantillonnage. Nous pourrons remédier au problème du ratio élevé et le temps de calcul sera réduit par rapport aux deux méthodes de rééchantillonnage. Nous utiliserons la matrice X_{1000}, comportant 1000 observations et 8 variables, construite de la même façon que la matrice X_{50}, c'est-à-dire une première colonne de 1 et les 7 autres colonnes constituées de réalisations de normale de moyenne 0 et de variance 1. Nous utiliserons un écart type de 0,447, 4,47 et 44,7 pour générer le vecteur d'erreur $\underset{\sim}{\epsilon}$. Ces valeurs illustrent un rapport signal-bruit élevé, moyen et faible. Puisque $(X'X)_{ii} \approx n$, nous avons, pour un même vecteur $\underset{\sim}{\beta}$, la relation suivante entre les écarts type de nos deux vecteurs $\underset{\sim}{\hat{\beta}}$

$$\sigma_{1000} \approx \sqrt{20} \times \sigma_{50}$$

où σ_n est approximativement l'écart type correspondant à un élément de $\underset{\sim}{\hat{\beta}}$ lorsque nous utilisons n observations. C'est pour cette raison que nous utilisons ces valeurs.

Nous commencerons par regarder les pourcentages de couverture des différents intervalles de confiance pour la matrice de design X_{1000}. Nous nous contenterons de sélectionner les modèles selon une méthode de sélection S_e convergente, BIC, et non convergente, CPM. Nous porterons une attention particulière sur l'influence de la taille du sous-échantillon. Pour ce faire, nous avons utilisé 7 tailles de sous-échantillon ($b = 30, 40, 50, 75, 100, 150, 200$). Par la suite, nous comparerons ces résultats aux méthodes de rééchantillonnage des deux sections précédentes.

3.4.1. Rapport signal-bruit élevé.

Examinons tout d'abord la qualité des modèles sélectionnés. Pour la méthode de sélection S_e CPM, la proportion de vrais modèles sélectionnés à partir des données originales est la même que celle des vrais modèles sélectionnés par sous-échantillonnage et ce peu importe la taille du sous-échantillon. Cette proportion est égale à 60%. Les 40% restants sont des modèles trop grands par chacune des méthodes. Pour la méthode de sélection S_e BIC, le modèle sélectionné à partir des données originales est le vrai modèle légèrement plus souvent que celui sélectionné par sous-échantillonnage. A partir des données originales, nous sélectionnons le vrai modèle 98% des fois. Le sous-échantillonnage sélectionne le vrai modèle dans des proportions variant de 84% à 96%. Cette proportion augmente en même temps que la taille du sous-échantillon.

Attardons-nous maintenant sur les pourcentages de couverture des intervalles de confiance obtenus par sous-échantillonnage. La figure 3.4.1 illustre les pourcentages de couverture des différents intervalles de confiance pour les coefficients β_3, β_4 et β_5 pour la méthode de sélection S_e BIC. Notons que les lignes "classique" représentent la moyenne des pourcentages de couverture des 3500 intervalles de confiance obtenus, c'est-à-dire 7 fois les 500 intervalles de confiance obtenus pour

une simulation. En effet, malgré que la même matrice X_{1000} soit utilisée, la génération des nombres aléatoires sera différente selon la taille du sous-échantillon utilisée. Les pourcentages de couverture des intervalles de confiance des coefficients nuls sont supérieurs à 99% autant par inférence classique que par sous-échantillonnage et ce, pour toutes les tailles de sous-échantillon. La très forte proportion de vrais modèles sélectionnés explique ces pourcentages de couverture. Aucun des coefficients non nuls n'est différent de la valeur prescrite de 95% par la méthode classique. Par sous-échantillonnage, les pourcentages de couverture de tous les intervalles de confiance décroissent avec la taille du sous-échantillon. Pour un sous-échantillon de taille inférieure à 50, les pourcentages sont supérieurs à la valeur prescrite, ils ne sont pas différents de cette valeur pour $b \in [50, 100]$ et ils sont inférieurs à la valeur prescrite pour un sous-échantillon de taille supérieure à 100.

La longueur moyenne des intervalles de confiance obtenus par sous-échantillonnage explique les différences entre les pourcentages de couverture. Pour les coefficients non nuls, les longueurs de ces intervalles sont supérieures aux longueurs des intervalles de confiance classiques pour un sous-échantillon de taille inférieure à 50, identiques pour $b \in [50, 100]$ et inférieures pour $b > 100$. La taille optimale du sous-échantillon se situe entre 50 et 100. Elle associe des pourcentages de couverture identiques par sous-échantillonnage et par inférence classique et des longueurs identiques d'intervalles de confiance pour les coefficients non nuls. Cependant, les longueurs des intervalles de confiance pour les coefficients nuls sont au moins 25 fois plus larges par sous-échantillonnage que par inférence classique. Pour cette raison, nous nous devons de prévilégier la méthode classique à la méthode du sous-échantillonnage.

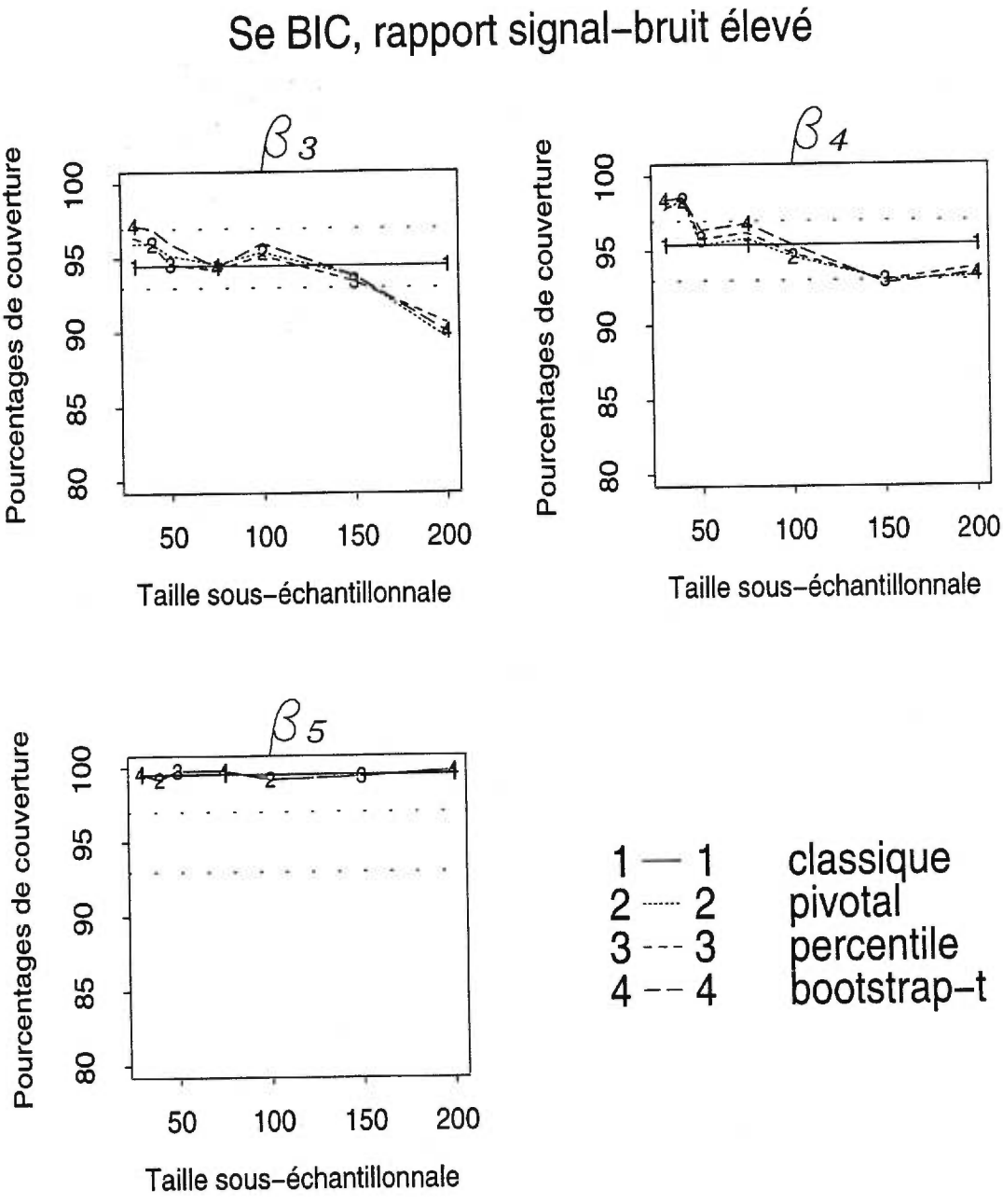

FIGURE 3.4.1. *Pourcentages de couverture des intervalles de confiance bilatéraux de β_3, β_4 et β_5 pour un rapport signal-bruit élevé et différentes tailles sous-échantillonnales. La méthode de sélection S_e BIC a été utilisée avec la méthode du sous-échantillonnage sur la matrice de design X_{1000}.*

Pour la méthode de sélection S_e CPM et un rapport signal-bruit élevé, le comportement des intervalles de confiance par sous-échantillonnage varie différemment selon que le vrai coefficient est ou n'est pas nul. La figure 3.4.2 illustre les pourcentages de couverture des 4 intervalles de confiance selon la taille du sous-échantillon pour les coefficients β_3 à β_5. Notons tout d'abord que pour la méthode classique, les moyennes des pourcentages de couverture sur les 7 simulations ne sont pas différentes de la valeur prescrite de 95%. Par sous-échantillonnage, pour un coefficient non nul, les pourcentages de couverture des intervalles de confiance décroissent légèrement à mesure que la taille du sous-échantillon augmente. Le coefficient β_3 illustré est moins représentatif à cet égard que les coefficients β_1, β_2 et β_4 (les résultats pour β_1 et β_2 ne sont pas présentés). Les pourcentages de couverture pour les intervalles de confiance de ces coefficients sont supérieurs à 95% pour $b = 30$ et inférieurs à 95% pour $b = 200$. Pour un coefficient nul, les intervalles de confiance pivotal et bootstrap-t-MSE ne sont pas différents de la valeur prescrite aussitôt que $b > 30$. Les pourcentages de couverture des intervalles de confiance percentile descendent sous la barre des 90% dès que $b \geq 75$.

Les longueurs des intervalles de confiance par sous-échantillonnage décroissent à mesure que la taille du sous-échantillon augmente. Pour $b \geq 100$, les longueurs des intervalles de confiance pivotal et percentile sont mêmes inférieures aux longueurs des intervalles de confiance classiques pour les coefficients non nuls. Cependant, pour les coefficients nuls, malgré un grand sous-échantillon ($b = 200$), les longueurs des intervalles de confiance pivotal et percentile restent au moins 4 fois supérieures aux longueurs des intervalles de confiance classiques pour les mêmes coefficients.

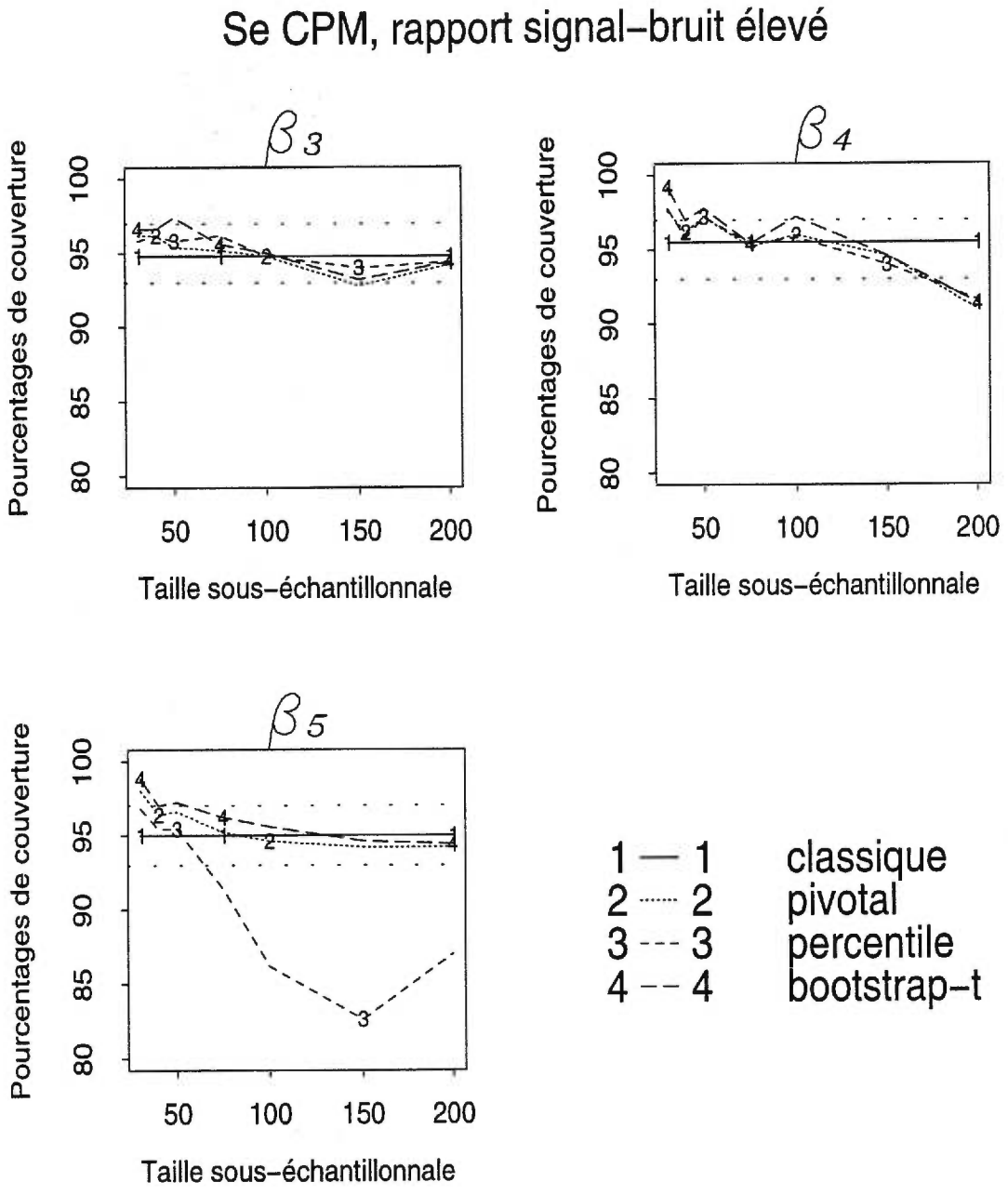

FIGURE 3.4.2. *Pourcentages de couverture des intervalles de confiance bilatéraux de β_3, β_4 et β_5 pour un rapport signal-bruit élevé et différentes tailles sous-échantillonnales. La méthode de sélection S_e CPM a été utilisée avec la méthode du sous-échantillonnage sur la matrice de design X_{1000}.*

La taille optimale du sous-échantillon dépendra de l'intervalle de confiance utilisé et de l'importance donnée à la longueur de l'intervalle. Pour l'intervalle de confiance percentile, nous n'avons d'autre choix que d'utiliser un sous-échantillon de taille inférieure à 50. Pour cette taille de sous-échantillon cependant, les intervalles de confiance classiques ont les mêmes pourcentages de couverture et sont plus courts que les intervalles de confiance percentile. Si nous utilisons les intervalles de confiance pivotal, plus courts et aussi performants que bootstrap-t-MSE, un sous-échantillon de grande taille nous donnera des intervalles de confiance pour les coefficients non nuls plus courts et aussi performants que l'inférence classique. Nous obtiendrons tout de même des intervalles de confiance bootstrap pour les coefficients nuls beaucoup plus larges pour un même pourcentage de couverture que les intervalles de confiance classiques. Le gain en temps de calcul est aussi à considérer. Puisqu'en général les intervalles de confiance classiques sont plus courts que les intervalles de confiance obtenus par sous-échantillonnage, nous devons privilégier l'inférence classique lorsque le rapport signal-brut est élevé.

3.4.2. Rapport signal-bruit moyen

Regardons maintenant les résultats de nos simulations pour un rapport signal-bruit moyen, c'est-à-dire un écart type de 4,47. Pour la méthode de sélection S_e BIC, nous sélectionnons, à partir des 500 matrices originales, le vrai modèle 75% des fois et un modèle trop petit dans une proportion de 25%. Les proportions de vrais modèles sélectionnés par sous-échantillonnage à partir des 500 000 modèles "bootstrap" lorsque la méthode S_e BIC est employée varient avec la taille du sous-échantillon de 0,72% à 16,08%. La proportion de modèles biaisés atteint 99,02% pour $b = 30$ et diminue jusqu'à 83,25% pour $b = 200$. Les différences entre les proportions des vrais modèles issus des données originales et "bootstrap" sont un

peu moins importantes lorsque la méthode de sélection S_e CPM est employée. Pour les données originales, la proportion de vrais modèles est de 60% et celle des modèles biaisés est de 40%. La qualité du modèle varie beaucoup selon la taille du sous-échantillon. Le tableau 3.4.1 contient les proportions des modèles biaisés, des vrais modèles et des modèles trop grands sélectionnés par la méthode de sélection S_e CPM lorsque la méthode du sous-échantillonnage est utilisée.

TABLEAU 3.4.1. *Qualité de la sélection des 500 000 modèles "bootstrap" par la méthode de sélection S_e CPM pour un rapport signal bruit moyen lorsque le sous-échantillonnage est employé.*

	Vrai modèle	Modèle trop grand	Modèle biaisé
$b = 30$	0,033	0,031	0,936
$b = 40$	0,058	0,047	0,895
$b = 50$	0,085	0,065	0,850
$b = 75$	0,148	0,105	0,747
$b = 100$	0,199	0,137	0,664
$b = 150$	0,271	0,189	0,540
$b = 200$	0,319	0,218	0,463

Nous concluons que pour un rapport signal-bruit moyen, la méthode classique sélectionne davantage le vrai modèle que le sous-échantillonnage et ce, peu importe que la méthode de sélection S_e soit convergente ou non.

Regardons maintenant les pourcentages de couverture pour la méthode de sélection S_e BIC. La figure 3.4.3 illustre les pourcentages de couverture des 4 intervalles de confiance des coefficients β_3, β_4 et β_5 pour ce rapport signal-bruit et cette méthode de sélection. Les lignes "classique" sont formées comme pour les

figures pour un rapport signal-bruit élevé. Les coefficients nuls ont des pourcentages de couverture au-dessus de 99%, peu importe la taille du sous-échantillon et que nous utilisions la méthode classique ou le sous-échantillonnage. Cependant, les intervalles de confiance classiques sont beaucoup plus courts. Lorsque $b = 200$, les longueurs moyennes des intervalles de confiance par sous-échantillonnage sont encore au moins 19 fois plus larges que les intervalles de confiance classiques.

Les coefficients $\beta_0, \beta_1, \beta_2$, et β_3 ont un comportement similaire entre eux. De façon générale, tous les pourcentages de couverture se retrouvent dans l'intervalle $[93\%, 97\%]$. Les longueurs des intervalles de confiance par sous-échantillonnage sont cependant plus larges que les longueurs des intervalles de confiance classiques par un facteur de 1,03 à 1,50 selon le coefficient et la taille du sous-échantillon considéré.

Le comportement du coefficient β_4 est différent de celui des autres coefficients. C'est le seul coefficient dont le pourcentage de l'intervalle de confiance classique est inférieur à la valeur prescrite. Les pourcentages de couverture des intervalles de confiance pivotal et bootstrap-t-MSE sont similaires à ceux de l'inférence classique. De plus, la longueur des intervalles de confiance classiques et par sous-échantillonnage est en moyenne la même. Seul l'intervalle de confiance percentile a un comportement différent des autres. Il faut attendre d'utiliser $b > 100$ pour obtenir des pourcentages de couverture comparables aux autres. Pour une valeur b inférieure à 100, les pourcentages de couverture sont beaucoup plus faibles.

Le comportement des intervalles de confiance par sous-échantillonnage est encore plus erratique lorsque les modèles sont sélectionnés par la méthode S_e CPM. La figure 3.4.4 illustre les différents intervalles de confiance pour les coefficients β_3 à β_5. La moyenne des intervalles de confiance classiques pour les

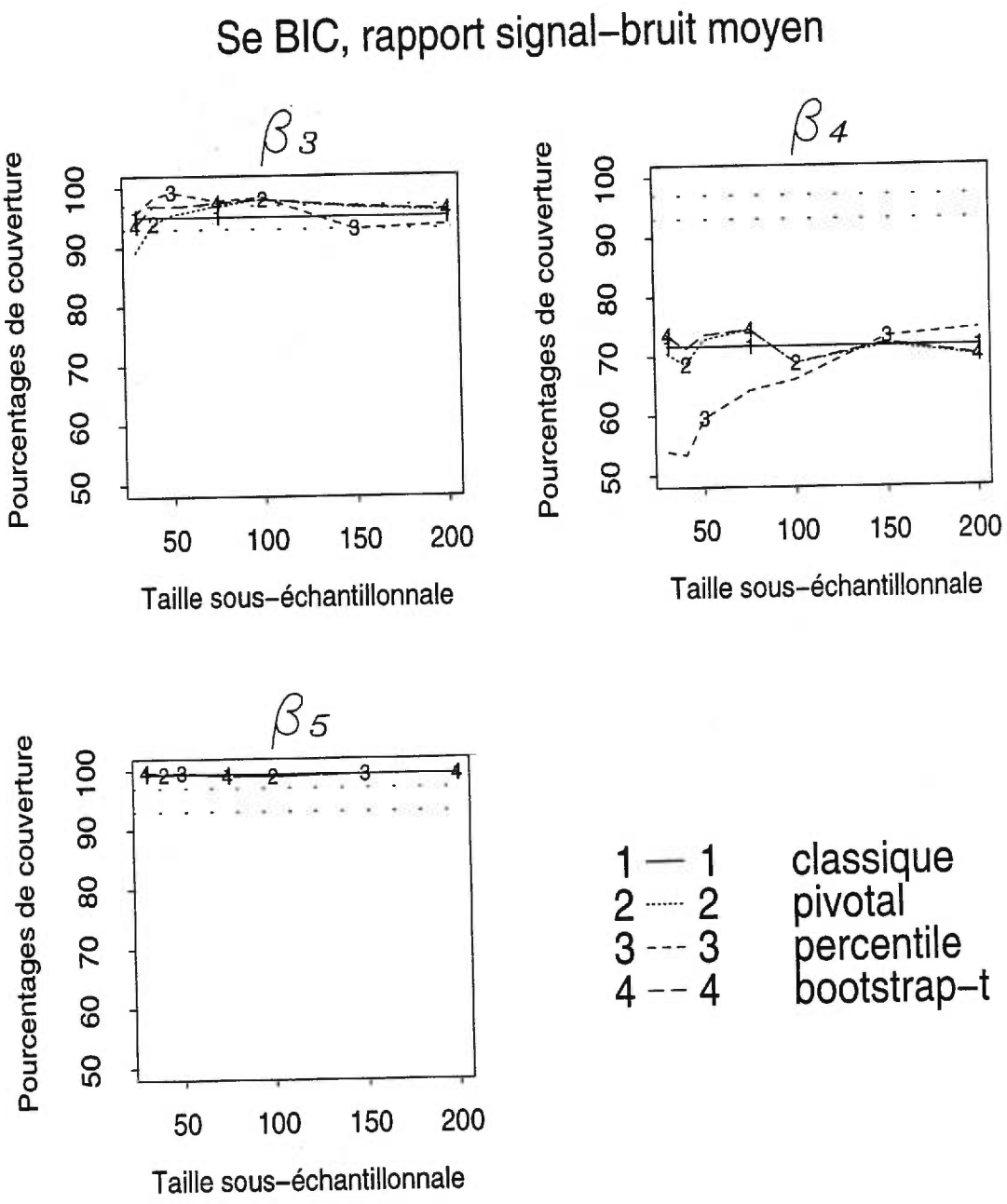

FIGURE 3.4.3. *Pourcentages de couverture des intervalles de confiance bilatéraux de β_3, β_4 et β_5 pour un rapport signal-bruit moyen et différentes tailles sous-échantillonnales. La méthode de sélection S_e BIC a été utilisée avec la méthode du sous-échantillonnage sur la matrice de design X_{1000}.*

7 simulations est illustrée par la ligne "classique". L'inférence classique donne d'excellents résultats. Tous les coefficients ont des pourcentages de couverture à l'intérieur de l'intervalle [93%, 97%]. Lorsque nous utilisons la méthode du sous-échantillonnage, le comportement des intervalles de confiance est différent selon le coefficient observé et le type d'intervalle de confiance. Pour un coefficient clairement non nul ($\beta_0, \beta_1, \beta_2, \beta_3$), les pourcentages de couverture des intervalles de confiance de type percentile et bootstrap-t-MSE par sous-échantillonnage sont supérieurs à 97% pour $b \leq 75$ et décroissent par la suite sous les 97%, mais au-dessus de la limite inférieure de 93%. L'intervalle de confiance pivotal a des pourcentages de couverture qui ne sont pas différents de la valeur prescrite pour presque toutes les tailles sous-échantillonnales. Ajoutons également que les longueurs moyennes des intervalles de confiance par sous-échantillonnage des coefficients clairement non nuls sont plus longs que les intervalles de confiance classiques par un facteur pouvant atteindre 1,3 jusqu'à $b = 100$. Pour une valeur supérieure à 100, les longueurs moyennes des intervalles de confiance par sous-échantillonnage sont plus courts que les classiques par un facteur pouvant atteindre 0,9.

Pour un coefficient presque nul, β_4, les pourcentages de couverture des intervalles de confiance pivotal et bootstrap-t-MSE ne sont pas différents de la valeur prescrite pour $b \leq 50$. Par la suite, les pourcentages de couverture décroissent. Au contraire, les pourcentages de couverture des intervalles de confiance percentile sont inférieurs à 95% pour $b < 100$, mais ils sont croissants. Dès que $b \geq 100$, les pourcentages de couverture de l'intervalle de confiance percentile ne sont pas différents de 95% (Coefficient β_4, figure 3.4.4). Les intervalles de confiance par sous-échantillonnage sont plus larges seulement pour $b = 30$. Pour $b > 30$, le

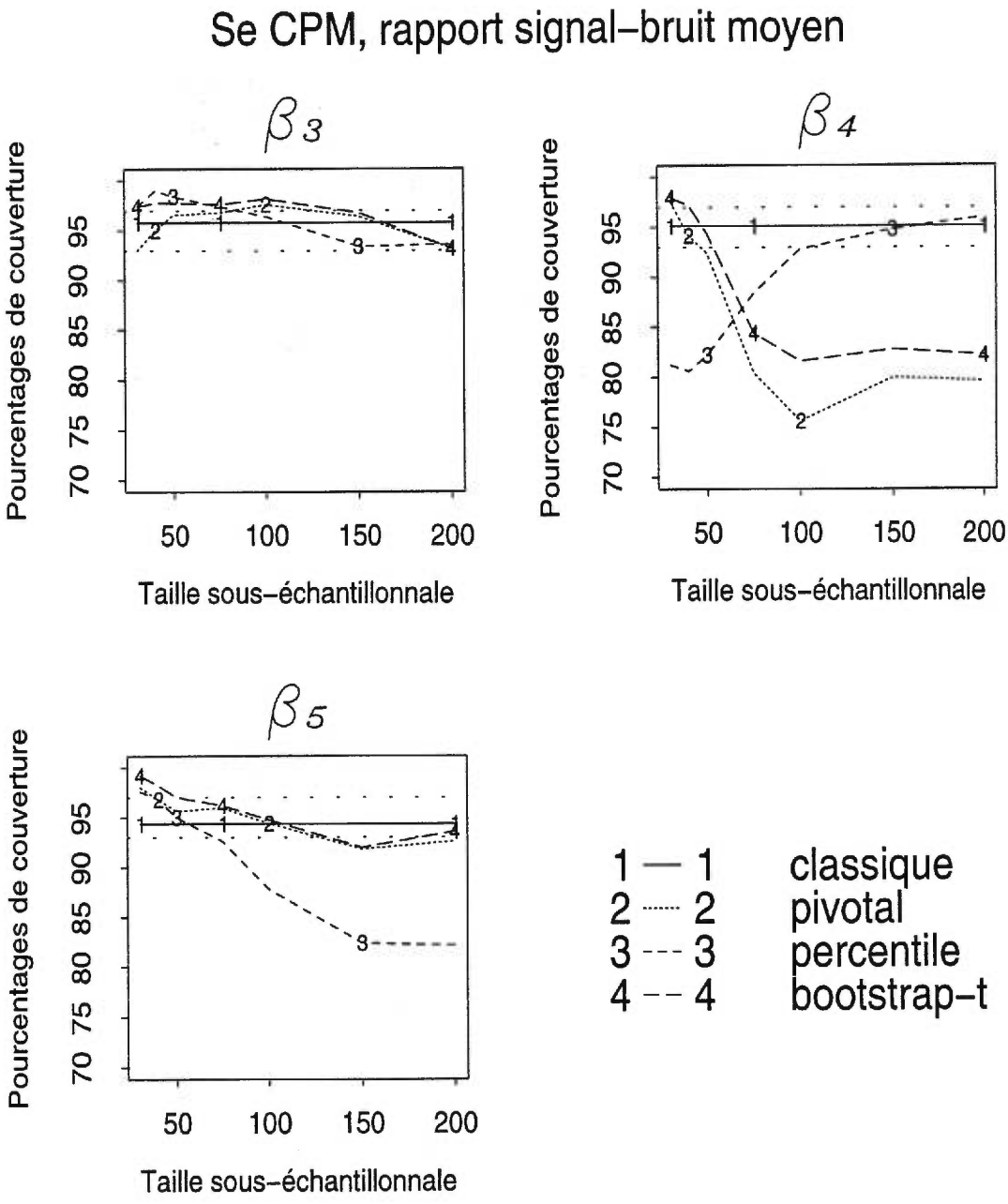

FIGURE 3.4.4. *Pourcentages de couverture des intervalles de confiance bilatéraux de β_3, β_4 et β_5 pour un rapport signal-bruit moyen et différentes tailles sous-échantillonnales. La méthode de sélection S_e CPM a été utilisée avec la méthode du sous-échantillonnage sur la matrice de design X_{1000}.*

ratio des longueurs des intervalles de confiance par sous-échantillonnage sur les classiques peut atteindre 0,8.

Pour un coefficient nul, $\beta_5, \beta_6, \beta_7$, les pourcentages de couverture des intervalles de confiance pivotal et bootstrap-t-MSE ne sont pas différents de la valeur prescrite. Cependant, les pourcentages de couverture des intervalles de confiance percentile descendent sous la valeur de 93% dès que $b \geq 75$ et décroissent par la suite. (Coefficient β_5, figure 3.4.4). De plus, même pour $b = 200$, les intervalles de confiance par sous-échantillonnage des coefficients nuls sont au moins 5 fois larges que leurs pendants classiques.

De façon générale, l'inférence classique devrait être privilégiée par rapport au sous-échantillonnage pour un rapport signal-bruit moyen. Les pourcentages de couverture des intervalles de confiance classiques ne sont pas inférieurs à ceux des intervalles de confiance par sous-échantillonnage et leurs longueurs sont généralement plus courtes. Parmi les intervalles de confiance par sous-échantillonnage, l'intervalle de confiance percentile est beaucoup moins stable que les 2 autres. En raison de cette instabilité, il est difficile de déterminer une taille de sous-échantillon optimale. Pour la méthode de sélection S_e BIC, si l'on désire employer l'intervalle de confiance percentile en plus des 2 autres, la taille optimale se situe entre 50 et 100. Elle nous permet d'obtenir des pourcentages de couverture pas différents des intervalles de confiance classiques, quoique, en général, plus longs. Pour la méthode de sélection S_e CPM, on doit privilégier une petite taille du sous-échantillon ($b = 50$) et ne pas utiliser l'intervalle de confiance percentile. Ce dernier ne permet pas d'obtenir des pourcentages de couverture qui ne sont pas différents de la valeur prescrite pour le coefficient β_4 et un coefficient nul pour une même valeur de b.

3.4.3. Rapport signal-bruit faible

Considérons les résultats des simulations portant sur la sélection de modèle en utilisant un rapport signal-bruit faible, c'est-à-dire un écart type de 44,7 ($\sqrt{20} \times 10$) et la matrice X_{1000}. Notons tout d'abord que presque la totalité des modèles sont biaisés et ce, pour toutes les méthodes de sélection S_e et que les données soient issues de la matrice originale ou de celle obtenue par sous-échantillonnage.

Puisque les modèles sont tous biaisés, pour la méthode de sélection S_e BIC, qui sélectionne pour une matrice de taille fixe des modèles plus courts, les variables dont les vrais coefficients sont nuls sont très rarement incluses dans le modèle. Ceci explique les pourcentages de couverture de plus de 99% des coefficients nuls autant pour les intervalles de confiance classiques que par sous-échantillonnage. Ces pourcentages de couverture illustrés par le coefficient β_5 ainsi que les pourcentages de couverture des coefficients β_4 et β_5 sont tous présentés dans la figure 3.4.5 pour la méthode de sélection S_e BIC et différentes tailles sous-échantillonnales. Malgré le fait que tous les intervalles de confiance ont des pourcentages de couverture identiques, les intervalles de confiance classiques sont plus de 30 fois plus courts que les intervalles de confiance par sous-échantillonnage pour $b = 200$, la taille sous-échantillonnale correspondant aux intervalles de confiance les plus courts.

Le comportement des coefficients β_3 et β_4 est représentatif de tous les coefficients non nuls. Les intervalles de confiance classiques ont des pourcentages de couverture très en deçà de la valeur prescrite. On remarque que pour les intervalles de confiance par sous-échantillonnage, les pourcentages de couverture des intervalles de confiance décroissent lorsque b augmente. Les intervalles de confiance pivotal et bootstrap-t-MSE ne font guère mieux que les intervalles de confiance classiques dès que $b = 50$. L'intervalle de confiance percentile performe

mieux, particulièrement pour une petite valeur de b. Evidemment, les intervalles de confiance par sous-échantillonnage sont 10 fois plus larges que leurs pendants classiques, mais ces derniers sont beaucoup trop optimistes. Les pourcentages de couverture des intervalles de confiance des coefficients nuls sont tous près de 100% en raison de la très faible sélection de ces variables. Puisque les pourcentages de couverture sont décroissants, la taille sous-échantillonnale optimale est $b = 30$.

Les pourcentages de couverture diffèrent davantage selon les coefficients observés lorsque la méthode de sélection S_e CPM est utilisée sur la matrice X_{1000} avec un rapport signal-bruit faible. Les résultats sont résumés dans la figure 3.4.6 qui illustre les différents intervalles de confiance des coefficients β_3 à β_5. Pour un coefficient clairement non nul, (β_0 à β_3), les intervalles de confiance classiques sont très mauvais, comparables à ceux obtenus avec la matrice X_{50} pour un rapport signal-bruit faible. Les intervalles de confiance par sous-échantillonnage font beaucoup mieux, particulièrement pour une taille sous-échantillonnale petite où les pourcentages de couverture ne diffèrent pas de la valeur prescrite. Les pourcentages de couverture des intervalles de confiance pivotal et bootstrap-t-MSE sont décroissants pour $b > 50$. Les intervalles de confiance percentile décroissent plus rapidement, mais remontent légèrement pour $b > 100$. (Coefficient β_3, figure 3.4.6).

Pour le coefficient β_4, le coefficient dont la vraie valeur est presque nulle, les intervalles de confiance classiques ont des pourcentages de couverture ne dépassant pas 20%. Les pourcentages de couverture des intervalles de confiance par sous-échantillonnage font beaucoup mieux. Ils sont tous les 3 décroissants, mais ne sont pas différents de 95% jusqu'à $b = 50$. (Coefficient β_4, figure 3.4.6).

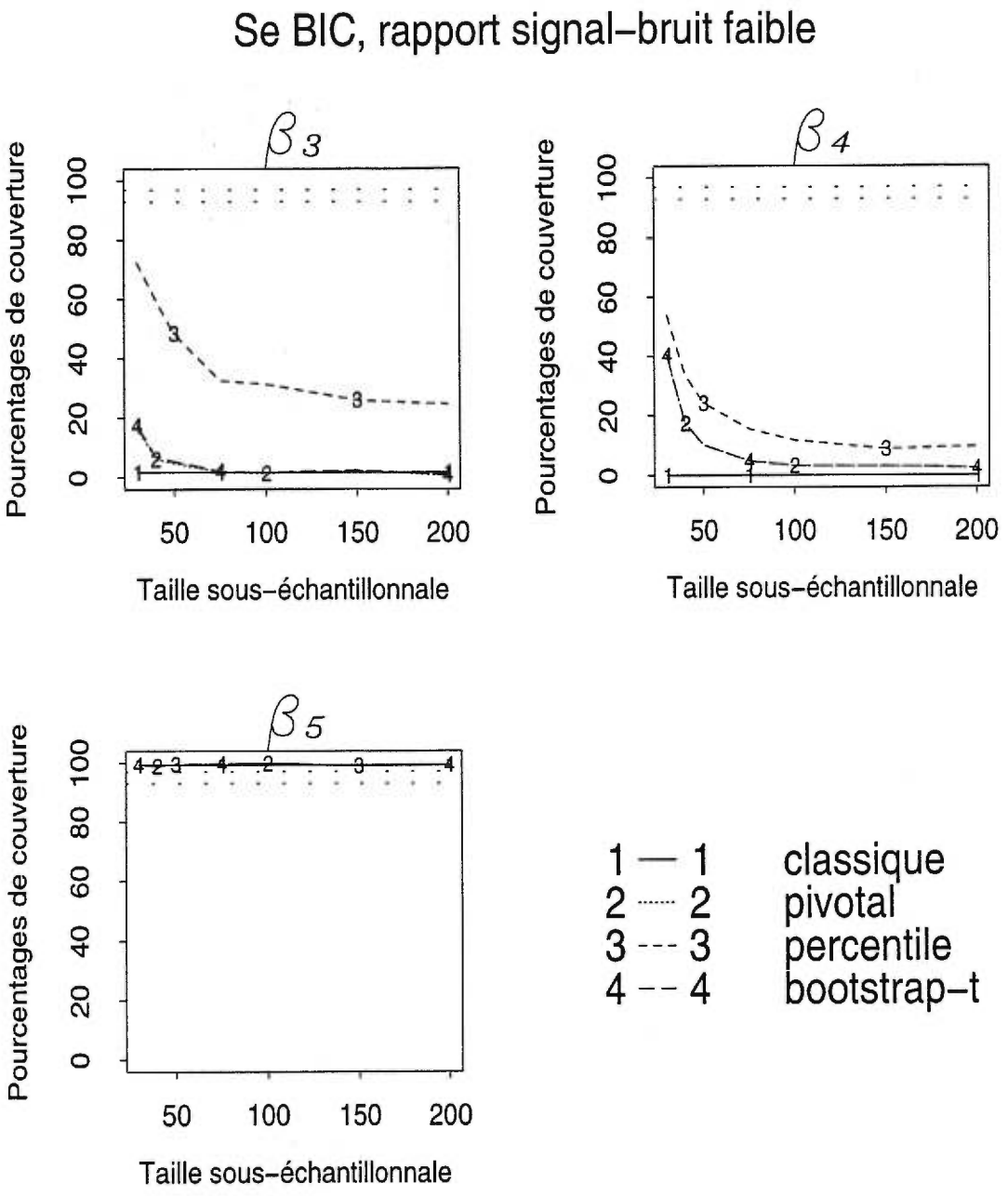

FIGURE 3.4.5. *Pourcentages de couverture des intervalles de confiance bilatéraux de β_3, β_4 et β_5 pour un rapport signal-bruit faible et différentes tailles sous-échantillonnales. La méthode de sélection S_e BIC a été utilisée avec la méthode du sous-échantillonnage sur la matrice de design X_{1000}.*

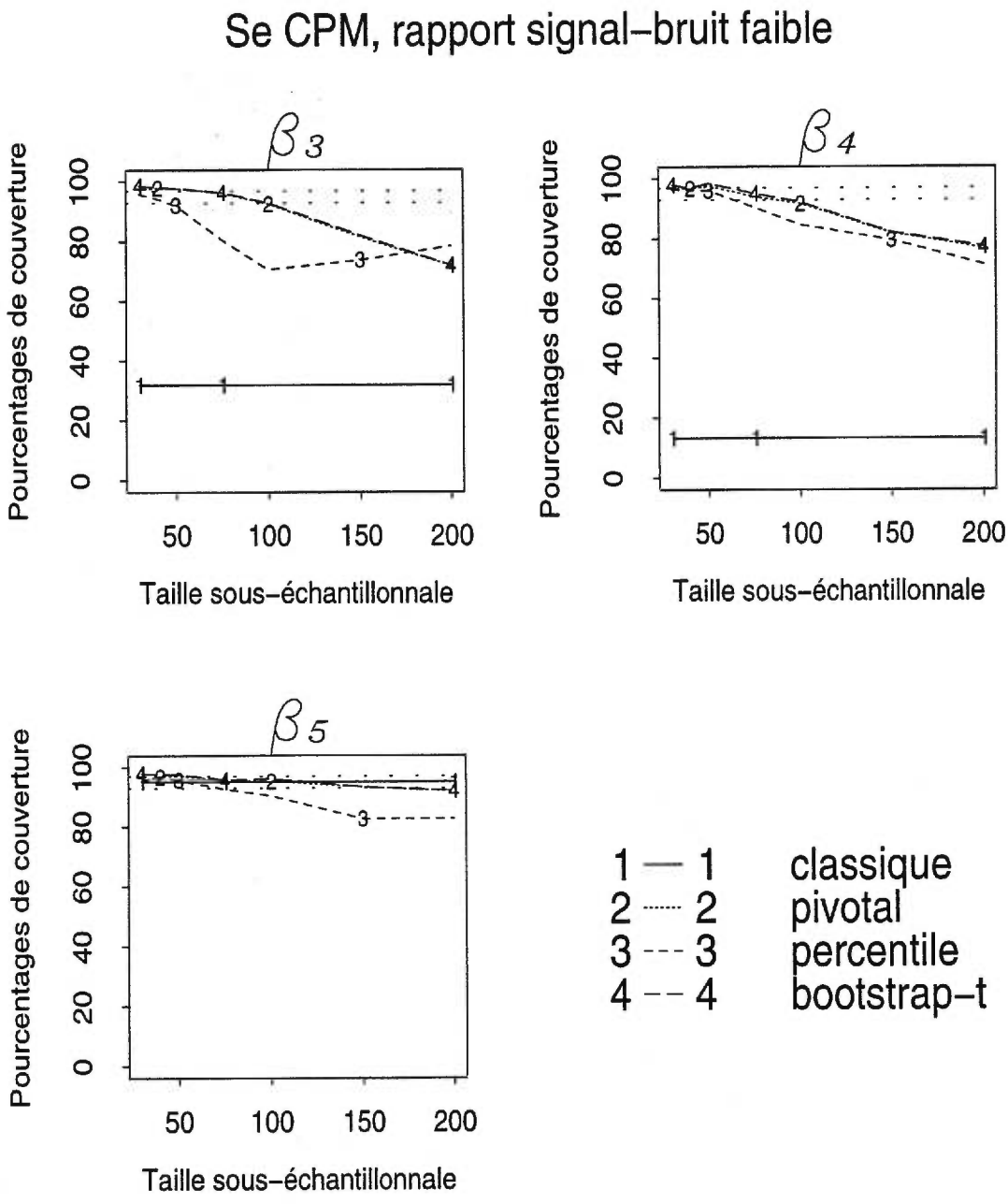

FIGURE 3.4.6. *Pourcentages de couverture des intervalles de confiance bilatéraux de β_3, β_4 et β_5 pour un rapport signal-bruit faible et différentes tailles sous-échantillonnales. La méthode de sélection S_e CPM a été utilisée avec la méthode du sous-échantillonnage sur la matrice de design X_{1000}.*

Finalement, les intervalles de confiance classiques des coefficients nuls ont des pourcentages de couverture qui ne sont pas différents de la valeur prescrite de 95%. Il en va de même pour les intervalles de confiance pivotal et bootstrap-t-MSE pour toutes les valeurs de b. Les intervalles de confiance percentile décroissent rapidement et sont différents de la valeur prescrite pour $b \geq 75$. (Coefficient β_5, figure 3.4.6).

Les intervalles de confiance par sous-échantillonnage sont plus larges, par un facteur de 6 environ, mais couvrent mieux la vraie valeur que les intervalles de confiance classiques. Vu la piètre performance des intervalles de confiance percentile pour une grande valeur de b, nous privilégions l'utilisation d'une petite taille de sous-échantillon, inférieure à 50. Pour une petite taille, tous les intervalles de confiance par sous-échantillonnage sont comparables lorsque la méthode de sélection S_e CPM est employée. Pour plus de stabilité, nous conseillons tout de même d'utiliser l'intervalle de confiance pivotal ou bootstrap-t-MSE.

3.4.4. Influence du rapport signal-bruit

Nous avons, dans les sous-sections précédentes, étudié l'influence de la taille sous-échantillonnale sur les pourcentages de couverture des intervalles de confiance par sous-échantillonnage pour différentes valeurs du rapport signal-bruit. Nous verrons ici l'influence de ce rapport sur l'intervalle de confiance pivotal qui s'avère, avec l'intervalle de confiance bootstrap-t-MSE, l'intervalle de confiance le plus stable.

Puisque nous avons observé des comportements différents selon le coefficient, nous allons, une fois de plus, considérer séparément les pourcentages de couverture selon que le vrai coefficient est clairement non nul, presque nul ou nul. La figure 3.4.7 illustre les pourcentages de couverture des coefficients β_3, β_4 et β_5,

représentatifs des coefficients clairement non nuls, presque nul et nuls pour la méthode de sélection S_e BIC et différentes valeurs du rapport signal-bruit. Pour un écart type inférieur à 5, l'intervalle de confiance pivotal couvre très bien la vraie valeur du coefficient β_3. Au contraire, pour un rapport signal-bruit faible, les pourcentages de couverture n'atteignent pas 20%. On remarque également que la taille du sous-échantillon influence peu les pourcentages de couverture.

Le coefficient β_4 varie davantage selon le rapport signal-bruit utilisé. Les pourcentages de couverture croissent régulièrement à mesure que le rapport signal-bruit augmente. La taille du sous-échantillon influence encore peu les pourcentages de couverture de l'intervalle de confiance pivotal sinon pour un rapport signal-bruit plus faible. Un rapport signal-bruit élevé est cependant requis pour obtenir des pourcentages de couverture qui ne sont pas différents de la valeur prescrite.

Pour la méthode de sélection S_e BIC, les pourcentages de couverture des intervalles de confiance pour les coefficients nuls sont tous plus élevés que 99% en raison de la forte proportion de modèles biaisés et de vrais modèles sélectionnés selon que le rapport signal-bruit est faible ou élevé.

La qualité du modèle sélectionné par sous-échantillonnage explique les différences des pourcentages de couverture. Les proportions de vrais modèles sélectionnés à partir des données originales passent de 98% à 75% à près de 0% pour des rapports signal-bruit élevé, moyen et faible. Par sous-échantillonnage, la proportion de vrais modèles passe, en moyenne, de 90% à 10% à près de 0%. Notons que puisque la sélection de modèle est faite à partir de b observations plutôt que n, avec b beaucoup plus petit que n, par sous-échantillonnage le rapport signal-bruit effectif est plus faible que pour toute autre méthode. Il devient donc encore

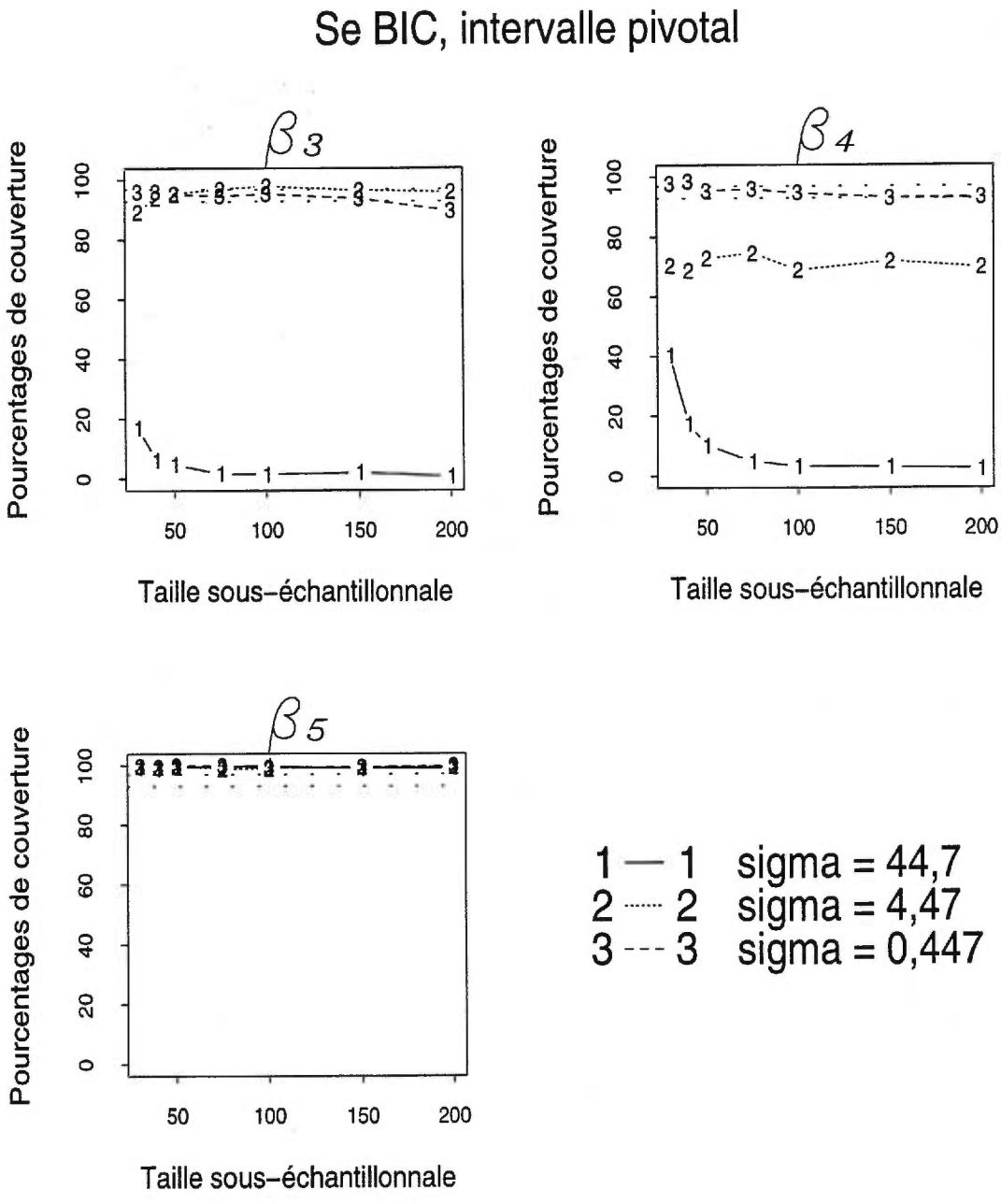

FIGURE 3.4.7. *Pourcentages de couverture des intervalles de confiance bilatéraux pivotal de β_3, β_4 et β_5 pour différents rapports signal-bruit et différentes tailles sous-échantillonnales. La méthode de sélection S_e BIC a été utilisée avec la méthode du sous-échantillonnage sur la matrice de design X_{1000}.*

plus difficile de sélectionner le vrai modèle par cette méthode. Les pourcentages de couverture pour les intervalles classiques sont supérieurs ou égaux aux pourcentages de couverture des intervalles de confiance par sous-échantillonnage pour des rapports signal-bruit élevé ou moyen. Pour un rapport signal-bruit faible, les pourcentages de couverture sont très faibles dans les 2 cas. La qualité de la sélection semble donc importante pour le sous-échantillonnage avec la méthode de sélection S_e BIC.

Lorsque nous sélectionnons le modèle par la méthode S_e CPM, les pourcentages de couverture des intervalles de type pivotal ne sont pas différents de la valeur prescrite pour une taille sous-échantillonnale $b \leq 50$. La figure 3.4.8 illustre ce phénomène en présentant les pourcentages de couverture des intervalles de type pivotal des coefficients β_3, β_4 et β_5 pour différentes valeurs du rapport signal-bruit lorsque la méthode de sélection S_e CPM est employée. Pour une taille sous-échantillonnale plus élevée ($b > 50$), les pourcentages de couverture des coefficients non nuls diminuent lorsque le rapport signal-bruit diminue comme pour les intervalles de confiance de type pivotal des coefficients non nuls lorsque la méthode S_e BIC est employée. L'importance de la taille du sous-échantillon est cependant plus marquée lors de l'emploi de la méthode de sélection S_e CPM que lors de l'emploi de la méthode S_e BIC. Toutefois dans ce dernier cas, lorsque le rapport signal-bruit est faible, les pourcentages de couverture sont uniformément mauvais.

La qualité du modèle sélectionné par sous-échantillonnage est encore la cause de la diminution des pourcentages de couverture pour des tailles sous-échantillonnales élevées lorsque la méthode S_e CPM est employée. En effet, les proportions de vrais modèles sélectionnés à partir des données originales passent de 60% pour

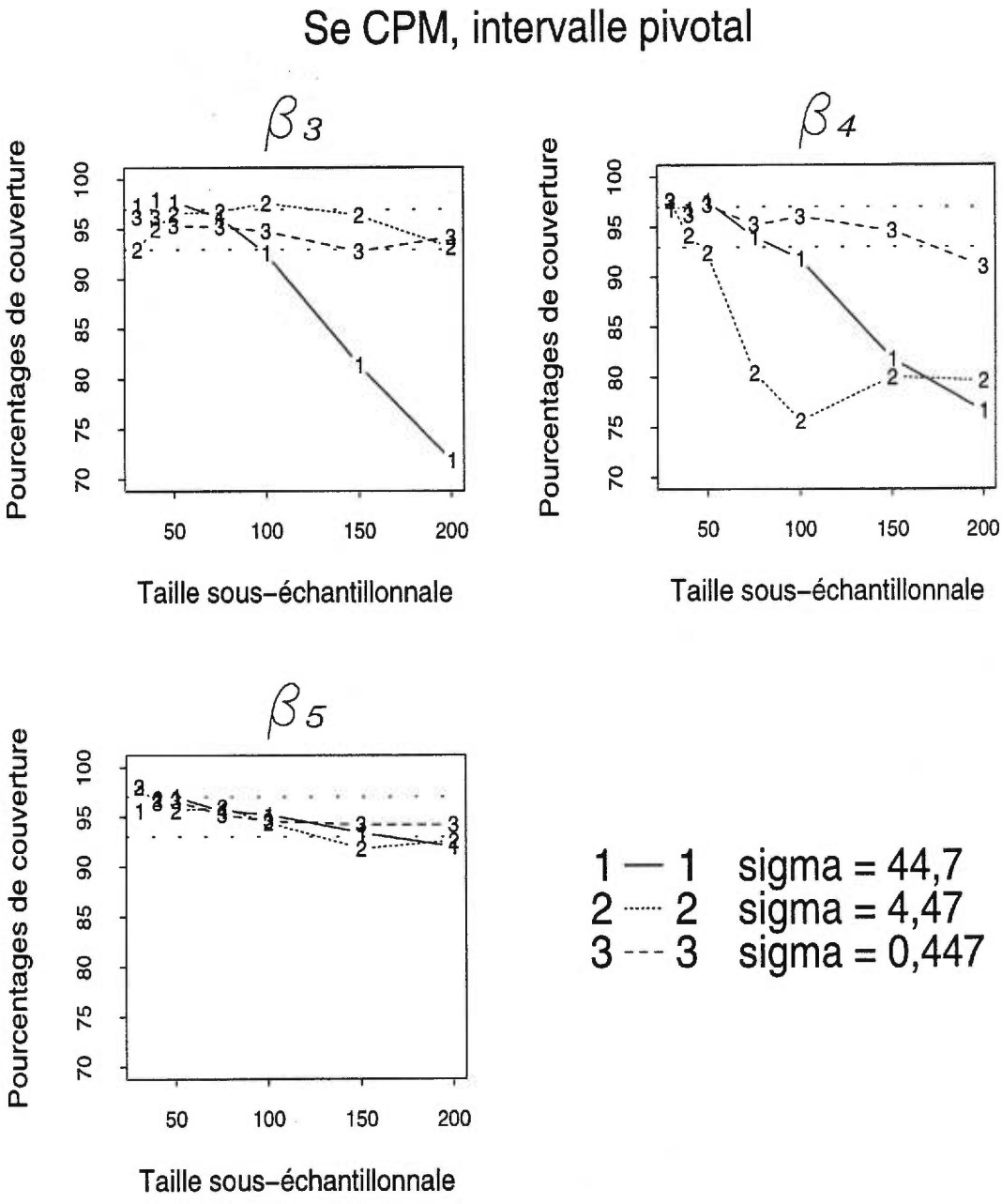

FIGURE 3.4.8. *Pourcentages de couverture des intervalles de confiance bilatéraux pivotal de β_3, β_4 et β_5 pour différents rapports signal-bruit et différentes tailles sous-échantillonnales. La méthode de sélection S_e CPM a été utilisée avec la méthode du sous-échantillonnage sur la matrice de design X_{1000}.*

des rapports signal-bruit élevé et moyen à près de 0% pour un rapport signal-bruit faible. Pour les données issues du sous-échantillonnage, ces proportions passent de 60% à 32% ($b = 200$) à près de 0%. Pour une taille de sous-échantillon inférieure ou égale à 50, les proportions de vrais modèles, plus faibles, n'influencent pas les pourcentages de couverture. A noter cependant, que pour des petites tailles sous-échantillonnales, la proportion de modèles biaisés est comparable à celle par inférence classique et beaucoup plus faible que pour des grandes tailles.

3.4.5. Comparaison des méthodes de rééchantillonnage et de sous-échantillonnage.

Nous mettrons fin à cette section et à ce chapitre en confrontant les trois méthodes de rééchantillonnage. Nous débuterons par comparer l'efficacité de chacune des méthodes en mettant en parallèle les temps de calcul requis pour faire de l'inférence avec ces méthodes. Nous résumerons rapidement les principaux résultats obtenus avec la méthode de rééchantillonnage des paires d'observations et des résidus lorsque la matrice de design employée est X_{1000}, c'est-à-dire une matrice comportant $n = 1000$ observations. Nous terminerons par la comparaison des pourcentages de couverture des meilleures combinaisons possibles des différentes méthodes de sélection et de types d'intervalle de confiance pour chacune des méthodes de rééchantillonnage.

Efficacité

Afin de pouvoir comparer la méthode du sous-échantillonnage aux deux autres méthodes de rééchantillonnage, nous devons utiliser la même matrice de design X_{1000}. Dans tous nos programmes, la sélection de modèle et l'estimation des coefficients sont effectuées à partir de la décomposition QR de la matrice X_{1000}. Cette décomposition demande le calcul $\mathcal{O}(n^2)$ opérations. Bien que l'utilisation

de Fortran 77 entraîne un temps de calcul linéaire en fonction du nombre de répétitions, ce temps de calcul est quadratique en fonction de la taille de la matrice de design.

Le sous-échantillonnage utilise des matrices de dimension beaucoup plus restreinte que le rééchantillonnage des paires d'observations ou des résidus. Le tableau 3.4.2 nous donne les temps de calcul d'une répétition des différents programmes de rééchantillonnage lorsque la matrice de design X_{1000} est employée. Nous avons établi que le temps de calcul pour effectuer une répétition du programme de rééchantillonnage des paires d'observations lorsque la matrice de design X_{50} est employée est égal à une unité de temps ($t = 1$). Les méthodes de sélection S_e CPM et S_b CPM ont été utilisées avec un rapport signal-bruit faible.

Par exemple, si le temps pour calculer une répétition par sous-échantillonnage en utilisant la matrice de design X_{1000} et $b = 30$ est de une minute, le même temps que pour une répétition par rééchantillonnage des paires en utilisant la matrice de design X_{50}, il faudra 210 minutes (3 heures et 30 minutes) pour obtenir la même chose avec le programme du rééchantillonnage des paires d'observations.

Nous pouvons noter que le temps de calcul pour le programme de rééchantillonnage des résidus est moins long que le programme de rééchantillonnage des paires d'observations. Dans ce dernier programme, nous effectuons, avant chaque régression un test pour savoir si la matrice de design X^* est singulière ou non. Ce test requiert une décomposition QR supplémentaire qui augmente le temps de calcul de ce programme.

Rééchantillonnage des paires d'observations avec la matrice de design X_{1000}.

Lorsque nous utilisons le rééchantillonnage des paires d'observations avec la matrice de design X_{1000}, les pourcentages de couverture peuvent être inférieurs

TABLEAU 3.4.2. *Efficacité des différents programmes de rééchantillonnage lorsque la matrice de design X_{1000} est utilisée. Une unité de temps correspond au temps de calcul du programme de rééchantillonnage des paires pour X_{50}.*

Programme rééchantillonnage des paires, X_{50}, t = 1 unité de temps	
Méthode	t
Rééchantillonnage des paires	210,83
Rééchantillonnage des résidus	98,22
Sous-échantillonnage	
$b = 30$	1,00
$b = 40$	1,11
$b = 50$	1,50
$b = 75$	2,28
$b = 100$	3,00
$b = 150$	6,33
$b = 200$	9,11

aux pourcentages de couverture pour la matrice de design X_{50}. Les conclusions demeurent cependant les mêmes qu'à la section 3.3.

Nous nous rappelons que pour un rapport signal-bruit élevé et la méthode de sélection S_e BIC, le rééchantillonnage des paires d'observations nous donnait des pourcentages de couverture qui n'étaient pas différents de 95% pour les coefficients non nuls. Pour les coefficients nuls, seul l'intervalle de confiance percentile nous donnait des pourcentages de couverture près de 100% tandis que les autres intervalles n'étaient pas différents de 95%. Lorsque nous utilisons la matrice de design X_{1000}, les pourcentages de couverture demeurent les mêmes, sauf que tous les intervalles de confiance des coefficients nuls sont près de 100%. Nous expliquons

cela par la meilleure qualité de sélection, due à la convergence de la méthode de sélection, lorsque X_{1000} est utilisée. En effet, la proportion de vrais modèles passe de 90,00% à 98,80% lorsque nous utilisons les données originales et de 63,29% à 86,9% lorsque nous utilisons les données bootstrap.

Lorsque nous utilisons la méthode de sélection S_e CPM, tous les pourcentages de couverture pour la matrice de design X_{1000} ne sont pas différents des pourcentages de couverture pour X_{50}.

Nous avions conclu et nous concluons encore que l'intervalle de confiance percentile est très légèrement supérieur aux deux autres intervalles de confiance bootstrap pour un rapport signal-bruit élevé.

Pour un rapport signal-bruit faible et la méthode de sélection S_e BIC, les pourcentages de couverture pour X_{1000} sont inférieurs aux pourcentages de couverture pour X_{50}. Ils sont cependant inférieurs pour chacun des intervalles. L'intervalle de confiance percentile demeure donc le meilleur.

Lorsque nous utilisons la méthode de sélection S_e CPM, les pourcentages de couverture pour X_{1000} sont légèrement inférieurs à ceux de X_{50} lorsque nous utilisons les intervalles de confiance pivotal et bootstrap-t-MSE et identiques pour les intervalles de confiance percentile. Ce dernier demeure encore le meilleur.

Pour les 2 méthodes de sélection S_e, presque tous les modèles demeurent biaisés lorsque $n = 1000$.

Pour comparer le rééchantillonnage des paires d'observations et le sous-échantillonnage, nous utiliserons seulement l'intervalle de confiance percentile.

Rééchantillonnage des résidus avec la matrice de design X_{1000}.

Nous obtenons également une certaine diminution des pourcentages de couverture des intervalles de confiance obtenus par rééchantillonnage des résidus.

Pour la méthode de sélection S_e BIC, nous obtenons une légère augmentation des pourcentages de couverture pour un rapport signal-bruit élevé. En effet, pour les coefficients non nuls, tous les pourcentages de couverture des intervalles de confiance pivotal ne sont pas différents de la valeur prescrite. Les pourcentages de couverture des intervalles de confiance pivotal pour la matrice X_{1000} sont en fait les mêmes que les pourcentages de couverture des intervalles de confiance bootstrap-t-MSE pour la matrice X_{50}. Pour les coefficients nuls, tous les pourcentages de couverture sont près de 100% en raison, encore une fois de la meilleure qualité de la sélection. La proportion de modèles correctement sélectionnés passe en effet de 72,7% pour X_{50} à 90,4% pour X_{1000} lorsque les données bootstrap et la méthode S_b CPM sont employées.

Pour un rapport signal-bruit faible, les pourcentages de couverture diminuent sensiblement de la même façon que par rééchantillonnage des paires d'observations. Les conclusions demeurent cependant les mêmes qu'à la section 3.2.

Pour la méthode de sélection S_e CPM, les résultats sont pratiquement identiques pour les matrices de design X_{50} et X_{1000}. Pour un rapport signal-bruit élevé, les pourcentages de couverture et la qualité de la sélection demeurent les mêmes. On peut simplement noter une légère augmentation des pourcentages de couverture pour un rapport signal-bruit faible. De la même façon que pour une méthode convergente et un rapport signal-bruit élevé, les intervalles de confiance pivotal conviennent mieux que les intervalles de confiance bootstrap-t-MSE.

Afin de comparer le rééchantillonnage des résidus au sous-échantillonnage, nous utiliserons uniquement les intervalles de confiance pivotal obtenus avec la méthode de sélection S_b CPM.

Comparaison

Nous avons choisi de ne comparer que le meilleur intervalle de confiance pour chacune des méthodes de rééchantillonnage. Par rééchantillonnage des paires d'observations, nous avons utilisé l'intervalle de confiance percentile. Par rééchantillonnage des résidus, nous avons utilisé l'intervalle de confiance pivotal avec une méthode de sélection pré-bootstrap S_b CPM. Finalement, par sous-échantillonnage, nous avons utilisé l'intervalle de confiance pivotal avec $b = 50$. Plusieurs tailles sous-échantillonnales pouvaient convenir à notre comparaison, mais nous avons choisi celle qui constituait le meilleur compromis. La figure 3.4.9 donne les pourcentages de couverture pour chacune de ces situations avec les lettres "p" pour indiquer les pourcentages de couverture par rééchantillonnage des paires d'observations, "r" par rééchantillonnage des résidus et "s" par sous-échantillonnage pour les coefficients β_3, β_4 et β_5. Nous avons également inclus les pourcentages de couverture des intervalles de confiance classiques que nous indiquerons par la lettre "c".

Pour une méthode de sélection convergente, S_e BIC, il est préférable d'utiliser l'intervalle de confiance percentile obtenu par rééchantillonnage des paires d'observations aux deux autres méthodes de rééchantillonnage. Le tableau 3.4.3 nous donne les pourcentages de couverture pour un rapport signal-bruit élevé pour chacune des trois méthodes de rééchantillonnage lorsque la matrice de design X_{1000} est utilisée. La ligne "Classique" indique les pourcentages de couverture obtenus par inférence classique, la ligne "Paires" indique les pourcentages de couverture pour l'intervalle de confiance percentile par rééchantillonnage des paires d'observations, la ligne "Résidus", les pourcentages de couverture de l'intervalle de confiance pivotal par rééchantillonnage de résidus avec la méthode S_b CPM

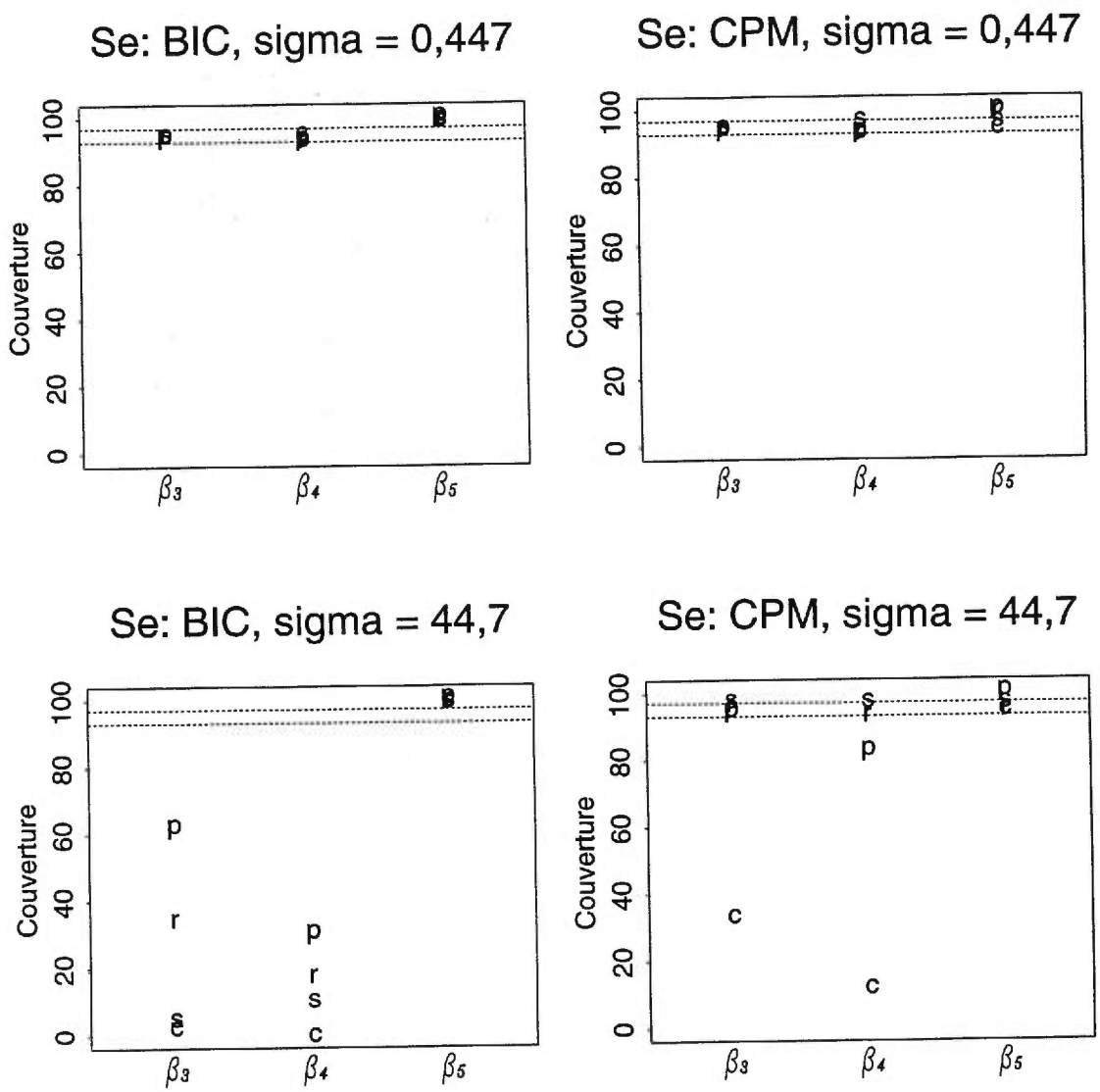

FIGURE 3.4.9. *Pourcentages de couverture des coefficients β_3, β_4 et β_5 pour les méthodes de sélection S_e BIC et CPM de l'intervalle de confiance classique, "c", de l'intervalle de confiance percentile lorsque la méthode du rééchantillonnage des paires, "p" est utilisée, de l'intervalle de confiance pivotal lorsque les méthodes du rééchantillonnage des résidus avec la méthode de sélection S_b CPM, "r", et du sous-échantillonnage avec $b = 50$, "s", sont utilisées.*

et la ligne "Sous-éch.", les pourcentages de couverture de l'intervalle de confiance pivotal par sous-échantillonnage avec $b = 50$. Il n'y a aucune différence entre les pourcentages de couverture des 3 méthodes de rééchantillonnage.

TABLEAU 3.4.3. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit élevé pour la méthode de sélection S_e BIC lorsque le rééchantillonnage des paires, des résidus et le sous-échantillonnage sont utilisés.*

Matrice de design X_{1000}, $\sigma = 0{,}447$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC									
Classique	Uni. gauche 2,5%	1,80	2,60	2,40	2,20	2,80	0,60	0,00	0,00
	Uni. droite 2,5%	2,80	2,40	2,80	2,80	3,60	0,20	0,40	0,00
	Bilatéral 95%	95,40	95,00	94,80	95,00	93,60	99,20	99,60	100,00
Paires	Uni. gauche 2,5%	2,20	2,80	2,40	2,80	2,80	0,00	0,00	0,00
	Uni. droite 2,5%	3,00	2,60	3,00	3,00	3,80	0,00	0,00	0,00
	Bilatéral 95%	94,80	94,60	94,60	94,20	93,40	100,00	100,00	100,00
Résidus	Uni. gauche 2,5%	2,40	2,80	2,60	2,40	3,00	0,60	0,00	0,00
	Uni. droite 2,5%	3,60	2,40	2,60	2,60	3,80	0,20	0,40	0,00
	Bilatéral 95%	94,00	94,80	94,80	95,00	93,20	99,20	99,60	100,00
Sous-éch.	Uni. gauche 2,5%	2,60	1,80	2,20	2,40	1,00	0,00	0,00	0,00
	Uni. droite 2,5%	2,20	1,40	2,40	2,40	3,60	0,20	0,20	0,20
	Bilatéral 95%	95,20	96,80	95,40	95,20	95,40	99,80	99,80	99,80

Le tableau 3.4.4 nous donne les rapports entre les longueurs moyennes des intervalles de confiance pour les méthodes choisies. Les longueurs moyennes des intervalles de confiance pivotal obtenus par sous-échantillonnage sont au dénominateur. La méthode du sous-échantillonnage donne des intervalles de confiance légèrement plus longs que les deux autres méthodes pour les coefficients non nuls. Pour les coefficients nuls, le rééchantillonnage des résidus donne les intervalles de confiance les plus courts et le rééchantillonnage des paires, les plus longs.

Le tableau 3.4.5 nous donne les pourcentages de couverture des intervalles de confiance des méthodes choisies pour un rapport signal-bruit faible et la méthode

TABLEAU 3.4.4. *Rapports entre les longueurs moyennes des intervalles de confiance par rééchantillonnage des paires d'observations, des résidus et par sous-échantillonnage lorsque la méthode S_e BIC est employée. Au dénominateur, les longueurs moyennes des intervalles de confiance pivotal par sous-échantillonnage.*

Matrice de design X_{1000}, S_e BIC								
Coefficients	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes	1,0	2,0	1,5	1,5	0,5	0	0	0
$\sigma = 0{,}447$								
Paires / Sous-éch.	0,97	0,96	0,95	0,94	0,95	1,20	1,35	1,23
Résidus / Sous-éch.	0,97	0,95	0,95	0,95	0,95	0,53	0,60	0,50
$\sigma = 44{,}7$								
Paires / Sous-éch.	1,01	1,87	1,87	1,91	1,99	1,87	1,57	1,70
Résidus / Sous-éch.	1,01	1,46	1,22	1,20	1,02	0,80	0,85	0,63

S_e BIC. Dans ce cas-ci, l'intervalle de confiance percentile obtenu par rééchantillonnage des paires a des pourcentages de couverture plus élevés que les deux autres méthodes. Nous avons déjà expliqué à la section 3.3.4 que la symétrie de la distribution des $\hat{\beta}_i^*$ ainsi que le faible pourcentage d'inclusion des variables par rééchantillonnage des résidus étaient la cause de la différence entre le rééchantillonnage des paires et des résidus. La distribution des $\hat{\beta}_i^\circ$ par sous-échantillonnage est de la même forme que par rééchantillonnage des résidus. En effet, pour le coefficient β_4, il y a 327 intervalles de confiance pivotal $\{0\}$ par sous-échantillonnage. Il devient excessivement difficile de sélectionner une variable lorsque le rapport signal-bruit est faible car par sous-échantillonnage nous utilisons seulement 50 observations ce qui nous donne une variance plus élevée que pour 1000 observations et un rapport signal-bruit effectif encore plus faible que ce que nous étudions.

Si nous regardons le tableau 3.4.4 nous voyons que l'intervalle de confiance percentile par rééchantillonnage des paires est légèrement plus large que les intervalles de confiance des deux autres méthodes.

Malgré le fait que le rééchantillonnage des paires d'observations est la méthode de rééchantillonnage demandant le plus de temps de calcul (tableau 3.4.2),

TABLEAU 3.4.5. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit faible pour la méthode de sélection S_e BIC lorsque le rééchantillonnage des paires, des résidus et le sous-échantillonnage sont utilisés.*

Matrice de design X_{1000}, $\sigma = 44{,}7$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e BIC									
Classique	Uni. gauche 2,5%	2,60	2,20	1,80	3,20	1,40	0,00	0,20	0,20
	Uni. droite 2,5%	2,60	92,60	97,80	94,60	98,60	0,40	0,00	0,40
	Bilatéral 95%	94,80	5,20	0,40	2,20	0,00	99,60	99,80	99,40
Paires	Uni. gauche 2,5%	2,60	0,20	0,00	0,00	0,00	0,00	0,00	0,00
	Uni. droite 2,5%	2,60	27,40	43,80	38,20	70,00	0,00	0,00	0,00
	Bilatéral 95%	94,80	72,40	56,20	61,80	30,00	100,00	100,00	100,00
Résidus	Uni. gauche 2,5%	2,60	2,80	2,00	2,20	1,80	0,60	0,00	0,00
	Uni. droite 2,5%	3,20	49,40	65,40	63,40	80,60	0,20	0,40	0,00
	Bilatéral 95%	94,20	47,80	32,60	34,40	17,60	99,20	99,60	100,00
Sous-éch.	Uni. gauche 2,5%	2,60	2,20	2,20	2,00	0,20	0,00	0,20	0,40
	Uni. droite 2,5%	3,00	89,20	93,40	93,20	89,60	0,60	0,20	0,60
	Bilatéral 95%	94,40	8,60	4,40	4,80	10,20	99,40	99,60	99,00

nous concluons qu'il s'agit de la meilleure méthode de rééchantillonnage afin d'avoir des pourcentages de couverture le plus près possible de la valeur prescrite lorsqu'une méthode de sélection S_e convergente est employée. Nous devons cependant utiliser l'intervalle de confiance percentile. Le sous-échantillonnage et l'inférence classique qui donnent de bons résultats pour un rapport signal-bruit élevé, sont les méthodes les moins performantes pour un rapport signal-bruit faible.

Pour une méthode de sélection non convergente, S_e CPM, il est préférable d'utiliser l'intervalle de confiance pivotal obtenu par sous-échantillonnage aux intervalles de confiance obtenus par les deux autres méthodes de rééchantillonnage. Pour un rapport signal-bruit élevé, toutes les méthodes donnent des pourcentages de couverture qui ne sont pas différents de 95%. Les résultats des méthodes choisies sont exposées dans le tableau 3.4.6. La méthode du sous-échantillonnage

donne des intervalles de confiance légèrement plus larges que les deux autres méthodes, mais une légère augmentation de la taille du sous-échantillon est en mesure de ramener les longueurs des intervalles de confiance par sous-échantillonnage aux longueurs des deux autres méthodes. Le tableau 3.4.7 donne plus de détails sur les ratios des longueurs moyennes des intervalles de confiance des méthodes choisies. Les longueurs moyennes des intervalles de confiance pivotal par sous-échantillonnage sont au dénominateur.

TABLEAU 3.4.6. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit élevé pour la méthode de sélection S_e CPM lorsque le rééchantillonnage des paires, des résidus et le sous-échantillonnage sont utilisés.*

Matrice de design X_{1000}, $\sigma = 0{,}447$								
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e CPM									
Classique	Uni. gauche 2,5%	1,80	2,60	2,40	2,20	2,80	2,60	3,20	1,40
	Uni. droite 2,5%	2,60	2,40	2,80	2,80	3,60	2,80	2,00	2,80
	Bilatéral 95%	95,60	95,00	94,80	95,00	93,60	94,60	94,80	95,80
Paires	Uni. gauche 2,5%	2,20	2,80	2,40	2,80	2,80	0,20	0,00	0,00
	Uni. droite 2,5%	3,20	2,60	3,00	3,00	3,80	0,00	0,20	0,00
	Bilatéral 95%	94,60	94,60	94,60	94,20	93,40	99,80	99,80	100,00
Résidus	Uni. gauche 2,5%	2,00	2,80	2,40	2,40	3,20	3,00	3,80	1,60
	Uni. droite 2,5%	3,20	2,40	2,80	2,60	3,80	2,40	1,80	3,00
	Bilatéral 95%	94,80	94,80	94,80	95,00	93,00	94,60	94,40	95,40
Sous-éch.	Uni. gauche 2,5%	3,00	2,20	1,60	2,40	1,00	1,80	3,20	1,00
	Uni. droite 2,5%	1,40	2,00	2,00	2,20	1,80	1,60	1,80	2,00
	Bilatéral 95%	95,60	95,80	96,40	95,40	97,20	96,60	95,00	97,00

Le sous-échantillonnage est préférable à utiliser lorsque le rapport signal-bruit est plus faible. En effet, pour un rapport signal-bruit faible, nous avons déjà vu que l'intervalle de confiance pivotal par rééchantillonnage des résidus était préférable au rééchantillonnage des paires d'observations en raison de la symétrie des distributions des $\hat{\beta}_i^*$ (section 3.3.4). Encore une fois, la distribution des $\hat{\beta}_i^\circ$ par sous-échantillonnage est la cause de la supériorité des pourcentages

TABLEAU 3.4.7. *Rapports entre les longueurs moyennes des intervalles de confiance par rééchantillonnage des paires d'observations, des résidus et par sous-échantillonnage lorsque la méthode S_e CPM est employée. Au dénominateur, les longueurs moyennes des intervalles de confiance pivotal par sous-échantillonnage.*

Matrice de design X_{1000}, S_e CPM								
Coefficients	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes	1,0	2,0	1,5	1,5	0,5	0	0	0
$\sigma = 0{,}447$								
Paires / Sous-éch.	0,96	0,94	0,94	0,94	0,94	0,84	0,85	0,83
Résidus / Sous-éch.	0,96	0,94	0,94	0,94	0,94	0,94	0,94	0,94
$\sigma = 44{,}7$								
Paires / Sous-éch.	0,98	0,95	0,89	0,91	0,86	0,85	0,86	0,85
Résidus / Sous-éch.	0,98	1.05	0,99	1,00	0,97	0,96	0,96	0,96

de couverture de ce dernier sur les pourcentages de couverture des intervalles de confiance par rééchantillonnage des paires d'observations. Le sous-échantillonnage et le rééchantillonnage des résidus ont des pourcentages de couverture et des longueurs moyennes d'intervalles de confiance comparables. Les résultats sont exposés respectivement dans les tableaux 3.4.8 et 3.4.7. Nous préférons la méthode de sous-échantillonnage car les calculs sont effectués plus rapidement. Pour une taille sous-échantillonnale de 50, il est 63 fois plus rapide d'utiliser la méthode du sous-échantillonnage.

3.4.6. Conclusion

Nous avons vu dans cette section les principaux résultats obtenus lorsque nous utilisons le sous-échantillonnage. Cette méthode trouve son utilité lorsque le nombre d'observations est élevé. En effet, pour un petit nombre d'observations, $n = 50$, les pourcentages de couverture sont inférieurs à la valeur prescrite et ce, même pour un rapport signal-bruit élevé.

TABLEAU 3.4.8. *Pourcentages de couverture des intervalles de confiance pour un rapport signal-bruit faible pour la méthode de sélection S_e CPM lorsque le rééchantillonnage des paires, des résidus et le sous-échantillonnage sont utilisés.*

Matrice de design X_{1000}, $\sigma = 44{,}7$									
Coefficients		β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Valeurs exactes		1,0	2,0	1,5	1,5	0,5	0	0	0
S_e CPM									
Classique	Uni. gauche 2,5%	2.20	2.00	2.40	3.20	2.20	2.40	3.00	3.00
	Uni. droite 2,5%	2.60	51.60	65.80	63.40	85.80	2.60	3.20	1.40
	Bilatéral 95%	95.20	46.40	31.80	33.40	12.00	95.00	93.80	95.60
Paires	Uni. gauche 2,5%	2.60	2.20	0.60	1.40	0.00	0.00	0.00	0.00
	Uni. droite 2,5%	2.00	3.00	3.60	4.20	17.60	0.00	0.00	0.00
	Bilatéral 95%	95.40	94.80	95.80	94.40	82.40	100.00	100.00	100.00
Résidus	Uni. gauche 2,5%	2,20	2,80	2,20	2,20	3,00	2,60	3,40	1,60
	Uni. droite 2,5%	3,40	0,20	0,40	1,00	3,80	2,60	1,40	3,40
	Bilatéral 95%	94,40	97,00	97,40	96,80	93,20	94,80	95,20	95,00
Sous-éch.	Uni. gauche 2,5%	3,00	1,20	1,00	1,60	1,20	1,20	2,00	3,60
	Uni. droite 2,5%	3,00	0,20	0,60	0,60	1,20	1,80	1,80	1,60
	Bilatéral 95%	94,00	98,60	98,40	97,80	97,60	97,00	96,20	94,80

Par la suite, nous avons utilisé un jeu de données de grande taille, soit $n = 1000$. Pour une méthode de sélection S_e convergente, la méthode du sous-échantillonnage performe moins bien que les autres méthodes de rééchantillonnage et sensiblement de la même façon que l'inférence classique. En effet, pour une taille de sous-échantillon fixe, aucun intervalle de confiance ne permet de supplanter les intervalles de confiance classiques. L'intervalle de confiance percentile fait mieux que l'intervalle de confiance classique pour un rapport signal-bruit élevé lorsque nous utilisons une petite taille sous-échantillonnale, mais moins bien que ce dernier pour la même taille de sous-échantillon et un rapport signal-bruit moyen. Pour un rapport signal-bruit élevé, les intervalles de confiance pivotal et bootstrap-t-MSE ne sont pas différents des intervalles de confiance classiques.

Pour une méthode de sélection non convergente, le sous-échantillonnage fait mieux que l'inférence classique. Pour un rapport signal-bruit élevé, les pourcentages de couverture sont les mêmes, mais pour un rapport signal-bruit faible, les pourcentages de couverture des intervalles de confiance pivotal et bootstrap-t-MSE par sous-échantillonnage sont supérieurs pour toutes les tailles sous-échantillonnales étudiées. L'intervalle de confiance percentile est très instable selon le coefficient observé et le rapport signal-bruit. Il n'est, par conséquent, pas recommandé.

Puisqu'en général les pourcentages de couverture diminuent lorsque la taille du sous-échantillon augmente, nous avons utilisé une taille sous-échantillonnale de 50 et l'intervalle de confiance pivotal afin de comparer le sous-échantillonnage aux deux autres méthodes. Nous avons conclu que pour une méthode de sélection S_e convergente il est préférable d'utiliser l'intervalle de confiance percentile obtenu par rééchantillonnage des paires dont les pourcentages de couverture sont plus élevées. Pour une méthode de sélection S_e non convergente, bien que le sous-échantillonnage et le rééchantillonnage des résidus donnent des résultats pratiquement identiques, nous avons avantagé le sous-échantillonnage en raison de sa plus grande rapidité de calcul.

CONCLUSION

Dans ce mémoire, nous nous sommes intéressés à la construction d'intervalles de confiance en régression linéaire multiple suite à la sélection de modèle. De façon "classique", nous avons vu que lorsque nous ne sélectionnons pas une variable, l'intervalle de confiance du coefficient associé à cette variable est l'ensemble $\{0\}$. Cette façon de procéder ne tient pas compte du caractère aléatoire de la sélection qui dépend des observations.

Pour remédier au problème, Carignan (1996) a suggéré d'utiliser le rééchantillonnage. Nous avons repris sa méthode qui consiste à sélectionner et à estimer un modèle par rééchantillonnage des résidus. Nous sélectionnons tout d'abord, comme dans le cas classique, un modèle $\hat{A}$ basé sur une méthode de sélection à évaluer S_e. Nous évaluons ensuite l'estimateur de $\underset{\sim}{\beta}$, $\underset{\sim}{\hat{\beta}}_{\hat{A}}$. Afin de calculer les quantiles bootstrap, nous procédons comme suit. Nous sélectionnons un modèle $\tilde{A}$ par une méthode de sélection S_b qui peut être différente de S_e. Nous calculons ensuite $\underset{\sim}{\hat{\beta}}_{\tilde{A}}$, l'estimateur du vecteur des coefficients pour ce modèle bootstrap. Nous calculons ensuite le vecteur des résidus $\underset{\sim}{\hat{\epsilon}}_{\tilde{A}}$ par la formule $\underset{\sim}{\hat{\epsilon}}_{\tilde{A}} = \underset{\sim}{y} - X_{\tilde{A}}\underset{\sim}{\hat{\beta}}_{\tilde{A}}$. Nous générons par la suite un grand nombre de vecteurs d'erreurs bootstrap en rééchantillonnant, avec remise, les résidus du modèle sélectionné avec la méthode de sélection S_b, que nous additionnons à $X_{\tilde{A}}\underset{\sim}{\hat{\beta}}_{\tilde{A}}$ afin d'obtenir les vecteurs $\underset{\sim}{y}^*$ d'observations bootstrap correspondants. Nous utilisons alors $X_{\tilde{A}}$ et $\underset{\sim}{y}^*$ pour sélectionner un modèle $\hat{A}^*$ par la méthode de sélection S_e. Nous calculons ensuite l'estimateur $\underset{\sim}{\hat{\beta}}^*_{\hat{A}^*}$, l'estimateur de $\underset{\sim}{\hat{\beta}}_{\tilde{A}}$, qui servira aux calculs des quantiles bootstrap.

Il ne s'agit pas de la seule façon d'appliquer le rééchantillonnage. Nous avons considéré deux nouvelles méthodes d'inférence: le rééchantillonnage des paires d'observations et le sous-échantillonnage.

Le rééchantillonnage des paires d'observations consiste à rééchantillonner, avec remise, directement les lignes de la matrice $Z = (X, \underset{\sim}{y})$, la matrice formée des variables indépendantes et dépendantes. Nous débutons comme dans le cas classique par sélectionner un modèle $\hat{A}$ par la méthode de sélection S_e. Nous évaluons ensuite $\underset{\sim}{\hat{\beta}}_{\hat{A}}$. Pour le calcul des quantiles bootstrap, nous tirons, avec remise, de nombreux échantillons de taille n. Nous obtenons des matrices $Z^* = (X^*, \underset{\sim}{y}^*)$ pour chacun de ces échantillons. Ces matrices, X^*, et vecteurs, $\underset{\sim}{y}^*$, nous serviront à obtenir les quantiles bootstrap suite à la sélection d'un modèle $\hat{A}^*$ par la méthode de sélection S_e et à l'estimation du vecteur $\underset{\sim}{\hat{\beta}}^*_{\hat{A}^*}$.

Le sous-échantillonnage ressemble au rééchantillonnage des paires d'observations. Après avoir estimé $\hat{A}$ et $\hat{\beta}_{\hat{A}}$, nous calculons les quantiles par sous-échantillonnage en tirant, sans remise, un sous-échantillon de taille inférieure au nombre d'observations mais supérieure au nombre de variables indépendantes. Nous obtenons des matrices $Z^\circ = (X^\circ, \underset{\sim}{y}^\circ)$ qui serviront aux calculs de $\hat{A}^\circ$ et $\underset{\sim}{\hat{\beta}}^\circ_{\hat{A}^\circ}$ et à l'obtention des quantiles par sous-échantillonnage qui serviront aux calculs des intervalles de confiance. Un ajustement tenant compte de la taille des échantillons et des sous-échantillons, de même que de la vitesse de convergence de l'estimateur doit être fait.

Nous voulions découvrir si les pourcentages de couverture des intervalles de confiance obtenus par une de ces méthodes étaient supérieurs aux pourcentages de couverture des intervalles de confiance obtenus par la méthode proposée par Carignan. Pour ce faire, nous avons utilisé des jeux de données de petite et de grande taille. Nous avons également considéré des méthodes de sélection S_e convergentes

(méthodes de Ducharme, 1997 et celle de BIC, Schwarz, 1978) et non convergentes (C_p de Mallows, 1973, méthodes d'addition et de retrait par étapes).

Nos simulations nous ont permis de classer les diverses méthodes de rééchantillonnage que nous avons étudiées selon la qualité de la couverture des différents intervalles de confiance bootstrap.

Pour un jeu de données de petite taille, $n = 50$ avec 7 prédicteurs plus une constante, la méthode du sous-échantillonnage ne donne pas de bons résultats. Le ratio b/n, la taille sous-échantillonale sur le nombre d'observations, trop élevé et l'obligation de choisir une taille sous-échantillonnale supérieure au nombre de variables contenues dans le modèle explique la piètre performance de cette méthode.

Pour une méthode de sélection S_e convergente, l'intervalle de confiance percentile obtenu par rééchantillonnage des paires d'observations est l'intervalle de confiance obtenu par rééchantillonnage qui donne les pourcentages de couverture les plus élevés. En fait, pour un rapport signal-bruit élevé, les deux méthodes de rééchantillonnage et les intervalles de confiance classiques ont des pourcentages de couverture identiques. L'inférence classique nous permet même d'obtenir les intervalles de confiance les plus courts. Cependant, pour un rapport signal-bruit faible, les pourcentages de couverture des intervalles de confiance classiques sont très faibles. Puisque nous ne connaissons pas *a priori* le rapport signal-bruit, nous devons avantager le rééchantillonnage qui donne des pourcentages de couverture beaucoup plus élevés pour un rapport signal-bruit faible.

La forme de la distribution des $\hat{\beta}^*_{i\hat{A}^*}$ et le faible taux d'inclusion des différentes variables expliquent la supériorité du rééchantillonnage des paires d'observations.

Ces distributions expliquent également pourquoi l'intervalle de confiance percentile est supérieur aux deux autres intervalles de confiance bootstrap. Les intervalles de confiance pivotal et bootstrap-t-MSE sont construits dans la mauvaise direction lorsque le coefficient $\hat{\beta}_{i\hat{A}}$ est nul.

Si nous devons utiliser le rééchantillonnage des résidus, l'intervalle de confiance bootstrap employé a peu d'importance lorsqu'une méthode de pré-sélection bootstrap est employée. Par contre, si le modèle complet est utilisé, nous pouvons obtenir des pourcentages de couverture inférieurs à ceux obtenus lorsqu'une méthode de sélection S_b est employée pour les intervalles de confiance pivotal et bootstrap-t-MSE.

Pour une méthode de sélection S_e non convergente, les taux d'inclusion plus élevés des variables augmentent les pourcentages de couverture des méthodes de rééchantillonnage. Le rééchantillonnage des résidus parvient même à détrôner le rééchantillonnage des paires d'observations. Tous les intervalles de confiance boostrap donnent des résultats identiques par rééchantillonnage des résidus lorsqu'une méthode de sélection S_b est employée. Si le modèle complet est employé, les pourcentages de couverture sont inférieurs à la valeur prescrite et inférieure aux pourcentages de couverture des intervalles de confiance obtenus par rééchantillonnage des paires d'observations. Si cette dernière méthode doit être utilisée, l'intervalle de confiance percentile demeure le meilleur intervalle de confiance bootstrap.

Lorsque nous utilisons un jeu de données de grande taille, $n = 1000$ toujours avec 7 prédicteurs plus une constante, le temps de calcul devient alors un facteur important à considérer. Tout d'abord, les conclusions tirées pour $n = 50$ observations demeurent les mêmes pour $n = 1000$ observations pour les méthodes

de rééchantillonnage des paires d'observations et des résidus. Les meilleurs intervalles de confiance pour ces méthodes sont respectivement l'intervalle de confiance percentile et pivotal.

Puisque le ratio b/n peut être plus faible lorsque $n = 1000$, le sous-échantillonnage, dont le temps de calcul est réduit, devient une méthode d'inférence intéressante à utiliser. Son comportement et les distributions des estimateurs $\hat{\beta}^{\circ}_{i\hat{A}^{\circ}}$ sont similaires au comportement et aux distributions des estimateurs $\hat{\beta}^{*}_{i\hat{A}^{*}}$ obtenus par rééchantillonnage des résidus. Puisqu'en général les pourcentages de couverture décroissent lorsque la taille du sous-échantillon augmente, nous avons établi la taille optimale de ce dernier à $b = 50$. De plus, l'intervalle de confiance percentile est très instable. Pour certains coefficients, les pourcentages de couverture peuvent être croissants tandis que pour d'autres, ils décroissent. Nous avons donc choisi d'utiliser l'intervalle de confiance pivotal afin de comparer cette méthode aux deux autres méthodes de rééchantillonnage.

Pour une méthode de sélection S_e convergente, le sous-échantillonnage ne fait pas mieux que le rééchantillonnage des résidus. En fait, il ne fait guère mieux que l'inférence classique. L'intervalle de confiance percentile obtenu par rééchantillonnage des paires d'observations demeure donc le meilleur choix. Cette méthode obtient les pourcentages de couverture les plus élevés.

Pour une méthode de sélection S_e non convergente, les pourcentages de couverture des intervalles de confiance pivotal obtenu par rééchantillonnage des résidus et par sous-échantillonnage sont essentiellement les mêmes. Ils sont tous deux supérieurs aux pourcentages de couverture obtenus par rééchantillonnage des paires d'observations. Cependant, le temps de calcul requis pour une répétition du programme de sous-échantillonnage est 63 fois plus rapide que le programme de rééchantillonnage des résidus. Pour cette raison, nous privilégions le

sous-échantillonnage pour une méthode de sélection S_e non convergente avec un grand nombre d'observations.

Il serait intéressant de pouvoir considérer une matrice de variables explicatives entièrement aléatoire. Bien que les méthodes de rééchantillonnage des paires d'observations et du sous-échantillonnage sont issues d'un modèle aléatoire, les simulations effectuées utilisaient tout de même une matrice fixe. Une autre avenue de recherche serait d'augmenter le nombre de variables indépendantes et le nombre d'observations. Ceci nous permettrait d'espérer de futures applications au "data mining", particulièrement pour la méthode du sous-échantillonnage.

Annexe A

PROGRAMMES INFORMATIQUES

```
************************************************************************
*                                                                      *
*     Suite a la selection d'un modele de regression, ce programme calcule *
*     le nombre de modeles trop petits, corrects, trop grands selon les  *
*     methodes dites classique et bootstrap. Le bootstrap s'effectue en  *
*     reechantillonnant les paires X et Y. On nous donne egalement le    *
*     nombre de matrices (X) qui ne sont pas de plein rang avant la      *
*     selection de modele et apres la selection de modele.               *
*     La longueur des intervalles de confiance est aussi donnee selon    *
*     4 types d'intervalles de confiance:                                *
*     classique, bootstrap percentile, pivotal et t-mse.                 *
*                                                                        *
*     On doit donner une matrice X, specifier le nombre d'observations   *
*     (NOBS), le nombre de variables incluant la dependante et l'intercept *
*     (NVAR), la methode de selection parmi les 5 methodes:              *
*     Cp de Mallow (CPM), de Ducharme (DUC), BIC (BIC), ajout par etapes *
*     (FWD) et retrait par etapes (BWD).                                 *
*                                                                        *
*     Parametres a changer uniquement dans le programme principal:       *
*     MAXBT, MAXREP, ALPHA, STDDEV, NBOOT, NREP.                         *
*                                                                        *
*     Les parametres sont a modifier dans le programme principal et dans *
*     la sous-routine REG (PMAX ET NMAX).                                *
*                                                                        *
*     Ce programme utilise la librairie NAG_MARK16 de Fortran.           *
*                                                                        *
************************************************************************

      PROGRAM PAIRES
*
*     Parametres
*
      INTEGER           MAXR, MAXW1, MAXW2, MAXW3,
```

```
     *                MAXBT, MAXREP, PMAX, NMAX, LWORK, NBOOT, NREP

      CHARACTER *10     DONNEES
      DOUBLE PRECISION  ALPHA, STDDEV
************************************************************************
*     Parametres a changer (globaux)
*
      PARAMETER         (PMAX = 16, NMAX = 1000)
      PARAMETER         (MAXBT = 1000, MAXREP = 500)
      PARAMETER         (DONNEES = "matrice1")
      PARAMETER         (ALPHA = .05, STDDEV = .1)
      PARAMETER         (NBOOT = 1000, NREP = 2)
************************************************************************
*2222222222222222222222222222222222222222222222222222222222222222222222*
*2    Parametre a ajouter pour le programme du sous-echantillonnage      *
*2    PARAMETER         (NSUB = 200)                                     *
*2222222222222222222222222222222222222222222222222222222222222222222222*
************************************************************************

************************************************************************
*                                                                      *
*     PMAX:    Nombre maximal de variables dans le modele incluant l'ordonne *
*              et la variable expliquee.                               *
*     NMAX:    Nombre maximal d'observations.                          *
*     MAXBT:   Nombre maximal de repetitions bootstrap.                *
*     MAXREP:  Nombre maximal de repetitions.                          *
*     DONNEES: Nom du jeu de donnees utilisees.                        *
*     ALPHA:   1-ALPHA est le niveau de confiance.                     *
*     STDDEV:  Ecart-type des erreurs.                                 *
*     NBOOT:   Nombre de repetitions bootstrap                         *
*     NREP:    Nombre de repetitions.                                  *
************************************************************************

*     Parametres pour CHOICP (ALLRQR et DGEQRF)
      PARAMETER         (MAXR = (PMAX+1)*(PMAX+2)/2,
     *                   MAXW1 = (PMAX-3)*(3+MAXR/3),
     *                   MAXW2 = (PMAX-3)*(PMAX+2)/2,
     *                   MAXW3 = (PMAX+1)*PMAX/2,
     *                   LWORK = 64*NMAX)

      IMPLICIT DOUBLE PRECISION (A-H, O-Z)

*     Variables locales
      INTEGER           IFAIL, IND, NVAR,NOBS,
     *                  X10, X11, X12, X20, X21, X22, I, J, K, KK,
     *                  V10, V11, V12, V20, V21, V22
```

```
************************************************************************
*                                                                      *
*     IND:      Nombre de variables contenues dans le modele selectionne     *
*               (incluant l'ordonnee).                                 *
*     NVAR:     Nombre de variables dans le modele incluant l'ordonnee et    *
*               la variable expliquee.                                 *
*     NOBS:     Nombre d'observations.                                 *
*     X**:      Pour generer les nombres aleatoires.                   *
*     V**:      Pour generer les nombres aleatoires.                   *
************************************************************************

      DOUBLE PRECISION MSE, QUANT, QT, MSECL
************************************************************************
*                                                                      *
*     MSE:      MSE du modele bootstrap                                *
*     MSECL:    MSE du modele classique                                *
*     QUANT:    Fonction pour le calcul de quantile.                   *
*     QT:       Quantiles de DCOEF ET DCOEFM                           *
************************************************************************


      DOUBLE PRECISION ABOOT(NMAX,PMAX),AORIG(NMAX, PMAX),A(NMAX, PMAX),
     *                 BETA(PMAX),MU(NMAX)

      INTEGER          BETAL(PMAX)

      CHARACTER        SELECT*3
************************************************************************
*                                                                      *
*     AORIG:  Matrice (X,Y) originale                                  *
*     A     : Matrice (X,Y) servant au calcul de beta chapeau classique    *
*     ABOOT:  Matrice (X,Y) servant au calcul de beta chapeau bootstrap    *
*     MU:     Resultat de X*BETA                                       *
*     BETA:   Vecteur des vraies valeurs de beta                       *
*     BETAL:  Vecteur de 0 et de 1, 1 si la vraie valeur de beta est   *
*             differente de 0.0                                        *
*     SELECT: Methode de selection de modeles:                         *
*             CPM: Cp de Mallow, DUC: Cp de Ducharme, BIC: BIC,        *
*             FWD: Forward, BWD: Backward                              *
*                                                                      *
************************************************************************

      DOUBLE PRECISION DCOEF(MAXBT,PMAX), DCOEFM(MAXBT,PMAX),
     *                 QBRL(PMAX,2),QBRU(PMAX,2),LONGIC(PMAX,MAXREP,4),
     *                 SUMIDC(PMAX, 4), SSQIDC(PMAX,4), MOYIDC(PMAX,4),
     *                 SEIDC(PMAX,4)
```

```
************************************************************************
*                                                                      *
*     DCOEF:  Difference entre COEF et COEFO                           *
*     DCOEFM: DCEOF divise par MSE du modele bootstrap choisi          *
*     QBRL:   Quantile inferieur de DCEOF ET DCOEFM                    *
*     QBRU:   Quantile superieur de DCEOF ET DCOEFM                    *
*     LONGIC: Longueur des 4 types d'intervalles de confiance.         *
*     SUMIDC: Somme de la longueur des intervalles de confiance        *
*     SSQIDC: Somme du carre des longueurs des intervalles de confiance*
*                                                                      *
************************************************************************

      INTEGER          IBOOT(NMAX), QLT(3), QLTBT(3), QLTCL(3),
     *                 CVT(PMAX, 3), CPTCVT(4, PMAX, 3), CTBETA(PMAX),
     *                 CTIND(PMAX)
************************************************************************
*                                                                      *
*     IBOOT:  Vecteur des indices de l'echantillon                     *
*     QLT:    Vecteur de 0 et 1 indiquant si le modele est trop petit, *
*             correct, ou trop grand (un seul 1 par vecteur)           *
*     QLTBT:  Compteur de QLT pour le modele bootstrap.                *
*     QLTCL:  Compteur de QLT pour le modele classique.                *
*     CVT:    Vecteur de 0 et 1 indiquant si les vraies valeurs de beta*
*             sont a gauche, a droite ou entre les bornes de l'intervalle*
*             de confiance.                                            *
*     CPTCVT: Compteur de CVT pour les 4 types d'IDC (classique, pivotal,*
*             percentile, t-mse.                                       *
*     CTBETA: Compteur de la presence de la variable i bootstrap       *
*     CTIND:  Compteur de la taille des modeles selectionnes bootstrap *
*                                                                      *
************************************************************************

      INTEGER          CTRANK, CTRK2, CTCLR, CTCLR2
************************************************************************
*                                                                      *
*     CTRANK: Compteur pour le nombre de matrice X de rang incomplet,  *
*             avant la selection de modele, methode bootstrap.         *
*     CTRK2:  Meme compteur que precedemment, apres la selection du modele*
*     CTCLR:  Idem a CTRANK pour le modele classique                   *
*     CTCLR2: Idem a CTRK2, pour le modele classique                   *
*                                                                      *
************************************************************************

      DOUBLE PRECISION COEF(PMAX), COEFO(PMAX), SECOEF(PMAX),
     *                 BINFCL(PMAX),BSUPCL(PMAX),
```

```
*                BINFPV(PMAX), BSUPPV(PMAX),
*                BINFPC(PMAX), BSUPPC(PMAX),
*                BINFTM(PMAX), BSUPTM(PMAX)
************************************************************************
*                                                                      *
*     COEF:   Estimation de beta etoile                                *
*     COEFO:  Estimation de beta                                       *
*     SECOEF: Ecart-type de beta chapeau                               *
*     BINF**: Bornes inferieures des intervalles de confiance pour les *
*             methodes: CL: classique, PV: pivotal, PC: percentile,    *
*             TM: t-mse.                                               *
*     BSUP**: Bornes superieures des intervalles de confiance.         *
************************************************************************

************************************************************************
*1111111111111111111111111111111111111111111111111111111111111111111111*
*1                                                                     *
*1  Variables a ajouter pour l'execution du programme du reechantillonnage *
*1  des residus.                                                       *
*1                                                                     *
*1    DOUBLE PRECISION MUSTAR(NMAX), COEFPB(PMAX)                      *
*1    CHARACTER        SEVAL*3, SBOOT*3                                *
*1    INTEGER          QLTPB(3), PBBETA(PMAX), PBIND(PMAX)             *
*1                                                                     *
*1                                                                     *
*1     MUSTAR: Resultat de X*COEFPB                                    *
*1     COEFPB: Estimation de beta pre-bootstrap                        *
*1     SEVAL : Methode de selection de modeles servant a l'evaluation: *
*1     SBOOT:  Meme chose que SEVAL, mais pour la determination des erreurs *
*1             a reechantilloner. Aucune selection de modele est aussi *
*1             disponible (NON)                                        *
*1     QLTPB:  Compteur de QLT pour le modele pre-bootstrap.           *
*1     PBIND:  Compteur de la taille des modeles selectionnes pre-bootstrap *
*1     PBBETA: Compteur de la presence de la variable i pre-bootstrap  *
*1                                                                     *
*1111111111111111111111111111111111111111111111111111111111111111111111*
************************************************************************

************************************************************************
*2222222222222222222222222222222222222222222222222222222222222222222222*
*2                                                                     *
*2  Variables a ajouter pour l'execution du programme du              *
*2  sous-echantillonnage                                              *
*2                                                                     *
*2    INTEGER          NSUB, ECH(NMAX)                                *
*2                                                                     *
```

```
*2     NSUB:    Taille des echantillons                                    *
*2     ECH:     Indices de 1 a NOBS                                        *
*2                                                                         *
*22222222222222222222222222222222222222222222222222222222222222222222222*
*************************************************************************

*     Variables de travail (programme principal)
      DOUBLE PRECISION WORK5(MAXBT), WORK6(MAXBT), WORK7(MAXBT),
     *                 WORK8(MAXBT), TEMP(MAXBT), YTMP, DF


*     Variables specifiques a REG.

      INTEGER          ISX(PMAX), VARMOD(PMAX)
*************************************************************************
*                                                                       *
*     ISX:    Vecteur de 0,1,2 indiquant quelles variables a considerer *
*             initialement dans le modele. 0: Variable non-considere,   *
*             1: considere, 2: force a etre comprise dans le modele.    *
*     VARMOD: Vecteur de 0 et de 1 indiquant les variables composant le *
*             modele selectionne.                                       *
*                                                                       *
*************************************************************************

      DOUBLE PRECISION WT(NMAX), B(PMAX),SE(PMAX),
     *                 COV((PMAX*PMAX+PMAX)/2),RES(NMAX)
*************************************************************************
*     WT:     Vecteur de poids.                                         *
*     B:      Coefficients non-nuls de la regression                    *
*     SE:     Ecart-type des coefficients beta chapeau etoile.          *
*     COV:    Matrice de variance-covariance de beta chapeau ou beta chapeau *
*             etoile.                                                   *
*     RES:    Vecteur des residus.                                      *
*************************************************************************

*************************************************************************
*11111111111111111111111111111111111111111111111111111111111111111111111*
*1                                                                       *
*1  Variables a ajouter pour l'execution du programme du reechantillonnage *
*1  des residus.                                                         *
*1                                                                       *
*1    DOUBLE PRECISION RESPB(NMAX)                                       *
*1                                                                       *
*1    RESPB: Vecteur des residus suite a la selection pre-bootstrap      *
*1                                                                       *
*11111111111111111111111111111111111111111111111111111111111111111111111*
```

```
************************************************************************

*     Variables de travail pour REG
      DOUBLE PRECISION H(NMAX),P(PMAX*(PMAX+2)),WK(NMAX)

*     Variables specifiques a BKWARD.

*     Variables de travail pour BKWARD
      DOUBLE PRECISION WKBK(5*(PMAX-1)+PMAX*PMAX), Q1(NMAX,PMAX+1),
     *                 QTMP(NMAX,PMAX+1), QMIN(NMAX,PMAX+1)


*     Variables specifiques a FRWARD
      INTEGER          IMVB(PMAX)
      CHARACTER*1      FREE(PMAX), MODEL(PMAX)
************************************************************************
*                                                                      *
*     IMVB:  Numeros des variables selectionnees.                      *
*     MODEL: Noms des variables selectionnees.                         *
************************************************************************

*     Variables de travail pour FRWARD
      DOUBLE PRECISION EXSS(PMAX),PFW(PMAX+1), WKFW(2*PMAX)

*     Variables specifiques a CHOICP
      INTEGER          IMVBCP(MAXW3)
      DOUBLE PRECISION R(MAXR), CP(PMAX), BRSS(PMAX)
************************************************************************
*                                                                      *
*     IMVBCP: Numeros des variables selectionnees.                     *
*     R:      Matrice R de la decomposition QR sous la forme de Stirling   *
*     CP:     Vecteur contenant les valeurs de CP (Mallow, Ducharme ou BIC) *
*             pour le meilleur modele de taille p.                     *
*     BRSS:   RSS pour le meilleur modele de taille p.                 *
************************************************************************

*     Variables de travail pour CHOICP
      INTEGER          IW1(MAXW2), IW2(PMAX), IW3(PMAX,2)
      DOUBLE PRECISION TAU(NMAX),WORK(LWORK),
     *                 W1(MAXW1),W2(MAXW2), W3(MAXW3), W4(PMAX,3)
*
      EXTERNAL         RNOR, RANDOM, REG, G01FBF, COUVRT

      OPEN(2,FILE = DONNEES)
      OPEN(3,FILE= "~/memoire/resultats/paires/try/res2.tex")
```

```
      OPEN(4,FILE= "~/memoire/resultats/paires/try/idccl.44")
      OPEN(5,FILE= "~/memoire/resultats/paires/try/idcpv.44")
      OPEN(6,FILE= "~/memoire/resultats/paires/try/idcpc.44")
      OPEN(7,FILE= "~/memoire/resultats/paires/try/idctm.44")
      OPEN(8,FILE ="~/memoire/resultats/paires/try/beta.44")
      OPEN(9,FILE ="~/memoire/resultats/paires/try/betas.44")
      OPEN(10,FILE="~/memoire/resultats/paires/try/b0.44")
      OPEN(11,FILE="~/memoire/compte")

*     Initialiser X10, X11, X12, X20, X21, X22 pour la generation
*     de nombres aleatoires

      DATA X10/1715124249/
      DATA X11/629026336/
      DATA X12/1117079947/
      DATA X20/932201972/
      DATA X21/1785785812/
      DATA X22/1827075957/

*     Initialiser V10, V11, V12, V20, V21, V22 pour la generation
*     de nombres aleatoires

      DATA V10/1715124241/
      DATA V11/629026338/
      DATA V12/1117079943/
      DATA V20/932201971/
      DATA V21/1785785811/
      DATA V22/1827075953/


*
*  Execution
*
*     Lire les constantes et la methode de selection

C*************************************************************************
      READ (2,*) NVAR, NOBS, SELECT                                  *  1A *
      IF (SELECT.NE.'CPM'.AND.SELECT.NE.'DUC'.AND.SELECT.NE.'BIC'.AND.   *  1A *
     *    SELECT.NE.'BWD'.AND.SELECT.NE.'FWD') THEN                  *  1A *
         WRITE(3,*)'Vous devez choisir parmi les methodes de selection:' *  1A *
         WRITE(3,*)'BWD, FWD, CPM, DUC ou BIC'                       *  1A *
         STOP                                                        *  1A *
      END IF                                                         *  1A *
C*************************************************************************

*     Lire le modele
```

```
      READ (2,*) (BETA(I), I = 1, NVAR-1)
*
*     Lire X

      READ (2,*) ((AORIG(I,J),J=1,NVAR-1),I=1,NOBS)

      WRITE(3,*) '\\documentclass{article}'
      WRITE(3,*) '\\begin{document}'
      WRITE(3,*) '% RESULTATS DE LA SIMULATION'
      WRITE(3,*)
      WRITE(3,*) '% JEU DE DONNEES EMPLOYE: ', DONNEES
      WRITE(3,*)
      WRITE(3,*) '% NOMBRE OBSERVATIONS: ',NOBS
      WRITE(3,*) '% NOMBRE DE COEFFICIENTS A ESTIMER: ', NVAR-1
      WRITE(3,99997) '% VRAI MODELE: ', (BETA(I), I= 1, NVAR-1)
      WRITE(3,*)
      WRITE(3,*) '% ERREURS: '
      WRITE(3,*) '% DISTRIBUTION: NORMALE'
      WRITE(3,99998) '% ECART-TYPE: ' , STDDEV
      WRITE(3,*)
      WRITE(3,99998) '% NIVEAU DE CONFIANCE: ', (1-ALPHA)*100, '%'
      WRITE(3,*) '% NOMBRE DE REPETITIONS: ' , NREP
      WRITE(3,*) '% NOMBRE DE REPETITIONS BOOTSTRAP: ', NBOOT
      WRITE(3,*)
      WRITE(3,*) '% METHODE DE SELECTION: ', SELECT
*************************************************************************
*2222222222222222222222222222222222222222222222222222222222222222222222*
*2                                                                      2*
*2    Ajouter pour le programme de sous-echantillonnage                 2*
*2                                                                      2*
*2    WRITE(3,*) '% TAILLE DES SOUS-ECHANTILLONS: ', NSUB               2*
*2    WRITE(3,*) '% NSUB/NOBS: ', REAL(NSUB)/REAL(NOBS)                 2*
*2                                                                      2*
*2222222222222222222222222222222222222222222222222222222222222222222222*
*************************************************************************


C     Initialiser les compteurs de rang plein
C     (Ces compteurs ne sont pas necessaires pour l'execution du
C      programme de reechantillonnage des residus)

C     Classique avant/apres
      CTCLR = 0
      CTCLR2 = 0
C     Bootstrap avant/apres
```

```
      CTRANK = 0
      CTRK2 = 0

C     Initialiser le compteur de beta
C
      DO 1 I = 1, NVAR-1
         CTBETA(I) = 0
 1    CONTINUE

C     Initialiser les compteurs pour la qualite du modele
C
C     Bootstrap
      QLTBT(1) = 0
      QLTBT(2) = 0
      QLTBT(3) = 0
C     Classique
      QLTCL(1) = 0
      QLTCL(2) = 0
      QLTCL(3) = 0
*********************************************************************************
*1111111111111111111111111111111111111111111111111111111111111111111111111111*
*1    Ajouter le compteur pre-bootstrap et pbbeta pour le programme de         *
*1    reechantillonnage des residus.                                           *
*1                                                                             *
*1    QLTPB(1) = 0                                                             *
*1    QLTPB(2) = 0                                                             *
*1    QLTPB(3) = 0                                                             *
*1                                                                             *
*1    DO 2 I = 1, NVAR-1                                                       *
*1       PBBETA(I) = 0                                                         *
*1 2  CONTINUE                                                                 *
*1                                                                             *
*1111111111111111111111111111111111111111111111111111111111111111111111111111*
*********************************************************************************

*     Initialiser BETAL
      DO 3 I = 1, NVAR-1
         IF (BETA(I) .EQ. 0.0) THEN
            BETAL(I) = 0
         ELSE
            BETAL(I) = 1
         END IF
 3    CONTINUE

C     Initialiser compteur pour la qualite de couverture
```

```
      DO 10 I = 1,4
         DO 20 J = 1, NVAR-1
            DO 30 K = 1, 3
               CPTCVT(I,J,K) = 0
 30         CONTINUE
 20      CONTINUE
 10   CONTINUE

*************************************************************************
*2222222222222222222222222222222222222222222222222222222222222222222222*
*2                                                                       *
*2  Ajouter le code suivant pour initialiser la variable ECH pour le    *
*2  programme du sous-echantillonnage                                    *
*2                                                                       *
*2*     Initialiser ECH                                                 *
*2                                                                       *
*2      DO 35 I = 1, NOBS                                               *
*2         ECH(I) = I                                                   *
*2 35   CONTINUE                                                        *
*2                                                                       *
*2222222222222222222222222222222222222222222222222222222222222222222222*
*************************************************************************

C     Calculer MU

      DO 40 I = 1, NOBS
         YTMP = 0.0
         DO 50 J = 1, NVAR-1
            YTMP = YTMP + BETA(J) * AORIG(I,J)
 50      CONTINUE
         MU(I) = YTMP
 40   CONTINUE

C     Repeter NREP fois

      DO 60 KK = 1, NREP

         write(11,*) kk
C     Calculer Y avec erreurs N(MU, STDDEV)

         DO 70 I = 1, NOBS
            AORIG(I,NVAR) = MU(I)+STDDEV*RNOR(X10,X11,X12,X20,X21,X22)
 70      CONTINUE

C     Initialiser  A a AORIG
```

```
         DO 100 J = 1,NVAR
            DO 200 I = 1,NOBS
               A(I,J) = AORIG(I, J)
 200        CONTINUE
 100     CONTINUE

C*************************************************************************
*     Choisir le modele, determiner la qualite du modele                 *  1B *
         CALL REG(A, NVAR, NOBS, SELECT, IND,QLT, COEFO, MSECL,          *  1B *
     *           SECOEF,BETA, CTCLR, CTCLR2,ISX,WT,VARMOD,BETAL,         *  1B *
     *           B,SE,COV,RES,H,P,WK,WKBK,Q1,QTMP,QMIN,EXSS,PFW,WKFW,    *  1B *
     *           IMVB,FREE,MODEL,IMVBCP,TAU,WORK,LWORK,R,CP,             *  1B *
     *           IW1,IW2,IW3,W1,W2,W3,W4,BRSS)                           *  1B *
C*************************************************************************

         WRITE(10,99991) (COEFO(I), I=1, NVAR-1)

C     Compter le nombre de modeles classiques petits, ok, grands

         QLTCL(1) = QLTCL(1) + QLT(1)
         QLTCL(2) = QLTCL(2) + QLT(2)
         QLTCL(3) = QLTCL(3) + QLT(3)

C     Intervalles de confiance classiques
         IFAIL = -1
         DF = NOBS - IND
         QT = G01FBF('L',1-ALPHA/2,DF, IFAIL)

         DO 500 I = 1, NVAR-1
            BINFCL(I) = COEFO(I) - SECOEF(I) * QT
            BSUPCL(I) = COEFO(I) + SECOEF(I) * QT
 500     CONTINUE

         DO 510 I = 1, NVAR-1
            WRITE(4,99996) BINFCL(I), BSUPCL(I)
 510     CONTINUE

C
C     Intervalles de confiance bootstrap
C
*************************************************************************
*1111111111111111111111111111111111111111111111111111111111111111111111*
*1    Ajouter les lignes pour l'estimation du modele et des coefficients  *
*1    pre-bootstrap.                                                      *
*1                                                                        *
*1*     Initialiser A a AORIG                                            *
```

```
*1                                                                              *
*1          DO 520 J = 1, NVAR                                                  *
*1             DO 530 I = 1, NOBS                                               *
*1                A(I,J) = AORIG(I,J)                                           *
*1 530         CONTINUE                                                         *
*1 520      CONTINUE                                                            *
*1                                                                              *
*1*   Choisir les residus du modele avec la methode de selection pre-bootstrap  *
*1*   (SBOOT)                                                                   *
*1                                                                              *
*1          CALL REG(A, NVAR, NOBS, SBOOT, IND,QLT, COEFPB, MSE,                *
*1     *            SECOEF,BETA, CTCLR, CTCLR2,ISX,WT,VARMOD,BETAL,           *
*1     *            B,SE,COV,RESPB,H,P,WK,WKBK,Q1,QTMP,QMIN,EXSS,PFW,WKFW,    *
*1     *            IMVB,FREE,MODEL,IMVBCP,TAU,WORK,LWORK,R,CP,               *
*1     *            IW1,IW2,IW3,W1,W2,W3,W4,BRSS)                             *
*1                                                                              *
*1*           Compter la presence des variables                                *
*1                                                                              *
*1            PBIND(IND) = PBIND(IND) + 1                                       *
*1                                                                              *
*1            DO 560 I = 1, NVAR-1                                              *
*1               PBBETA(I) = PBBETA(I) + VARMOD(I)                              *
*1 560        CONTINUE                                                          *
*1                                                                              *
*1C     Compter le nombre de modeles pre-bootstrap petits, ok, grands           *
*1                                                                              *
*1         QLTPB(1) = QLTPB(1) + QLT(1)                                         *
*1         QLTPB(2) = QLTPB(2) + QLT(2)                                         *
*1         QLTPB(3) = QLTPB(3) + QLT(3)                                         *
*1                                                                              *
*1111111111111111111111111111111111111111111111111111111111111111111111111111*
*********************************************************************************

        DO 550 K = 1, NBOOT
C*********************************************************************************
*           Choisir un echantillon                                        *  2A *
            DO 600 I = 1, NOBS                                            *  2A *
               IBOOT(I) = INT(RANDOM(X10,X11,X12,X20,X21,X22)*NOBS + 1)   *  2A *
 600        CONTINUE                                                      *  2A *
C*********************************************************************************

C*********************************************************************************
*          Initialiser ABOOT                                                * 1C-2B *
             DO 700 J = 1,NVAR                                             * 1C-2B *
                DO 800 I = 1,NOBS                                          * 1C-2B *
                   ABOOT(I,J) = AORIG(IBOOT(I), J)                         * 1C-2B *
```

```
 800          CONTINUE                                                  * 1C-2B *
 700        CONTINUE                                                    * 1C-2B *
C*************************************************************************
*************************************************************************
*1111111111111111111111111111111111111111111111111111111111111111111111*
*1    Ajouter le calcul du modele boostrap pour le programme de          *
*1    reechantillonnage des residus                                      *
*1                                                                       *
*1*          Calculer MUSTAR                                            *
*1                                                                       *
*1           DO 850 I = 1, NOBS                                         *
*1              YTMP = 0.0                                              *
*1              DO 860 J = 1, NVAR-1                                    *
*1                 YTMP = YTMP + COEFPB(J) * AORIG(I,J)                 *
*1 860          CONTINUE                                                *
*1              MUSTAR(I) = YTMP                                        *
*1 850        CONTINUE                                                  *
*1                                                                       *
*1*          Calculer Y* avec les residus du modele pre-bootstrap       *
*1                                                                       *
*1           DO 900 I = 1, NOBS                                         *
*1              ABOOT(I,NVAR) = MUSTAR(I) + RESPB(IBOOT(I))             *
*1 900        CONTINUE                                                  *
*1                                                                       *
*1111111111111111111111111111111111111111111111111111111111111111111111*
*************************************************************************

*
C*************************************************************************
        Selection du modele, determiner sa qualite                      * 1D-2C *
            CALL REG(ABOOT, NVAR, NOBS, SELECT, IND,                    * 1D-2C *
     *            QLT, COEF, MSE,SECOEF, BETA, CTRANK, CTRK2,           * 1D-2C *
     *            ISX,WT,VARMOD,BETAL,B,SE,COV,RES,H,P,                 * 1D-2C *
     *            WK,WKBK,Q1,QTMP,QMIN,EXSS,PFW,WKFW,IMVB,FREE,         * 1D-2C *
     *            MODEL,IMVBCP,TAU,WORK,LWORK,R,CP,IW1,                 * 1D-2C *
     *            IW2,IW3,W1,W2,W3,W4,BRSS)                             * 1D-2C *
C*************************************************************************

*          Compter la presence des variables

           CTIND(IND) = CTIND(IND) + 1

           DO 950 I = 1, NVAR-1
              CTBETA(I) = CTBETA(I) + VARMOD(I)
 950       CONTINUE
```

```
C*************************************************************************
*           Calculer Jn                                                 * 1E-2D *
*                                                                       * 1E-2D *
            DO 1000 I = 1, NVAR-1                                       * 1E-2D *
               DCOEF(K,I) = COEF(I) - COEF0(I)                          * 1E-2D *
 1000       CONTINUE                                                    * 1E-2D *
C*************************************************************************

            IF (KK.EQ.2) THEN
               WRITE(8,99993) (COEF(I),I=1,NVAR-1)
            END IF

*           Calculer Ln
*
            DO 1100 I = 1, NVAR-1
               DCOEFM(K,I) = DCOEF(K,I) / SQRT(MSE)
 1100       CONTINUE

            IF (KK.EQ.2) THEN
               WRITE(9,99992) (DCOEFM(K,I),I=1, NVAR-1), SQRT(MSE)
            END IF

C     Compter le nombre de modeles petits, corrects, grands
C
            QLTBT(1) = QLTBT(1) + QLT(1)
            QLTBT(2) = QLTBT(2) + QLT(2)
            QLTBT(3) = QLTBT(3) + QLT(3)

 550     CONTINUE

C
C     Calculer les quantiles alpha/2 et 1-alpha/2 de dcoef (Jn)
C
         DO 1300 I = 1, NVAR-1
            DO 1350 K = 1, NBOOT
               TEMP(K) = DCOEF(K,I)
 1350       CONTINUE

            QBRL(I,1)=QUANT(ALPHA/2, TEMP, NBOOT, WORK5,WORK6,
     *                 WORK7,WORK8,V10,V11,V12,V20,V21,V22)
            QBRU(I,1)=QUANT(1-ALPHA/2, TEMP, NBOOT,WORK5,WORK6,
     *                 WORK7,WORK8,V10,V11,V12,V20,V21,V22)

 1300    CONTINUE
C
```

```
C     Calculer les quantiles alpha/2 et 1-alpha/2 de dcoefm (Ln)
C
         DO 1400 I = 1, NVAR-1
            DO 1450 K = 1, NBOOT
               TEMP(K) = DCOEFM(K,I)
 1450       CONTINUE

            QBRL(I,2)=QUANT(ALPHA/2, TEMP, NBOOT, WORK5,WORK6,
     *                 WORK7,WORK8,V10,V11,V12,V20,V21,V22)
            QBRU(I,2)=QUANT(1-ALPHA/2, TEMP, NBOOT, WORK5,WORK6,
     *                 WORK7,WORK8,V10,V11,V12,V20,V21,V22)

 1400    CONTINUE

C     Calcul des intervalles de confiance bootstrap

C     bootstrap pivotal = colonne 1
C     bootstrap percentile = colonne 2
C     bootstrap t-MSE = colonne 3

C*************************************************************************
         DO 1500 I = 1, NVAR-1                                          *  1F *
            BINFPV(I) = COEFO(I) - QBRU(I,1)                            *  1F *
            BINFPC(I) = COEFO(I) + QBRL(I,1)                            *  1F *
            BINFTM(I) = COEFO(I) - SQRT(MSECL)*QBRU(I,2)                *  1F *
            BSUPPV(I) = COEFO(I) - QBRL(I,1)                            *  1F *
            BSUPPC(I) = COEFO(I) + QBRU(I,1)                            *  1F *
            BSUPTM(I) = COEFO(I) - SQRT(MSECL)*QBRL(I,2)                *  1F *
 1500    CONTINUE                                                       *  1F *
C*************************************************************************

         DO 1520 I = 1, NVAR-1
            WRITE(5,99996) BINFPV(I), BSUPPV(I)
            WRITE(6,99996) BINFPC(I), BSUPPC(I)
            WRITE(7,99996) BINFTM(I), BSUPTM(I)
 1520    CONTINUE

C     Evaluer la qualite de couverture (1-classique, 2-pivotal,
C     3-percentile, 4-t-mse)

         CALL COUVRT(BINFCL,BSUPCL,BETA, NVAR-1,CVT,pmax)

         DO 1600 I = 1, NVAR-1
            DO 1650 J = 1,3
               CPTCVT(1,I,J) = CPTCVT(1,I,J) + CVT(I,J)
 1650       CONTINUE
```

```
1600    CONTINUE


        CALL COUVRT(BINFPV,BSUPPV,BETA, NVAR-1,CVT,pmax)
        DO 1700 I = 1,NVAR-1
           DO 1750 J = 1,3
              CPTCVT(2,I,J) = CPTCVT(2,I,J) + CVT(I,J)
1750       CONTINUE
1700    CONTINUE

        CALL COUVRT(BINFPC,BSUPPC,BETA, NVAR-1,CVT,pmax)
        DO 1800 I = 1,NVAR-1
           DO 1850 J=1,3
              CPTCVT(3,I,J) = CPTCVT(3,I,J) + CVT(I,J)
1850       CONTINUE
1800    CONTINUE

        CALL COUVRT(BINFTM,BSUPTM,BETA, NVAR-1,CVT,pmax)
        DO 1900 I = 1, NVAR-1
           DO 1950 J = 1, 3
              CPTCVT(4,I,J) = CPTCVT(4,I,J) + CVT(I,J)
1950       CONTINUE
1900    CONTINUE


C     Calculer la longueur des intervalles de confiance (1-classique,
C     2-pivotal, 3-percentile, 4-t-mse)

        DO 2600 I = 1, NVAR-1
           LONGIC(I,KK,1) = BSUPCL(I) - BINFCL(I)
           LONGIC(I,KK,2) = BSUPPV(I) - BINFPV(I)
           LONGIC(I,KK,3) = BSUPPC(I) - BINFPC(I)
           LONGIC(I,KK,4) = BSUPTM(I) - BINFTM(I)
2600    CONTINUE

60    CONTINUE

C     Calculer la moyenne et les ecart-types des longueurs des
C     intervalles de confiance.

      DO 2700 K = 1,4
         DO 2710 I = 1, NVAR - 1
            DO 2720 J = 1, NREP
               SUMIDC(I,K) = SUMIDC(I,K) + LONGIC(I,J,K)
               SSQIDC(I,K) = SSQIDC(I,K) + LONGIC(I,J,K)**2
2720        CONTINUE
```

```
2710    CONTINUE
2700 CONTINUE

     DO 2730 K = 1, 4
        DO 2740 I = 1, NVAR-1
           MOYIDC(I,K)= SUMIDC(I,K)/NREP
           SEIDC(I,K)=SQRT((SSQIDC(I,K)-NREP*MOYIDC(I,K)**2)/(NREP-1))
2740    CONTINUE
2730 CONTINUE

*     Imprimer les resultats

C*************************************************************************
                                                                      * 1G-2E *
      WRITE (3,*) '% QUALITE DU MODELE CLASSIQUE'                     * 1G-2E *
      WRITE (3,*)                                                     * 1G-2E *
      WRITE (3,99995) '% TROP PETITS', REAL(QLTCL(1)) / NREP *100, '%'  * 1G-2E *
      WRITE (3,99995) '% CORRECTS', REAL(QLTCL(2)) / NREP * 100, '%'    * 1G-2E *
      WRITE (3,99995) '% TROP GRANDS', REAL(QLTCL(3)) / NREP * 100, '%' * 1G-2E *
      WRITE (3,*) '% PAS DE PLEIN RANG AVANT ', CTCLR                 * 1G-2E *
      WRITE (3,*) '% PAS DE PLEIN RANG APRES', CTCLR2                 * 1G-2E *
      WRITE (3,*)                                                     * 1G-2E *
      WRITE (3,*) '% PRESENCE DES VARIABLES BOOTSTRAP'                * 1G-2E *
      WRITE (3,*) '%', (CTBETA(I), I=1, NVAR-1)                       * 1G-2E *
      WRITE (3,*) '% TAILLE DES MODELES BOOTSTRAP'                    * 1G-2E *
      WRITE (3,*) '%', (CTIND(I), I = 1, NVAR-1)                      * 1G-2E *
      WRITE (3,*) ' '                                                 * 1G-2E *
      WRITE (3,*) '% QUALITE DU MODELE BOOTSTRAP'                     * 1G-2E *
      WRITE (3,99995)'%TROP PETITS', REAL(QLTBT(1))/(NREP*NBOOT)*100,'%'* 1G-2E *
      WRITE (3,99995)'%CORRECTS', REAL(QLTBT(2))/(NREP*NBOOT) * 100,'%' * 1G-2E *
      WRITE (3,99995)'%TROP GRANDS', REAL(QLTBT(3))/(NREP*NBOOT)*100,'%'* 1G-2E *
      WRITE (3,*) '% PAS DE PLEIN RANG AVANT ', CTRANK                * 1G-2E *
      WRITE (3,*) '% PAS DE PLEIN RANG APRES', CTRK2                  * 1G-2E *
      WRITE (3,*) ' '                                                 * 1G-2E *
      WRITE (3,*) '% QUALITE DE COUVERTURE'                           * 1G-2E *
      WRITE (3,*) '% IDC EN FORMAT LATEX'                             * 1G-2E *
      WRITE (3,*)                                                     * 1G-2E *
      WRITE (3,*) '\\begin{center}'                                   * 1G-2E *
      WRITE (3,*) '\\begin{tabular}{|l|l|rrrrrrrr|} \\hline'          * 1G-2E *
      WRITE (3,99989) '\\multicolumn{10}{|c|}{Matrice 1,$\\sigma =$',  * 1G-2E *
     *                STDDEV,                                        * 1G-2E *
     *               '} \\\\ \\hline \\hline'                        * 1G-2E *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Coefficients} & $\\beta_0$ &', * 1G-2E *
     *            '$\\beta_1$&$\\beta_2$&$\\beta_3$&$\\beta_4$&$',   * 1G-2E *
     *            '\\beta_5$&$\\beta_6$ & $\\beta_7$ \\\\'           * 1G-2E *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Valeurs exactes}',           * 1G-2E *
```

```
*           '& 1,0&2,0&1,5&1,5&0,5&0&0&0\\\\ \\hline \\hline'     * 1G-2E *
     WRITE (3,*) '\\multicolumn{10}{|c|}{',SELECT,'}\\\\ \\hline'      * 1G-2E *
                                                                      * 1G-2E *
     DO 4990 K = 1,4                                                  * 1G-2E *
        DO 5000 J = 1, 3                                              * 1G-2E *
           IF (J.EQ.1.AND.K.EQ.1) THEN                                * 1G-2E *
              WRITE(3,99994) 'classique & Uni. gauche 2,5\\%',        * 1G-2E *
*               ('&', REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\' * 1G-2E *
           ELSE IF (J.EQ.1.AND.K.EQ.2) THEN                           * 1G-2E *
              WRITE(3,99994) 'pivotal & Uni. gauche 2,5\\%',          * 1G-2E *
*               ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'  * 1G-2E *
           ELSE IF (J.EQ.1.AND.K.EQ.3) THEN                           * 1G-2E *
              WRITE(3,99994) 'percentile & Uni. gauche 2,5\\%',       * 1G-2E *
*               ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'  * 1G-2E *
           ELSE IF (J.EQ.1.AND.K.EQ.4) THEN                           * 1G-2E *
              WRITE(3,99994) 'bootstrap-t & Uni. gauche 2,5\\%',      * 1G-2E *
*               ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'  * 1G-2E *
           ELSE IF (J.EQ.2) THEN                                      * 1G-2E *
              WRITE(3,99994) '& Uni. droite 2,5\\%',                  * 1G-2E *
*               ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'  * 1G-2E *
           ELSE                                                       * 1G-2E *
              WRITE(3,99994) '& Bilateral 95\\%',                     * 1G-2E *
*               ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),        * 1G-2E *
*               '\\\\ \\hline'                                        * 1G-2E *
                                                                      * 1G-2E *
           END IF                                                     * 1G-2E *
                                                                      * 1G-2E *
5000    CONTINUE                                                      * 1G-2E *
4990 CONTINUE                                                         * 1G-2E *
     WRITE (3,*) '\\end{tabular}'                                     * 1G-2E *
     WRITE (3,*) '\\end{center}'                                      * 1G-2E *
                                                                      * 1G-2E *
     WRITE (3,*)                                                      * 1G-2E *
     WRITE (3,*) '% LONGUEUR DES INTERVALLES DE CONFIANCE CLASSIQUE'  * 1G-2E *
     WRITE (3,*) '% EN FORMAT LATEX'                                  * 1G-2E *
                                                                      * 1G-2E *
     WRITE (3,*) '\\begin{center}'                                    * 1G-2E *
     WRITE (3,*) '\\begin{tabular}{|l|l|rrrrrrrr|} \\hline'           * 1G-2E *
     WRITE (3,99989) '\\multicolumn{10}{|c|}{Matrice 1,$\\sigma =$',  * 1G-2E *
*                    STDDEV,                                          * 1G-2E *
*                   '} \\\\ \\hline \\hline'                          * 1G-2E *
     WRITE (3,*) '\\multicolumn{2}{|c|}{Coefficients} & $\\beta_0$ &', * 1G-2E *
*               '$\\beta_1$&$\\beta_2$&$\\beta_3$&$\\beta_4$&$',     * 1G-2E *
*               '\\beta_5$&$\\beta_6$ & $\\beta_7$ \\\\'              * 1G-2E *
     WRITE (3,*) '\\multicolumn{2}{|c|}{Valeurs exactes}',            * 1G-2E *
*               '& 1,0&2,0&1,5&1,5&0,5&0&0&0\\\\ \\hline \\hline'     * 1G-2E *
```

```
      WRITE (3,*) '\\multicolumn{10}{|c|}{',SELECT,'}\\\\ \\hline'      * 1G-2E *
                                                                        * 1G-2E *
                                                                        * 1G-2E *
      DO 5040 K=1,4                                                     * 1G-2E *
         IF (K.EQ.1) THEN                                               * 1G-2E *
            WRITE (3,99990) 'classique & moyenne ',                     * 1G-2E *
     *           ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'                * 1G-2E *
            WRITE (3,99990) '          & ecart-type ',                   * 1G-2E *
     *           ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'          * 1G-2E *
         ELSE IF (K.EQ.2) THEN                                          * 1G-2E *
            WRITE (3,99990) 'pivotal & moyenne ',                       * 1G-2E *
     *           ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'                * 1G-2E *
            WRITE (3,99990) '          & ecart-type ',                   * 1G-2E *
     *           ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'          * 1G-2E *
         ELSE IF (K.EQ.3) THEN                                          * 1G-2E *
            WRITE (3,99990) 'percentile & moyenne ',                    * 1G-2E *
     *           ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'                * 1G-2E *
            WRITE (3,99990) '          & ecart-type ',                   * 1G-2E *
     *           ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'          * 1G-2E *
         ELSE                                                           * 1G-2E *
            WRITE (3,99990) 'bootstrap-t & moyenne ',                   * 1G-2E *
     *           ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'                * 1G-2E *
            WRITE (3,99990) '          & ecart-type ',                   * 1G-2E *
     *           ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'          * 1G-2E *
         END IF                                                         * 1G-2E *
                                                                        * 1G-2E *
 5040 CONTINUE                                                          * 1G-2E *
      WRITE (3,*) '\\end{tabular}'                                      * 1G-2E *
      WRITE (3,*) '\\end{center}'                                       * 1G-2E *
      WRITE (3,*) '\\end{document}'                                     * 1G-2E *
99989 FORMAT (A, F5.2,A)                                                * 1G-2E *
*************************************************************************
99990 FORMAT (A, 8(A, F6.4),A)
99991 FORMAT (8(F9.5))
99992 format (9(f9.4))
99993 FORMAT (8(F9.4))
99994 FORMAT (A, 8(A, F6.2), A)
99995 FORMAT (X,A, F6.2,A)
99996 FORMAT (2F12.8)
99997 FORMAT (X,A,8F5.2)
99998 FORMAT (X,A,F4.1, A)
99999 FORMAT (9F10.5)
      END

C cette fontion genere des normale[0-1]
      double precision function rnor(x10,x11,x12,x20,x21,x22)
```

```
C- returns random variable, ignores argument.
C- This algorithm is due to Marsaglia and Tsang in JSSC(1984)
      implicit double precision (a-h,o-z)
      integer x10,x11,x12,x20,x21,x22
      double precision v(65),random,aa,b,c,c1,c2,pc,xn
      double precision s,x,y
      data aa,b/12.3758602991705507,0.487899177760379418/
      data c/12.6770580788654669/
      data c1,c2,pc,xn/0.968925454496254195,1.30119777969491621,
     +0.19583033395554691425d-1,2.77699426966287549361/
      data (v(i),i=1,56)/0.3409450287039778957,0.457314591866933539 2,
     +0.539779281611666942,0.606242679653048906,0.663169062764524856,
     +0.713697459056025892,0.759612474933920606,0.802035600355531314,
     +0.841722667978955422,0.879210223208313640,0.914894804386750588,
     +0.949079113753090251,0.982000481239888201,1.013849238029941735,
     +1.044781036740173519,1.074925382028553514,1.104391702268125812,
     +1.133273776243940793,1.161653030133931729,1.189601040838737990,
     +1.217181470700871216,1.244451587898246833,1.271463480572119694,
     +1.298265041883197512,1.324900782180861096,1.351412509933371297,
     +1.377839912870011814,1.404221063559975407,1.430592868502691314,
     +1.456991476137671579,1.483452656603219312,1.510012164318519916,
     +1.536706093359520992,1.563571235037691041,1.590645447014253524,
     +1.617968043674446984,1.645580218369081616,1.673525509567038672,
     +1.701850325062740553,1.730604541317783191,1.759842199038301201,
     +1.789622321566574501,1.820009890130691769,1.851077020230275900,
     +1.882904397592872815,1.915583051943032025,1.949216574916360603,
     +1.983923928905685773,2.019843052906235554,2.057135559990096169,
     +2.095992956249391618,2.136645022544389866,2.179371340398135660,
     +2.224517507216017842,2.272518554850147800,2.323933820094302571/
      data (v(i),i=57,65)/
     +2.379500774082828298,2.440221797979943404,2.507511701865317013,
     +2.583465835225429570,2.671391590320836021,2.776994269662875493,
     +2.776994269662875493,2.776994269662875493,2.776994269662875493/
C- Fast Part
      j=int(64.0d0*random(x10,x11,x12,x20,x21,x22))+1
      rnor=(2.0d0*random(x10,x11,x12,x20,x21,x22)-1.0d0)*v(j+1)
      if(dabs(rnor) .le. v(j)) return
C-------------Slow part; aa is a*f(0)
      x=(dabs(rnor)-v(j))/(v(j+1)-v(j))
      y=random(x10,x11,x12,x20,x21,x22)
      s=x+y
      if(s .gt. c2) go to 11
      if(s .le. c1) return
      if(y .gt. c-aa*dexp(-0.5*(b-b*x)**2)) go to 11
      if(dexp(-0.5*v(j+1)**2)+y*pc/v(j+1) .le. dexp(-0.5*rnor**2))
     +        return
```

```
C-------------Tail part: 0.3601015713011892763 is 1.0/xn
22            x=0.36010157130118927630*
     +              dlog(random(x10,x11,x12,x20,x21,x22))
              if(-2.0*dlog(random(x10,x11,x12,x20,x21,x22))
     +          .le. x**2)    go to 22
33            rnor=dsign(xn-x,rnor)
              return
11     rnor=dsign(b-b*x,rnor)
       return
       end
C*************************************************************
C
C  Adaptation du code en C de l'algorithme de L'Ecuyer (1995):
C  "Combined Multiple Recursive Random Number Generators".
C
C  Ceci genere des Uniforme[0-1].  Nous avons besoin d'une racine
C  de 6 entiers entre 1 et 2145483478, inclusivement.
C
C*************************************************************
       double precision function random (x10,x11,x12,x20,x21,x22)
       implicit double precision (a-h,o-z)
       integer m1,m2,a12,q12,r12,a13,q13,r13,a21,q21,r21
       integer a23,q23,r23,x10,x11,x12,x20,x21,x22,irandom
       double precision Invmp1
       integer h,p12,p13,p21,p23,irandom
       data m1/2147483647/
       data m2/2145483479/
       data a12/63308/
       data q12/33921/
       data r12/12979/
       data a13/-183326/
       data q13/11714/
       data r13/2883/
       data a21/86098/
       data q21/24919/
       data r21/7417/
       data a23/-539608/
       data q23/3976/
       data r23/2071/
       data Invmp1/4.656612873077393D-10/
C*****************************
C Composante 1
C*****************************
C       write(6,*) "x10=",x10," x11=",x11," x12=",x12
C       write(6,*) "x20=",x20," x21=",x21," x22=",x22
       h = x10 / q13
```

```
      p13 = -a13 * (x10 - h * q13) - h * r13
      h = x11 / q12
      p12 = a12 * (x11 - h * q12) - h * r12
      if(p13 .lt. 0) p13 = p13 + m1
      if(p12 .lt. 0) p12 = p12 + m1
      x10 = x11
      x11 = x12
      x12 = p12 - p13
      if(x12 .lt. 0) x12 = x12 + m1
C*****************************
C Composante 2
C*****************************
      h = x20 / q23
      p23 = -a23 * (x20 - h * q23) - h * r23
      h = x22 / q21
      p21 = a21 * (x22 - h * q21) - h * r21
      if(p23 .lt. 0) p23 = p23 + m2
      if(p21 .lt. 0) p21 = p21 + m2
      x20 = x21
      x21 = x22
      x22 = p21 - p23
      if(x22 .lt. 0) x22 = x22 + m2
C*****************************
C Combinaison
C*****************************
      irandom = x12 - x22
      if(irandom .lt. 0) irandom = irandom + m1
      if(irandom .eq. 0) irandom = m1
      random = dble(irandom) * Invmp1
      return
      end

*************************************************************************
*                                                                       *
*     Cette sous-routine selectionne un modele selon une des 5 methodes *
*     de selection (Cp de Mallow, de Ducharme, BIC, "forward" ou "backward"),*
*     evalue la qualite du modele (trop petit, correct, trop grand),    *
*     evalue si la matrice X est de plein rang avant la selection et apres *
*     celle-ci et calcule les coefficients de la regression du modele   *
*     selectionne ainsi que l'ecart-type de ces coefficients,           *
*                                                                       *
*************************************************************************

      SUBROUTINE REG(A, NVAR, NOBS, SELECT, IND,QLT,
     *               COEF, MSE, SECOEF, BETA, CTRANK,CTRK2,ISX,
     *               WT,VARMOD,BETAL,B,SE,COV,RES,H,P,WK,WKBK,
```

```
     *              Q1,QTMP,QMIN,EXSS,PFW,WKFW,IMVB,FREE,MODEL,
     *              IMVBCP,TAU,WORK,LWORK,R,CP,IW1,IW2,IW3,
     *              W1,W2,W3,W4,BRSS)

      INTEGER          PMAX, NMAX, NMAX, LWORK
      PARAMETER       (PMAX=16, NMAX=1000)

      CHARACTER        MEAN*1, WEIGHT*1, SELECT*3
************************************************************************
*                                                                      *
*     MEAN:   Indique si la regression est calcule avec ('M') ou sans ('Z') *
*             constante.                                               *
*     WEIGHT: Indique si la regression est calcule avec ('W') ou sans ('U') *
*             poids.                                                   *
*     SELECT: Methode de selection de modeles:                         *
*             CPM: Cp de Mallow, DUC: Cp de Ducharme, BIC: BIC,        *
*             FWD: Forward, BWD: Backward                              *
*                                                                      *
************************************************************************

      INTEGER          NVAR, NOBS, NVARI, IND, I, J, IDF, IRANK,
     *                 IFAIL, CTRANK, CTRK2
************************************************************************
*                                                                      *
*     NVAR:    Nombre de variables dans le modele incluant l'ordonnee et *
*              la variable expliquee.                                  *
*     NOBS:    Nombre d'observations.                                  *
*     NVARI:   Nombre de variables dans le modele excluant la variable *
*              explicative. (NVAR-1)                                   *
*     IND:     Nombre de variables contenues dans le modele selectionne *
*              (incluant l'ordonnee).                                  *
*     IDF:     Degres de liberte associe avec RSS.                     *
*     IRANK:   Le rang des variables independantes.                    *
*     CTRANK:  Compteur pour le nombre de matrice X de rang incomplet, *
*              avant la selection de modele.                           *
*     CTRK2:   Meme compteur que precedemment, apres la selection du modele *
*                                                                      *
************************************************************************

      INTEGER          ISX(NVAR-1),VARMOD(NVAR-1),BETAL(NVAR-1),QLT(3)
************************************************************************
*                                                                      *
*     ISX:     Vecteur de 0,1,2 indiquant quelles variables a considerer *
*              initialement dans le modele. 0: Variable non-considere, *
*              1: considere, 2: force a etre comprise dans le modele.  *
```

```
*     VARMOD: Vecteur de 0 et de 1 indiquant les variables composant le     *
*             modele selectionne.                                           *
*     BETAL:  Vecteur de 0 et de 1, 1 si la vraie valeur de beta est        *
*             differente de 0.0                                             *
*     QLT:    Vecteur de 0 et 1 indiquant si le modele est trop petit,      *
*             correct, ou trop grand (un seul 1 par vecteur)                *
*                                                                           *
*****************************************************************************

      DOUBLE PRECISION MSE, RSS, TOL
*****************************************************************************
*                                                                           *
*     MSE:    Erreur quadratique moyenne.                                   *
*     RSS:    Somme des erreurs au carre.                                   *
*     TOL:    Niveau de tolerance pour determiner si la matrice X est de    *
*             plein rang. (Valeur suggere TOL = 0.000001)                   *
*                                                                           *
*****************************************************************************

      DOUBLE PRECISION A(NMAX,NOBS), WT(NOBS), B(NVAR-1), SE(NVAR-1),
     *                 COV(((NVAR-1)*(NVAR-1)+(NVAR-1))/2),
     *                 RES(NOBS), Y(NMAX), X(NMAX,PMAX), COEF(NVAR-1),
     *                 SECOEF(NVAR-1), BETA(NVAR-1)
*****************************************************************************
*                                                                           *
*     A:      Matrice (X,Y)                                                 *
*     WT:     Vecteur de poids                                              *
*     B:      Coefficients de la regression pour les variables choisies     *
*     SE:     Ecart-type de B pour les variables choisies                   *
*     COV:    Matrice de variance-covariance pour les variables choisies    *
*     RES:    Vecteur de residus                                            *
*     COEF:   Coefficients de la regression pour toutes les variables       *
*     SECOEF: Ecart-type de COEF pour toutes les variables                  *
*     BETA:   Vecteur des vraies valeurs de tous les coefficients.          *
*                                                                           *
*****************************************************************************

*     Variables de travail (G02DAF)
      DOUBLE PRECISION H(NOBS), Q(NMAX, PMAX+2),
     *                 P((NVAR-1)*(NVAR+1)), WK(NOBS)

*     Variables specifiques a BKWARD
      DOUBLE PRECISION FOUT
*****************************************************************************
*                                                                           *
*     FOUT:   Valeur critique de la statistique F pour le retrait d'une     *
```

```
*            variable.                                                          *
*                                                                               *
*********************************************************************************

*     Variables de travail
      DOUBLE PRECISION WKBK(5*(NVAR-2)+(NVAR-1)*(NVAR-1)),
     *                 Q1(NOBS,NVAR), QTMP(NOBS,NVAR),
     *                 QMIN(NOBS,NVAR)

*     Variables specifiques a FRWARD
      INTEGER          IMVB(NVAR-1)

      DOUBLE PRECISION FIN

      CHARACTER*1      NAME(PMAX), FREE(NVAR-1), MODEL(NVAR-1)
*********************************************************************************
*                                                                               *
*     IMVB:  Numeros des variables selectionnees.                               *
*     FIN:   Valeur critique de la statistique F pour l'ajout d'une             *
*            variable.                                                          *
*     NAME:  Nom des variables.                                                 *
*     MODEL: Noms des variables selectionnees.                                  *
*                                                                               *
*********************************************************************************

*     Variables de travail
      DOUBLE PRECISION EXSS(NVAR-1), PFW(NVAR-1),WKFW(2*(NVAR-1))

*     Variables specifiques a CHOICP
      INTEGER          IMVBCP(NVAR*(NVAR-1)/2)

      DOUBLE PRECISION R(NVAR*(NVAR+1)/2),
     *                 CP(NVAR-1), BRSS(NVAR-1)
      LOGICAL          SVD
*********************************************************************************
*                                                                               *
*     IMVBCP: Numeros des variables selectionnees.                              *
*     R:      Matrice R de la decomposition QR sous la forme de Stirling        *
*     CP:     Vecteur contenant les valeurs de CP (Mallow, Ducharme ou BIC)     *
*             pour le meilleur modele de taille p.                              *
*     BRSS:   RSS pour le meilleur modele de taille p.                          *
*     SVD:    Retourne "T" si la matrice X n'est pas de rang complet            *
*                                                                               *
*********************************************************************************
```

```
*     Variables de travail
      INTEGER           IW1((NVAR-4)*(NVAR+1)/2),
     *                  IW2(NVAR-1), IW3(NVAR-1,2)
      DOUBLE PRECISION TAU(NOBS), WORK(LWORK),
     *                  W1((NVAR-4)*(3+(NVAR*(NVAR+1)/2)/3)),
     *                  W2((NVAR-4)*(NVAR+1)/2,2),
     *                  W3(NVAR*(NVAR-1)/2),
     *                  W4(NVAR-1,3)

      EXTERNAL          CHOICP, G02DAF, FRWARD, BKWARD
*
*  Execution
*

*  Etablir la dimension des matrices

      NVARI = NVAR - 1

*  Pour la regression sans poids et sans constante (colonne de 1 comprise)

      MEAN = "Z"
      WEIGHT = "U"
*
*     Separer X et Y
*
      DO 11 I = 1, NOBS
         DO 12 J = 1, NVAR-1
            X(I,J) = A(I,J)
 12      CONTINUE
         Y(I) = A(I,NVAR)
 11   CONTINUE

C     Fixer la tolerance
C
      TOL = 0.00001D0
      IFAIL = 0

C     Initialiser ISX

      DO 20 I = 1, NVARI
         ISX(I) = 1
 20   CONTINUE

*************************************************************************
*     Il n'est pas necessaire de calculer le rang de la matrice pour le *
*     programme de reechantillonnage des residus.                       *
```

```
*******************************************************************************

*     Effectuer la regression dans le but de determiner le rang de la
*     matrice.

      CALL G02DAF(MEAN,WEIGHT,NOBS,X,NMAX,NVARI,ISX,NVARI,Y,
     +            WT,RSS,IDF,B,SE,COV,RES,H,Q,NMAX,SVD,IRANK,
     +            P,TOL,WK,IFAIL)

*     Compter le nombre de matrice X de rang incomplet
      IF (SVD) THEN
         CTRANK = CTRANK + 1
      END IF

C
C     Choisir le modele selon la methode de selection
C     "backward", "forward", "cp": "mallow" ou "ducharme"
C
      IF (SELECT .EQ. "BWD") THEN
*     Fixer FOUT
         FOUT = 2.0
         CALL BKWARD(MEAN, WEIGHT, NOBS, NVARI, X, Y, WT, FOUT,
     *               VARMOD, IND,B, COV,H,P,RES,ISX,WT,SE,WKBK,
     *               PMAX,NMAX,Q1,QTMP,QMIN)
      END IF

      IF (SELECT .EQ. "FWD") THEN
C        Fixer FIN
         FIN = 2.0
         CALL FRWARD(MEAN, WEIGHT, NOBS, NVARI, X, Y, WT, FIN,
     *               IND, VARMOD,EXSS,PFW,WKFW,WT,ISX,FREE,MODEL,
     *               NAME,IMVB,Q,NMAX,PMAX)
      END IF

      IF (SELECT.EQ."CPM".OR.SELECT.EQ."DUC".OR.SELECT.EQ."BIC") THEN

         CALL CHOICP(NOBS, NVAR, A, NMAX,NVARI,
     *               SELECT, VARMOD, IND,IMVBCP,
     *               TAU,WORK,LWORK,R,CP,IW1,IW2,IW3,W1,W2,W3,W4,BRSS)
      END IF

*******************************************************************************
*1111111111111111111111111111111111111111111111111111111111111111111111111111*
*1    Ajouter la selection du modele complet pour le programme du             *
*1    reechantillonnage des residus.                                          *
*1      IF (SELECT.EQ."NON") THEN                                             *
```

```
*1        DO 29 I = 1, NVARI                                                    *
*1           VARMOD(I) = 1                                                      *
*1 29      CONTINUE                                                            *
*1         IND = NVARI                                                         *
*1      END IF                                                                 *
*11111111111111111111111111111111111111111111111111111111111111111111111111*
*******************************************************************************

C     Evaluer la qualite du modele (comparer les variables choisies
C     a la presence ou non de ces variables dans le vrai modele).

      CALL QUALIT(VARMOD, NVARI, QLT, BETAL)

C     Pour effectuer la regression

      IFAIL = 0

C     Initialiser SE

      DO 30 I=1,IND
         SE(I) = 0.0
 30   CONTINUE

*******************************************************************************
*                                                                             *
*     G02DAF: Sous-routine de la librairie NAG_MARK16 permettant de           *
*             determiner le rang de la matrice, les coefficients (B) et       *
*             l'ecart-type des coefficients (SE) pour le modele VARMOD        *
*             demande ainsi que le vecteur des residus (RES), la somme        *
*             des erreurs au carre (RSS).                                     *
*                                                                             *
*******************************************************************************

      CALL G02DAF(MEAN,WEIGHT,NOBS,X,NMAX,NVARI,VARMOD,IND,Y,
     +            WT,RSS,IDF,B,SE,COV,RES,H,Q,NMAX,SVD,IRANK,
     +            P,TOL,WK,IFAIL)

C     Compter le nombre de matrices qui ne sont pas de rang complet
C     apres la selection de modeles (pas necessaire par reechantillonnage
C     des residus).

      IF (SVD) THEN
         CTRK2 = CTRK2+1
      END IF

C     Calculer MSE
```

```
      MSE = RSS / (NOBS-IND)

C     Initialiser COEF

      DO 130 I = 1, NVARI
         COEF(I) = 0.0
 130  CONTINUE

C     Ajouter la valeur 0.0 pour les coefficients des variables
C     non-selectionnees

      J = 1
      DO 140 I = 1, NVARI
         IF (VARMOD(I) .EQ. 1) THEN
            COEF(I) = B(J)
            J = J + 1
         END IF
 140  CONTINUE

C     Initialiser SECOEF

      DO 150 I = 1, NVARI
         SECOEF(I) = 0.0
 150  CONTINUE

C     Ajouter la valeur 0.0 pour l'ecart-type des coefficients des
C     variables non-selectionnees.

      J = 1
      DO 160 I = 1, NVARI
         IF (VARMOD(I) .EQ. 1) THEN
            SECOEF(I) = SE(J)
            J = J + 1
         END IF
 160  CONTINUE

      RETURN
      END

*****************************************************************************
*                                                                           *
*     Cette sous-routine selectionne le modele selon la methode Cp de       *
*     Mallow, de Ducharme ou BIC                                            *
*                                                                           *
*****************************************************************************
```

```
      SUBROUTINE CHOICP(NOBS,NVAR,A,NMAX, NVARI,
     *                  SELECT, VARMOD, IND,IMVB,
     *                  TAU,WORK,LWORK,R,CP,IW1,IW2,IW3,W1,W2,W3,W4,
     *                  BRSS)
C
C     Declaration des variables
C

      CHARACTER SELECT*3
      INTEGER          NR, NVAR, NW1, NW2, NW3, NVARI, NW5, IORD, IFIN,
     *                 I, J, NOBS, IND, ITOT, K, L, INFO
      INTEGER          IMVB(NVAR*(NVAR-1)/2), VARMOD(NVARI)
************************************************************************
*                                                                      *
*     NVAR:   Nombre de variables incluant la constante et la variable *
*             expliquee.                                               *
*     NR, NW1, NW2, NW3, NVARI, NW5:                                   *
*             Taille de differents vecteurs et matrices.               *
*     IORD, IFIN:                                                      *
*             Etablir l'ordre de calcul.                               *
*     NOBS:   Nombre d'observation.                                    *
*     IND:    Nombre de variables contenues dans le modele selectionne *
*             (incluant l'ordonnee).                                   *
*     ITOT:   Position du modele selectionne dans IMVB.                *
*     IMVB:   Numeros des variables selectionnees.                     *
*     VARMOD: Vecteur de 0 et de 1 indiquant les variables composant le*
*             modele selectionne.                                      *
*                                                                      *
************************************************************************


      DOUBLE PRECISION SIGMA, BCP, DN
      DOUBLE PRECISION A(NMAX,NOBS), R(NVAR*(NVAR+1)/2), CP(NVARI)
************************************************************************
*                                                                      *
*     SIGMA:  Estime de sigma-carre pour le modele complet             *
*     BCP:    CP minimum                                               *
*     DN:     Exposant du Cp de Ducharme                               *
*     A:      Matrice (X,Y)                                            *
*     R:      Matrice R de la decomposition QR sous la forme de Stirling *
*     CP:     Vecteur contenant les valeurs de CP (Mallow, Ducharme ou BIC) *
*             pour le meilleur modele de taille p.                     *
*                                                                      *
************************************************************************
```

```
*     Variables de travail
      INTEGER           IW1((NVAR-4)*(NVAR+1)/2),
     *                  IW2(NVAR-1), IW3(NVAR-1,2)
      DOUBLE PRECISION W1((NVAR-4)*(3+(NVAR*(NVAR+1)/2)/3)),
     *                  W2((NVAR-4)*(NVAR+1)/2,2),
     *                  W3(NVAR*(NVAR-1)/2),
     *                  W4(NVAR-1,3),BRSS(NVAR-1), TAU(NOBS), WORK(LWORK)

      EXTERNAL ALLRQR, OUTR, DGEQRF


*  Etablir les dimensions des matrices

      NR  = NVAR*(NVAR+1)/2

      IF (NVAR .GE. 5) THEN
         NW1 = (NVAR-4)*(3+NR/3)
      ELSE
         NW1 = 1
      END IF

      IF (NVAR .GE. 5) THEN
         NW2 = (NVAR-4)*(NVAR+1)/2
      ELSE
         NW2 = 1
      END IF

      NW3 = NVAR*(NVAR-1)/2

      IF (NVAR .GE. 5) THEN
         NW5 = (NVAR-4)
      ELSE
         NW5 = 1
      END IF


*  Etablir l'ordre du calcul des modeles (subroutine ALLRQR)

      IORD = 2
      IFIN = 2

*************************************************************************
*                                                                       *
*        DGEQRF: Sous-routine de la librairie NAG_MARK16 pour effectuer *
*                la factorisation QR de A.                              *
*                                                                       *
```

```
*************************************************************************

        CALL DGEQRF(NOBS,NVAR,A,NMAX,TAU,WORK,LWORK,INFO)

*       Store R sous la forme de Stirling

      DO 2 I = 1, NVAR
         DO 3 J = 1, I
            IF (I .EQ. J) THEN
               R(I*(I+1)/2) = 1 / (A(I,J)**2)
            ELSE
               R(I*(I-1)/2 + J) = A(J,I) / A(J,J)
            END IF
 3       CONTINUE
 2    CONTINUE


C     Selection de modele

*************************************************************************
*                                                                       *
*     ALLRQR: Algorithme AS 268.1 de APPL.STATIST. (1991), Vol.40, No.3 *
*             Calcule le meilleur modele pour chaque taille p du modele *
*             (IMVB) ainsi que la somme des erreurs au carre minimale   *
*             pour chaque modele de taille p (BRSS)                     *
*                                                                       *
*************************************************************************

      CALL ALLRQR(R, NR, NVAR, W1, NW1, W2, IW1, NW2, W3, NW3, W4,
     *           IW2, NVARI, IW3, NW5, IORD, IFIN, OUTR, IFAULT,
     *           IMVB, BRSS)

C  Trouver le meilleur modele

C     Calcul du Cp

      IF (SELECT.EQ."CPM".OR.SELECT.EQ."DUC") THEN

         SIGMA = BRSS(NVARI) / (NOBS - NVARI)

C     Calcul du Cp de Mallow

         DO 4 J = 1, NVARI
C           CP(J) = (BRSS(J) / SIGMA) - NOBS + 2*J
            CP(J) = BRSS(J) + 2*J*SIGMA
 4       CONTINUE
```

```
C     Methode de Ducharme

         IF (SELECT .EQ. "DUC") THEN

C     Trouver le min pour le calcul de DN
            IND = 1
            BCP = CP(1)
            DO 5 I = 2, NVARI
               IF (CP(I).LT.BCP) THEN
                  BCP = CP(I)
                  IND = I
               END IF
 5          CONTINUE

            DN = 0.5 + 0.5*((NVARI - IND)/NVARI)

C     Calcul du Cp de Ducharme

            DO 6 J = 1, NVARI
               CP(J) = BRSS(J) + (NOBS**DN)*J*SIGMA
 6          CONTINUE

         END IF

      END IF

C     Methode BIC

      IF (SELECT .EQ. "BIC") THEN
         DO 9 J = 1, NVARI
            CP(J) = NOBS*(LOG(BRSS(J))-LOG(REAL(NOBS-J)))
     *           + J*LOG(REAL(NOBS))
 9       CONTINUE
      END IF

C     Calcul du minimum de Cp (peu importe la methode)

      IND = 1
      BCP = CP(1)
      DO 10 I = 2, NVARI
         IF (CP(I).LT.BCP) THEN
            BCP = CP(I)
            IND = I
         END IF
```

```
 10   CONTINUE

*     Trouver la position du modele selectionne dans IMVB

      ITOT = 0
      DO 15 K = 1, IND-1
         ITOT = ITOT + K
 15   CONTINUE


C     Initialiser VARMOD

      DO 90 I = 1, NVARI
         VARMOD(I) = 0.0
 90   CONTINUE

C     Affecter la valeur 1 si la variable est selectionne ou 0 sinon.

      DO 100 L = ITOT + 1, ITOT + IND
         VARMOD(IMVB(L)) = 1
 100  CONTINUE


      RETURN
      END

      SUBROUTINE ALLRQR(R, NR, N, W1, NW1, W2, IW1, NW2, W3, NW3, W4,
     *                  IW2, NW4, IW3, NW5, IORD, IFIN, OUTR, IFAULT,
     *                  IMVB, BRSS)
C
C        ALGORITHM AS 268.1  APPL.STATIST. (1991), VOL.40, NO.3
C
C        Calculates statistics for all possible subset regressions using
C        the R* matrix from a QR decomposition of (X|Y)
C
C        Auxiliary routine required: AS 164 must be called first to form
C        the orthogonal reduction.
C

      INTEGER NR, N, NW1, NW2, IW1(NW2), NW3, IW2(NW4), NW4,
     *        IW3(NW5, 2), NW5, IORD, IFIN, IFAULT
      DOUBLE PRECISION R(NR), W1(NW1), W2(NW2, 2), W3(NW3),
     +                 W4(NW4, 3)
      INTEGER I, J, J1, J2, J3, K, L, N1, N2, N3, N4, NB, NC,
     +        NF, NI,
     *        NL, NS1, NS2, NS3, IMVB(NW3)
```

```
      DOUBLE PRECISION TSS, GC, GS, E1, E2, E3, E4, E5, ONE,
     *                 BRSS(NW4)
      EXTERNAL OUTR
      DATA ONE / 1.E+0 /
C
      IFAULT = 0
C
C        Checks on input arguments
C
      I = N * (N + 1) / 2
      IF (I .NE. NR) THEN
         IFAULT = 1
         RETURN
      END IF
      N1 = N - 1
      IF (N1 .LT. 2) THEN
         IFAULT = 2
         RETURN
      END IF
      N2 = I - N
      IF (N2 .GT. NW3) THEN
         IFAULT = 3
         RETURN
      END IF
      IF (N1 .GT. NW4) THEN
         IFAULT = 4
         RETURN
      END IF
      IF (N .LT. 5) THEN
         J1 = 1
         J2 = 1
         J3 = 1
      ELSE
         J1 = N - 4
         J2 = J1 * (3 + NR / 3)
         J3 = I - 2 * (N + 1)
      END IF
      IF (J1 .GT. NW5) THEN
         IFAULT = 5
         RETURN
      END IF
      IF (J2 .GT. NW1) THEN
         IFAULT = 6
         RETURN
      END IF
      IF (J3 .GT. NW2) THEN
```

```
            IFAULT = 7
            RETURN
         END IF
         IF (IORD .LT. 1 .OR. IORD .GT. 2) THEN
            IFAULT = 8
            RETURN
         END IF
         IF (IFIN .LT. 1 .OR. IFIN .GT. (N1 - 1)) THEN
            IFAULT = 9
            RETURN
         END IF
C
C           Calculate RSS's & TSS from original R* matrix
C
         DO 10 I = 1, N1
            IW2(I) = I
      10 CONTINUE
         TSS = ONE/R(NR)
         CALL OUTR(IW2, TSS, R, N2, R(N2 + 1), N1, IMVB, BRSS,
        *          NW3, NW4)
         J = N2
         DO 20 I = N1, 1, -1
C*************************************************************************
            W4(I, 3) = R(N2 + I) * R(N2 + I) / R(J)
C*************************************************************************
            TSS = TSS + W4(I, 3)
            J = J - I
            IF (I .GE. IFIN .AND. I .GT. 1) THEN
               CALL OUTR(IW2, TSS, R, J, R(N2 + 1), (I - 1), IMVB, BRSS,
     *                   NW3, NW4)
            END IF
      20 CONTINUE
         DO 30 I = 2, N1
            W4(I, 3) = W4(I, 3) + W4(I - 1, 3)
      30 CONTINUE
C
C           Proceed dropping each column in turn
C
         IF (IORD .EQ. 1) THEN
            NB = 1
            NI = 1
            NF = N1 - IFIN
            NL = 0
         ELSE
            NB = N1 - IFIN
            NI = -1
```

```
         NF = 1
      END IF
   40 NC = NB
      IF (IORD .EQ. 2) THEN
         NL = NC
      END IF
C
C        Initialise temporary storage markers
C
      NS1 = 0
      NS2 = 0
      NS3 = 0
C
C        Start with original matrix.
C
      DO 50 I = 1, N2
         W3(I) = R(I)
   50 CONTINUE
      DO 60 I = 1, N1
         W4(I, 1) = R(N2 + I)
         W4(I, 2) = W4(I, 3)
         IW2(I) = I
   60 CONTINUE
C
C        Calculate givens rotation factors, perform rotations & fill
C        in elements of work matrix dropping the appropriate column.
C
      N4 = N - NC
   70 N3 = (N4 - 1) * (N4 - 2) / 2
   80 K = N3
      DO 110 L = 1, NC
         J = N4 + L - 2
         K = K + J
         J1 = N4 - NI * NC - NL - 2
C
C        Shift unchanged elements of R matrix
C
         IF (J1 .GE. 1) THEN
            J2 = K - J
            DO 90 I = 1, J1
               W3(J2 + I) = W3(K + I)
   90       CONTINUE
         END IF
C
C        Calculate first diagonal element
C
```

```
         J2 = K + J
         GC = W3(K)
         GS = W3(J2 + 1)
         IF (L .EQ. 1) THEN
            E1 = W3(J2)
C*****************************************************************************
            E5 = E1 * E1 * GS + GC                                      *  3A  *
C*****************************************************************************
         ELSE
C*****************************************************************************
            E5 = GS + GC                                                *  4A  *
C*****************************************************************************
         END IF
C*****************************************************************************
         W3(K) = GC * GS / E5                                           *      *
         GC = GC / E5                                                   *  5A  *
         GS = GS / E5                                                   *      *
C*****************************************************************************
         IF (L .EQ. 1) THEN
C*****************************************************************************
            GS = E1 * GS                                                *  6A  *
C*****************************************************************************
         END IF
C
C        Calculate second diagonal element & first off diagonal element
C
         IF (L .LT. NC) THEN
            J1 = J2 + J + 1
            E3 = W3(J1 + 1)
            IF (L .EQ. 1) THEN
               E2 = W3(J1)
               E4 = -E2 + E1 * E3
            ELSE
               E2 = W3(J2)
               E4 = -E2 + E3
            END IF
C*****************************************************************************
            W3(J2) = E2 * GS + E3 * GC                                  *  7A  *
            W3(J2 + 1) = E5 / (E4 * E4)                                 *      *
C*****************************************************************************
         END IF
C
C        Calculate  elements of rest of R matrix
C
         IF (L .LT. (NC - 1)) THEN
            J2 = K + 2 * J + 1
```

```
            DO 100 I = 1, NC - L - 1
               J1 = J2 + J + I + 1
               E3 = W3(J1 + 1)
               IF (L .EQ. 1) THEN
                  E2 = W3(J1)
                  W3(J2 + 1) = (-E2 + E1 * E3) / E4
               ELSE
                  E2 = W3(J2)
                  W3(J2 + 1) = (-E2 + E3) / E4
               END IF
C*******************************************************************************
               W3(J2) = E2 * GS + E3 * GC                                  *  8A  *
C*******************************************************************************
               J2 = J1
  100       CONTINUE
         END IF
C
C        Apply rotation to vector of C's
C
         E2 = W4(J, 1)
         E3 = W4(J + 1, 1)
         IF (L .LT. NC) THEN
            IF (L .EQ. 1) THEN
               W4(J + 1, 1) = (-E2 + E1 * E3) / E4
            ELSE
               W4(J + 1, 1) = (-E2 + E3) / E4
            END IF
         END IF
C*******************************************************************************
         W4(J, 1) = E2 * GS + E3 * GC                                   *  9A  *
         W4(J, 2) = W4(J, 1) * W4(J, 1) / W3(K)                         *      *
C*******************************************************************************
         IF (J .GT. 1) W4(J, 2) = W4(J, 2) + W4(J - 1, 2)
  110 CONTINUE
C
C        Remove deleted variable from list of variables in model
C
      DO 120 L = N4 - 1, J
         IW2(L) = IW2(L + 1)
  120 CONTINUE
C
C        Print RSS's etc.
C
C*******************************************************************************
      IF ((TSS - W4(J,2)).LT.BRSS(J)) THEN
         CALL OUTR(IW2, (TSS - W4(J, 2)), W3, K, W4(1, 1), J, IMVB,
```

```
     *            BRSS, NW3, NW4)
      END IF

C*************************************************************************
      IF (NC .GT. 1) THEN
C*************************************************************************
         IF ((TSS - W4(J,2)).GT.BRSS(J-NC+1)) GO TO 170                 *  10B *
C*************************************************************************
         J2 = K
         J1 = J + 1
         DO 130 L = 1, NC - 1
            I = J - L
            J2 = J2 - J1 + L
C*************************************************************************
         IF ((TSS - W4(I,2)).LT.BRSS(I)) THEN
            CALL OUTR(IW2, (TSS - W4(I, 2)), W3, J2, W4(1, 1),
     *                 I, IMVB, BRSS, NW3, NW4)
         END IF

C*************************************************************************
  130    CONTINUE
      END IF
C
C        Storing  required R matrices in temporary storage
C
      IF (NC .LE. 2) GO TO 160
      DO 140 J1 = 1, K
         NS1 = NS1 + 1
         W1(NS1) = W3(J1)
  140 CONTINUE
      DO 150 J1 = 1, J
         NS2 = NS2 + 1
         W2(NS2, 1) = W4(J1, 1)
         W2(NS2, 2) = W4(J1, 2)
         IW1(NS2) = IW2(J1)
  150 CONTINUE
      NS3 = NS3 + 1
      IF (IORD .EQ. 1) THEN
         IW3(NS3, 1) = 2
         N4 = N1 - NS3
         IW3(NS3, 2) = NC
         NC = 2
      ELSE
         IW3(NS3, 1) = NC - 2
         IW3(NS3, 2) = 1
      END IF
```

```
  160 NC = NC - 1
      IF (IORD .EQ. 1 .AND. NC .EQ. 1) THEN
         GO TO 70
      END IF
      IF (IORD .EQ. 2 .AND. NC .NE. 0) THEN
         GO TO 80
      END IF
C
C        Working through R matrices in temporary storage
C
  170 IF (NS3 .EQ. 0) GO TO 200
      J = N1 - NS3
      J1 = NS2
      DO 180 L = J, 1, -1
         W4(L, 1) = W2(J1, 1)
         W4(L, 2) = W2(J1, 2)
         IW2(L) = IW1(J1)
         J1 = J1 - 1
  180 CONTINUE
      K = J * (J + 1) / 2
      J2 = NS1
      DO 190 L = K, 1, -1
         W3(L) = W1(J2)
         J2 = J2 - 1
  190 CONTINUE
      NC = IW3(NS3, 1)
      N4 = N - NS3 - NC
      IF (NC .EQ. IW3(NS3, 2)) THEN
         NS1 = J2
         NS2 = J1
         NS3 = NS3 - 1
         IF (IORD .EQ. 1) THEN
            GO TO 170
         END IF
      ELSE
         IW3(NS3, 1) = IW3(NS3, 1) + NI
      END IF
      GO TO 70
  200 IF (NB .NE. NF) THEN
         NB = NB + NI
         GO TO 40
      END IF
      RETURN
      END

      SUBROUTINE OUTR(IMV, RSS, R, NR, C, NC, IMVB, BRSS,
```

```
     *                NW3, NW4)
C
C        ALGORITHM AS 268.3  APPL.STATIST. (1991), VOL.40, NO.3
C
C        Version B:  Best subsets
C
      INTEGER NR, NC, IMV(NC)
      DOUBLE PRECISION R(NR), C(NC), RSS
      INTEGER IMVB(NW3)
      DOUBLE PRECISION BRSS(NW4)
      INTEGER I, J
C
      J = NC * (NC - 1) / 2
      DO 10 I = 1, NC
         IMVB(J + I) = IMV(I)
   10 CONTINUE
      BRSS(NC) = RSS
      RETURN
      END

*****************************************************************************
*                                                                           *
*     Cette sous-routine selectionne le modele selon la methode "forward"   *
*                                                                           *
*****************************************************************************

      SUBROUTINE FRWARD(MEAN, WEIGHT, N, M, X, Y, WT, FIN,
     *                  NTERM, VARMOD,EXSS,P,WK,WT,ISX,FREE,
     *                  MODEL,NAME,IMVB,Q,NMAX,PMAX)

      INTEGER           NMAX, PMAX, I, IDF, IFAIL, IFR, ISTEP, J, M,
     *                  N, NTERM
*****************************************************************************
*                                                                           *
*     NMAX:   Nombre d'observations maximales                               *
*     PMAX:   Nombre de variables maximales (incluant la constante et la    *
*             variable expliquee                                            *
*     IDF:    Nombre de degres de liberte                                   *
*     IFR:    Nombre de variables inutilisees                               *
*     ISTEP:  Valeur a 0 pour initialiser le processus                      *
*     M:      Nombre de variables independantes  (incluant la constante)    *
*     N:      Nombre d'observations                                         *
*     NTERM:  Nombre de variables dans le modele selectionne                *
*                                                                           *
*****************************************************************************
```

```
      INTEGER          ISX(M), IMVB(M), VARMOD(M)
*****************************************************************************
*                                                                           *
*     ISX:    Vecteur de 0,1,2 indiquant quelles variables a considerer     *
*             initialement dans le modele. 0: Variable non-considere,       *
*             1: considere, 2: force a etre comprise dans le modele.        *
*     VARMOD: Vecteur de 0 et de 1 indiquant les variables composant le     *
*             modele selectionne.                                           *
*     IMVB:   Numeros des variables selectionnees.                          *
*                                                                           *
*****************************************************************************

      LOGICAL          ADDVAR
      CHARACTER        MEAN, WEIGHT
      CHARACTER*1      NEWVAR
      CHARACTER*1      FREE(M), MODEL(M), NAME(M)
*****************************************************************************
*                                                                           *
*     ADDVAR: Vrai si on ajoute une variable a cette etape.                 *
*     NEWVAR: Nom de la vraiable ajoutee su ADDVAR est vrai                 *
*     FREE:   Nom des variables non selectionnees                           *
*     MODEL:  Nom des variables selectionnees                               *
*     NAME:   Nom des variables                                             *
*                                                                           *
*****************************************************************************

      DOUBLE PRECISION CHRSS, F, FIN, RSS
      DOUBLE PRECISION WT(N), X(N,M), Y(N)
*****************************************************************************
*                                                                           *
*     CHRSS: Changement de RSS en ajoutant NEWVAR                           *
*     F:     Statistique F pour l'ajout de NEWVAR                           *
*     FIN:   Valeur critique pour le test F                                 *
*                                                                           *
*****************************************************************************

*     Variables de travail
      DOUBLE PRECISION EXSS(M), P(M+1), Q(N,M+2), WK(2*M)

      EXTERNAL         G02EEF

      IF (M.LE.PMAX .AND. N.LE.NMAX) THEN

C
C     Initialiser ISX
```

```
C
        ISX(1) = 2
        DO 45 J = 2, M
           ISX(J) = 1
 45     CONTINUE
C
C       Initialiser NAME
C
        DO 50 J = 33, 32+M
           NAME(J-32) = CHAR(J)
 50     CONTINUE

C       Initialiser VARMOD

        DO 51 J = 1, M
           VARMOD(J) = 0
 51     CONTINUE

*       Initialiser ISTEP

        ISTEP = 0

*       Selectionner le modele

        DO 60 I = 1, M
           IFAIL = 0
*
           CALL G02EEF(ISTEP,MEAN,WEIGHT,N,M,X,NMAX,NAME,ISX,PMAX,Y,WT,
     +                 FIN,ADDVAR,NEWVAR,CHRSS,F,MODEL,NTERM,RSS,IDF,
     +                 IFR,FREE,EXSS,Q,NMAX,P,WK,IFAIL)
*
           IF (IFAIL.NE.0) GO TO 80
           IF ( .NOT. ADDVAR) THEN
C
C     Change les character en integer
C
              DO 65 J = 1, NTERM
                 IMVB(J) = ICHAR(MODEL(J))-32
 65           CONTINUE

              DO 66 J = 1, NTERM
                 VARMOD(IMVB(J)) = 1
 66           CONTINUE

              GO TO 80
            ELSE
```

```
               IF (IFR.EQ.0) THEN
                  DO 67 J = 1, M
                     VARMOD(J) = 1
67                CONTINUE
                  GO TO 80
               END IF
            END IF
  60     CONTINUE
  80     CONTINUE
      END IF
      RETURN

      END

*************************************************************************
*                                                                       *
*     Cette sous-routine selectionne le meilleur modele selon la methode *
*     "backward"                                                        *
*                                                                       *
*************************************************************************

      SUBROUTINE BKWARD (MEAN, WEIGHT, N, M, X, Y, WT,FOUT,
     *                   VARMOD, IP, B, COV, H,P,RES,ISX,WT,SE,
     *                   WK,PMAX,NMAX,Q1,QTMP,QMIN)

      INTEGER          PMAX, NMAX, I, IDF, IFAIL, INDX, IP, IRANK, J,
     *                 M, N,INDMIN, K, L
*************************************************************************
*                                                                       *
*     PMAX:    Nombre maximal de variables (incluant la constante et la *
*              variable expliquee                                       *
*     NMAX:    Nombre maximal d'observations                            *
*     IDF:     Nombre de degres de liberte                              *
*     IP:      Nombre de variables disponibles                          *
*     INDMIN:  Indice de la variable ayant la plus petite valeur de la  *
*              statistique F                                            *
*                                                                       *
*************************************************************************

      INTEGER          ISX(M), VARMOD(M)
*************************************************************************
*                                                                       *
*     ISX:     Vecteur de 0,1,2 indiquant quelles variables a considerer *
*              initialement dans le modele. 0: Variable non-considere,  *
*              1: considere, 2: force a etre comprise dans le modele.   *
*     VARMOD: Vecteur de 0 et de 1 indiquant les variables composant le *
```

```
*            modele selectionne.                                              *
*                                                                             *
*******************************************************************************

      LOGICAL           SVD
      CHARACTER         MEAN, WEIGHT

      DOUBLE PRECISION RSS, TOL, RSSTMP,RSSMIN,FTEST, FMAX, FMIN
      PARAMETER        (FMAX = 10 000 000.0)
*******************************************************************************
*                                                                             *
*     RSS:    Somme au carre des residus                                      *
*     RSSMIN: RSS de la regression ayant la plus petite valeur de la          *
*             statistique F                                                   *
*     FTEST:  Valeur de la statistique F                                      *
*     FMIN:   Valeur minimale de la statistique F                             *
*                                                                             *
*******************************************************************************

*     Variables de travail
      DOUBLE PRECISION B(M), COV((M*M+M)/2), H(N),
     +                 P(M*(M+2)), Q1(NMAX,PMAX+1), RES(N),
     +                 SE(M), WK(5*(M-1)+M*M), WT(N),
     +                 X(N,M), Y(N), QTMP(NMAX,PMAX+1),
     +                 QMIN(NMAX,PMAX+1)

      EXTERNAL          G02DAF, G02DDF, G02DFF

*     Initialiser ISX

      DO 60 I = 1, M
         ISX(I) = 1
   60 CONTINUE

      IP = M

      IF (MEAN.EQ.'M' .OR. MEAN.EQ.'m') IP = IP + 1

*     Fixer la tolerance
      TOL = 0.00001D0
      IFAIL = 0

*     Effectuer une premiere regression (modele complet)
      CALL G02DAF(MEAN,WEIGHT,N,X,NMAX,M,ISX,IP,Y,WT,RSS,IDF,B,SE,
     +            COV,RES,H,Q1,NMAX,SVD,IRANK,P,TOL,WK,IFAIL)
*
```

```
*     Initialiser VARMOD

      DO 75 I = 1, M
         VARMOD(I) = 1
 75   CONTINUE

 80   IF (IP .LE. 1) THEN
         RETURN
      ELSE

         IFAIL = 0
*
*     Initialiser les matrices temporaires
         RSSTMP = RSS
         DO 81 I =1, N
            DO 82 J = 1, M+1
               QTMP(I,J) = Q1(I,J)
 82         CONTINUE
 81      CONTINUE

         RSSMIN = 0
         INDMIN = 0
         FMIN = FMAX

*     Enlever une variable a la fois

         DO 90 INDX = 2, IP
            CALL G02DFF(IP,Q1,NMAX,INDX,RSS,WK,IFAIL)
            FTEST = (RSS - RSSTMP) / (RSSTMP/(N-IP))

*     Tester la modification

            IF (FTEST .LT. FMIN .AND. FTEST .LT. FOUT) THEN
               FMIN = FTEST
               INDMIN = INDX
               DO 83 I =1, N
                  DO 84 J = 1, M+1
                     QMIN(I,J) = Q1(I,J)
 84               CONTINUE
 83            CONTINUE
               RSSMIN = RSS
            END IF
            RSS = RSSTMP
            DO 85 I =1, N
               DO 86 J = 1, M+1
                  Q1(I,J) = QTMP(I,J)
```

```
 86           CONTINUE
 85         CONTINUE

 90       CONTINUE

*     Choisir la meilleure modification
         IF (INDMIN .NE. 0) THEN
            J = 0
            DO 95 I = 1,M
               IF (VARMOD(I) .EQ. 1) J = J+1
               IF (INDMIN .EQ. J) THEN
                  VARMOD(I) = 0
                  IP = IP - 1
                  RSS = RSSMIN
                  DO 91 K =1, N
                     DO 92 L = 1, M+1
                        Q1(K,L) = QMIN(K,L)
 92                  CONTINUE
 91                  CONTINUE

                     GO TO 80
                  END IF
 95            CONTINUE
            ELSE
               RETURN
            END IF
      END IF

      RETURN
*
      END

*************************************************************************
*                                                                       *
*     Evaluer la qualite du modele en comparant VARMOD et BETAL.        *
*     La sousroutine retourne 1 a QLT(i) si le modele est i=1, trop petit,  *
*     i=2, correct, i=3, trop grand                                     *
*                                                                       *
*************************************************************************

      SUBROUTINE QUALIT (VARMOD, NVARI, QLT, BETAL)

C
C     Evalue la qualite du modele par rapport au modele de reference beta
C
```

```
      INTEGER VARMOD(NVARI), NVARI, QLT(3), I
      INTEGER UNDER, OK, BETAL(NVARI)
*************************************************************************
*                                                                       *
*     VARMOD: Vecteur de 0 et de 1, 1 lorsque la variable est dans le   *
*             modele selectionne                                        *
*     NVARI:  Nombre de variables independantes (excluant la constante) *
*     QLT:    Vecteur de 0 et de 1, 1 dans la position i lorsque le modele *
*             est trop petit (i=1), correct (i=2), trop grand (i=3)     *
*     UNDER:  Nombre de variables presentes dans le vrai modele, mais pas *
*             le modele selectionne                                     *
*     OK:     Nombre de 0 et de 1 identique dans le vrai modele et le   *
*             modele selectionne                                        *
*     BETAL:  Vecteur de 0 et de 1, 1 lorsque la variable est dans le   *
*             vrai modele                                               *
*                                                                       *
*************************************************************************

*     Initialiser les compteurs et QLT
      UNDER = 0
      OK = 0

      QLT(1) = 0
      QLT(2) = 0
      QLT(3) = 0

*     Evaluer la qualite du modele
      DO 10 I = 1, NVARI
         IF (VARMOD(I) .EQ. BETAL(I)) THEN
            OK = OK + 1
         ELSE
            IF (VARMOD(I) .EQ. 0) UNDER = UNDER + 1
         END IF
 10   CONTINUE

      IF (OK .EQ. NVARI)  THEN
         QLT(2) = 1
      ELSEIF (UNDER .NE. 0) THEN
         QLT(1) = 1
      ELSE
         QLT(3) = 1
      END IF
      RETURN
      END
```

```
C
C fonction servant a calculer les quantiles
      double precision function quant(alpha,y,ssize,data,left,low,
     + high,x10,x11,x12,x20,x21,x22)
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      double precision alpha,data(ssize),low(ssize),high(ssize),
     +  y(ssize),middle,large,left(ssize)
      integer ssize,size,lsize,usize,target,index,x10,x11,x12
     + ,x20,x21,x22
      size=ssize
      index=alpha*(ssize+1)
c      write(*,*) 'la valeur de index est'
c      write(*,*) index
      target=0
      do 1 i=1,ssize
         data(i)=y(i)
1     continue
2     call select(middle,data,low,high,size,lsize,usize,ssize,x10,x11,
     + x12,x20,x21,x22)
      if(target+lsize .gt. index) then
          do 6 i=1,lsize
             data(i)=low(i)
6         continue
          size=lsize
        else
          do 5 i=1,lsize
             left(target+i)=low(i)
5         continue
          if(target+lsize .lt. index) then
              do 4 i=1,usize
                 data(i)=high(i)
4             continue
              target=target+lsize+1
              left(target)=middle
              size=usize
            else
              goto 3
           endif
       endif
       goto 2
3     large = left(1)
c      write(*,*) 'large'
c      write(*,*) large
c      write(*,*) 'index'
c      write(*,*) index
```

```
      do 7 i=1,index
         if(left(i) .gt. large) large=left(i)
7     continue
c      write(*,*) 'large'
c      write(*,*) large
c      write(*,*) alpha
c      write(*,*) ssize
      quant = large + ((alpha * (ssize+1)) - index)
     +                * (middle - large)
c      write(*,*) 'quant'
c      write(*,*) quant
      return
      end
C
      subroutine select(middle,data,low,high,
     +       size,lsize,usize,ssize,x10,x11,x12,x20,x21,x22)
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      double precision data(ssize),low(ssize),high(ssize),middle
      integer rindex,size,lsize,usize,ssize,x10,x11,x12,x20,x21,x22
      double precision donnees,random
      lsize=0
      usize=0
      donnees=random(x10,x11,x12,x20,x21,x22)
c      write(*,*) 'la valeur de donnee'
c      write(*,*) donnees
      rindex=int(donnees*size)+1
c      write(*,*) 'la valeur de rindex'
c      write(*,*) rindex
      middle=data(rindex)
c      write(*,*) middle
c      write(*,*) size
      do 1 i=1,rindex-1
         if(data(i) .lt. middle) then
             lsize=lsize+1
             low(lsize)=data(i)
c            write(*,*) 'lsize1'
c            write(*,*) lsize
           else
             usize=usize+1
             high(usize)=data(i)
c            write(*,*) 'usize1'
c            write(*,*) usize
         endif
1     continue
      do 2 i=rindex+1,size
         if(data(i) .lt. middle) then
```

```
            lsize=lsize+1
            low(lsize)=data(i)
c           write(*,*) 'lsize2'
c           write(*,*) lsize
          else
            usize=usize+1
            high(usize)=data(i)
c           write(*,*) 'usize2'
c           write(*,*) usize
        endif
2     continue
c      write(*,*) 'la valeur de lsize et usize'
c      write(*,*) lsize
c      write(*,*) usize
      end

*************************************************************************
*                                                                       *
*     Sous-routine retournant 1 la vraie valeur de beta est entre inf et *
*     sup, 0 sinon sous forme de tableau (Chaque ligne est un coefficient, *
*     colonne 1: plus petit que la borne inferieure, 2: plus grand que la *
*     borne superieure, 3: entre les deux bornes.                       *
*                                                                       *
*************************************************************************

      SUBROUTINE COUVRT(INF, SUP, BETA, N, CVT, PMAX)

      INTEGER N, I, J,PMAX, CVT(PMAX,3)
      DOUBLE PRECISION INF(N), SUP(N), BETA(N)

*************************************************************************
*     N:     Nombre d'observations                                      *
*     PMAX:  Nombre maximal de variables independantes                  *
*     CVT:   Matrice de 0 et de 1, 1 dans la colonne i si la vrai valeur *
*            de BETA est a gauche de l'IDC (i=1), a droite (i=2), entre *
*            les bornes (i=3)                                           *
*     INF:   Borne inferieure de l'IDC                                  *
*     SUP:   Borne superieure de l'IDC                                  *
*     BETA:  Vecteur contenant les vrais valeurs des coefficients du modele *
*                                                                       *
*************************************************************************

*     Initialiser CVT

      DO 10 I = 1, N
         DO 20 J = 1, 3
```

```
            CVT(I,J) = 0
 20       CONTINUE
 10    CONTINUE

*      Calcul de la qualite de couverture
       DO 30 I = 1, N
          IF (BETA(I) .LT. INF(I)) THEN
             CVT(I,1) = 1
          ELSEIF (BETA(I) .GT. SUP(I)) THEN
             CVT(I,2) = 1
          ELSE
             CVT(I,3) = 1
          END IF
 30    CONTINUE

       RETURN
       END
```

```
*************************************************************************
*                                                                       *
*  Modifications a apporter au programme pour excuter le programme du   *
*  reechantillonnage des residus. Les lignes 1X doivent etre modifiees de *
*  la facon suivante.                                                   *
*                                                                       *
*                                                                       *
```

```
C*************************************************************************
      READ (2,*) NVAR, NOBS, SBOOT, SEVAL                          *  1A *
                                                                   *  1A *
      IF (SEVAL.NE.'CPM'.AND.SEVAL.NE.'DUC'.AND.SEVAL.NE.'BIC'.AND.*  1A *
     *    SEVAL.NE.'BWD'.AND.SEVAL.NE.'FWD') THEN                  *  1A *
         WRITE(3,*)'Vous devez choisir parmi les methodes de selection:' *  1A *
         WRITE(3,*)'BWD, FWD, CPM, DUC ou BIC'                     *  1A *
         WRITE(3,*)'pour evaluer le modele (SEVAL)'                *  1A *
         STOP                                                      *  1A *
      END IF                                                       *  1A *
                                                                   *  1A *
      IF (SBOOT.NE.'CPM'.AND.SBOOT.NE.'DUC'.AND.SBOOT.NE.'BIC'.AND.*  1A *
     *    SBOOT.NE.'BWD'.AND.SBOOT.NE.'FWD'.AND.SBOOT.NE.'NON') THEN *  1A *
         WRITE(3,*)'Vous devez choisir parmi les methodes de selection:' *  1A *
         WRITE(3,*)'BWD, FWD, CPM, DUC, BIC ou NON '               *  1A *
         WRITE(3,*)'pour choisir les erreurs a reechantillonner (SBOOT)' *  1A *
         STOP                                                      *  1A *
      END IF                                                       *  1A *
C*************************************************************************
```

```
C*************************************************************************
*     Choisir le modele, determiner la qualite du modele                *  1B *
*     (Meme methode de selection que SEVAL)                             *  1B *
        CALL REG(A, NVAR, NOBS, SEVAL, IND,QLT, COEFO, MSECL,           *  1B *
     *           SECOEF,BETA, CTCLR, CTCLR2,ISX,WT,VARMOD,BETAL,        *  1B *
     *           B,SE,COV,RES,H,P,WK,WKBK,Q1,QTMP,QMIN,EXSS,PFW,WKFW,   *  1B *
     *           IMVB,FREE,MODEL,IMVBCP,TAU,WORK,LWORK,R,CP,            *  1B *
     *           IW1,IW2,IW3,W1,W2,W3,W4,BRSS)                          *  1B *
C*************************************************************************

C*************************************************************************
*          Initialiser ABOOT                                            *  1C *
           DO 700 J = 1,NVAR-1                                          *  1C *
              DO 800 I = 1,NOBS                                         *  1C *
                 ABOOT(I,J) = AORIG(I, J)                               *  1C *
 800          CONTINUE                                                  *  1C *
 700       CONTINUE                                                     *  1C *
C*************************************************************************

C*************************************************************************
*          Selection du modele, determiner sa qualite                   *  1D *
           CALL REG(ABOOT, NVAR, NOBS, SEVAL, IND,                      *  1D *
     *           QLT, COEF, MSE,SECOEF, BETA, CTRANK, CTRK2,            *  1D *
     *           ISX,WT,VARMOD,BETAL,B,SE,COV,RES,H,P,                  *  1D *
     *           WK,WKBK,Q1,QTMP,QMIN,EXSS,PFW,WKFW,IMVB,FREE,          *  1D *
     *           MODEL,IMVBCP,TAU,WORK,LWORK,R,CP,IW1,                  *  1D *
     *           IW2,IW3,W1,W2,W3,W4,BRSS)                              *  1D *
C*************************************************************************

C*************************************************************************
*          Calculer Jn                                                  *  1E *
*                                                                       *  1E *
           DO 1000 I = 1, NVAR-1                                        *  1E *
              DCOEF(K,I) = COEF(I) - COEFPB(I)                          *  1E *
 1000      CONTINUE                                                     *  1E *
C*************************************************************************

C*************************************************************************
         DO 1500 I = 1, NVAR-1                                          *  1F *
           BINFPV(I) = COEFO(I) - QBRU(I,1)                             *  1F *
           BINFPC(I) = COEFPB(I) + QBRL(I,1)                            *  1F *
           BINFTM(I) = COEFO(I) - SQRT(MSECL)*QBRU(I,2)                 *  1F *
           BSUPPV(I) = COEFO(I) - QBRL(I,1)                             *  1F *
           BSUPPC(I) = COEFPB(I) + QBRU(I,1)                            *  1F *
           BSUPTM(I) = COEFO(I) - SQRT(MSECL)*QBRL(I,2)                 *  1F *
 1500    CONTINUE                                                       *  1F *
```

```
C*************************************************************************

C*************************************************************************
      WRITE (3,*) '% QUALITE DU MODELE CLASSIQUE'                          *  1G *
      WRITE (3,*)                                                          *  1G *
      WRITE (3,99995) '% TROP PETITS', REAL(QLTCL(1)) / NREP *100, '%'     *  1G *
      WRITE (3,99995) '% CORRECTS', REAL(QLTCL(2)) / NREP * 100, '%'       *  1G *
      WRITE (3,99995) '% TROP GRANDS', REAL(QLTCL(3)) / NREP * 100, '%'    *  1G *
      WRITE (3,*)                                                          *  1G *
      WRITE (3,*) '% QUALITE DU MODELE PRE-BOOTSTRAP'                      *  1G *
      WRITE (3,99995) '% TROP PETITS', REAL(QLTPB(1)) / NREP *100, '%'     *  1G *
      WRITE (3,99995) '% CORRECTS', REAL(QLTPB(2)) / NREP * 100, '%'       *  1G *
      WRITE (3,99995) '% TROP GRANDS', REAL(QLTPB(3)) / NREP * 100, '%'    *  1G *
      WRITE (3,*)                                                          *  1G *
      WRITE (3,*) '% QUALITE DU MODELE BOOTSTRAP'                          *  1G *
      WRITE (3,99995)'% TROP PETITS',REAL(QLTBT(1))/(NREP*NBOOT)*100,'%'   *  1G *
      WRITE (3,99995) '% CORRECTS',REAL(QLTBT(2))/(NREP*NBOOT) * 100,'%'   *  1G *
      WRITE (3,99995)'% TROP GRANDS',REAL(QLTBT(3))/(NREP*NBOOT)*100,'%'   *  1G *
      WRITE (3,*) ' '                                                      *  1G *
      WRITE (3,*) '% PRESENCE DES VARIABLES PRE-BOOTSTRAP'                 *  1G *
      WRITE (3,*) '%', (PBBETA(I), I=1, NVAR-1)                            *  1G *
      WRITE (3,*) '% TAILLE DES MODELES PRE-BOOTSTRAP'                     *  1G *
      WRITE (3,*) '%', (PBIND(I), I= 1, NVAR-1)                            *  1G *
      WRITE (3,*) '% PRESENCE DES VARIABLES BOOTSTRAP'                     *  1G *
      WRITE (3,*) '%', (CTBETA(I), I=1, NVAR-1)                            *  1G *
      WRITE (3,*) '% TAILLE DES MODELES BOOTSTRAP'                         *  1G *
      WRITE (3,*) '%', (CTIND(I), I = 1, NVAR-1)                           *  1G *
      WRITE (3,*) ' '                                                      *  1G *
      WRITE (3,*) '% QUALITE DE COUVERTURE'                                *  1G *
      WRITE (3,*) '% IDC EN FORMAT LATEX'                                  *  1G *
      WRITE (3,*)                                                          *  1G *
      WRITE (3,*) '\\begin{center}'                                        *  1G *
      WRITE (3,*) '\\begin{tabular}{|l|l|rrrrrrrr|} \\hline'               *  1G *
      WRITE (3,99989) '\\multicolumn{10}{|c|}{Matrice 1,$\\sigma =$',      *  1G *
     *                STDDEV,                                             *  1G *
     *               '} \\\\ \\hline \\hline'                             *  1G *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Coefficients} & $\\beta_0$ &',    *  1G *
     *           '$\\beta_1$&$\\beta_2$&$\\beta_3$&$\\beta_4$&$',         *  1G *
     *           '\\beta_5$&$\\beta_6$ & $\\beta_7$ \\\\'                  *  1G *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Valeurs exactes}',                *  1G *
     *           '& 1,0&2,0&1,5&1,5&0,5&0&0&0\\\\ \\hline \\hline'        *  1G *
      WRITE (3,*) '\\multicolumn{10}{|c|}{',SEVAL,'}\\\\ \\hline'          *  1G *
      WRITE (3,*) '\\multicolumn{2}{|c|}{single, S$_b$:',SBOOT,            *  1G *
     *           '} & & & & & && &\\\\ \\hline'                           *  1G *
      DO 4990 K = 1,4                                                      *  1G *
         DO 5000 J = 1, 3                                                  *  1G *
```

```
            IF (J.EQ.1.AND.K.EQ.1) THEN                                    *  1G *
               WRITE(3,99994) 'classique & Uni. gauche 2,5\\%',            *  1G *
     *           ('&', REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'     *  1G *
            ELSE IF (J.EQ.1.AND.K.EQ.2) THEN                               *  1G *
               WRITE(3,99994) 'pivotal & Uni. gauche 2,5\\%',              *  1G *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'      *  1G *
            ELSE IF (J.EQ.1.AND.K.EQ.3) THEN                               *  1G *
               WRITE(3,99994) 'percentile & Uni. gauche 2,5\\%',           *  1G *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'      *  1G *
            ELSE IF (J.EQ.1.AND.K.EQ.4) THEN                               *  1G *
               WRITE(3,99994) 'bootstrap-t & Uni. gauche 2,5\\%',          *  1G *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'      *  1G *
            ELSE IF (J.EQ.2) THEN                                          *  1G *
               WRITE(3,99994) '& Uni. droite 2,5\\%',                      *  1G *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'      *  1G *
            ELSE                                                           *  1G *
               WRITE(3,99994) '& Bilateral 95\\%',                         *  1G *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),           *  1G *
     *           '\\\\ \\hline'                                           *  1G *
                                                                           *  1G *
            END IF                                                         *  1G *
                                                                           *  1G *
 5000    CONTINUE                                                          *  1G *
 4990 CONTINUE                                                             *  1G *
      WRITE (3,*) '\\end{tabular}'                                         *  1G *
      WRITE (3,*) '\\end{center}'                                          *  1G *
                                                                           *  1G *
      WRITE (3,*)                                                          *  1G *
      WRITE (3,*) '% LONGUEUR DES INTERVALLES DE CONFIANCE'                *  1G *
      WRITE (3,*) '% EN FORMAT LATEX'                                      *  1G *
      WRITE (3,*)                                                          *  1G *
      WRITE (3,*) '\\begin{center}'                                        *  1G *
      WRITE (3,*) '\\begin{tabular}{|l|l|rrrrrrrr|} \\hline'               *  1G *
      WRITE (3,99989) '\\multicolumn{10}{|c|}{Matrice 1,$\\sigma =$',      *  1G *
     *                 STDDEV,                                           *  1G *
     *                '} \\\\ \\hline \\hline'                           *  1G *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Coefficients} & $\\beta_0$ &',   *  1G *
     *            '$\\beta_1$&$\\beta_2$&$\\beta_3$&$\\beta_4$&$',       *  1G *
     *            '\\beta_5$&$\\beta_6$ & $\\beta_7$ \\\\'                *  1G *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Valeurs exactes}',                *  1G *
     *            '& 1,0&2,0&1,5&1,5&0,5&0&0&0\\\\ \\hline \\hline'       *  1G *
      WRITE (3,*) '\\multicolumn{10}{|c|}{',SEVAL,'}\\\\ \\hline'          *  1G *
      WRITE (3,*) '\\multicolumn{2}{|c|}{single, S$_b$:',SBOOT,            *  1G *
     *            '} & & & & & && &\\\\ \\hline'                          *  1G *
                                                                           *  1G *
      DO 5040 K=1,4                                                        *  1G *
```

```
      IF (K.EQ.1) THEN                                                  *  1G *
         WRITE (3,99990) 'classique & moyenne ',                        *  1G *
     *        ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'                   *  1G *
         WRITE (3,99990) '          & ecart-type ',                     *  1G *
     *        ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'             *  1G *
      ELSE IF (K.EQ.2) THEN                                             *  1G *
         WRITE (3,99990) 'pivotal & moyenne ',                          *  1G *
     *        ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'                   *  1G *
         WRITE (3,99990) '          & ecart-type ',                     *  1G *
     *        ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'             *  1G *
      ELSE IF (K.EQ.3) THEN                                             *  1G *
         WRITE (3,99990) 'percentile & moyenne ',                       *  1G *
     *        ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'                   *  1G *
         WRITE (3,99990) '          & ecart-type ',                     *  1G *
     *        ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'             *  1G *
      ELSE                                                              *  1G *
         WRITE (3,99990) 'bootstrap-t & moyenne ',                      *  1G *
     *        ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'                   *  1G *
         WRITE (3,99990) '          & ecart-type ',                     *  1G *
     *        ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'             *  1G *
      END IF                                                            *  1G *
                                                                        *  1G *
 5040 CONTINUE                                                          *  1G *
      WRITE (3,*) '\\end{tabular}'                                      *  1G *
      WRITE (3,*) '\\end{center}'                                       *  1G *
      WRITE (3,*) '\\end{document}'                                     *  1G *
                                                                        *  1G *
99989 FORMAT (A, F5.2,A)                                                *  1G *
C*************************************************************************
```

```
*************************************************************************
*  Modifications a apporter au programme pour excuter le programme du    *
*  sous-echantillonnage. Les lignes 2X doivent etre modifiees de         *
*  la facon suivante.                                                    *
*                                                                        *
```

```
C*************************************************************************
*          Choisir un echantillon de taille NSUB sans remise.          *  2A *
           IFAIL = 0                                                    *  2A *
           CALL G05EJF(ECH, NOBS, IBOOT, NSUB, IFAIL)                   *  2A *
*************************************************************************  2A *
*                                                                       *  2A *
*     G05EJF: Sous-routine de la librairie NAG_MARK16 de FORTRAN        *  2A *
*             permettant de choisir un echantillon IBOOT de taille NSUB *  2A *
*             sans remise a partir du vecteur ECH de taille NOBS.       *  2A *
*                                                                       *  2A *
```

```
*************************************************************************** 2A *
C***************************************************************************

C***************************************************************************
*          Initialiser ABOOT                                              *  2B *
           DO 700 J = 1,NVAR                                              *  2B *
              DO 800 I = 1,NSUB                                           *  2B *
                 ABOOT(I,J) = AORIG(IBOOT(I), J)                          *  2B *
 800          CONTINUE                                                    *  2B *
 700       CONTINUE                                                       *  2B *
C***************************************************************************

C***************************************************************************
*          Selection du modele, determiner sa qualite                     *  2C *
           CALL REG(ABOOT, NVAR, NSUB, SELECT, IND,                       *  2C *
     *            QLT, COEF, MSE,SECOEF, BETA, CTRANK, CTRK2,             *  2C *
     *            ISX,WT,VARMOD,BETAL,B,SE,COV,RES,H,P,                   *  2C *
     *            WK,WKBK,Q1,QTMP,QMIN,EXSS,PFW,WKFW,IMVB,FREE,           *  2C *
     *            MODEL,IMVBCP,TAU,WORK,LWORK,R,CP,IW1,                   *  2C *
     *            IW2,IW3,W1,W2,W3,W4,BRSS)                               *  2C *
C***************************************************************************

C***************************************************************************
          Calculer Jn                                                     *  2D *
*                                                                         *  2D *
           DO 1000 I = 1, NVAR-1                                          *  2D *
              DCOEF(K,I) = SQRT(REAL(NSUB)/REAL(NOBS))*                   *  2D *
    *                      (COEF(I) - COEFO(I))                           *  2D *
 1000      CONTINUE                                                       *  2D *
C***************************************************************************

C***************************************************************************
      WRITE (3,*) '% QUALITE DU MODELE CLASSIQUE'                         *  2E *
      WRITE (3,*)                                                         *  2E *
      WRITE (3,99995) '% TROP PETITS', REAL(QLTCL(1)) / NREP *100, '%'    *  2E *
      WRITE (3,99995) '% CORRECTS', REAL(QLTCL(2)) / NREP * 100, '%'      *  2E *
      WRITE (3,99995) '% TROP GRANDS', REAL(QLTCL(3)) / NREP * 100, '%'   *  2E *
      WRITE (3,*) '% PAS DE PLEIN RANG AVANT ', CTCLR                     *  2E *
      WRITE (3,*) '% PAS DE PLEIN RANG APRES', CTCLR2                     *  2E *
      WRITE (3,*)                                                         *  2E *
      WRITE (3,*) '% QUALITE DU MODELE BOOTSTRAP'                         *  2E *
      WRITE (3,99995)'% TROP PETITS',REAL(QLTBT(1))/(NREP*NBOOT)*100,'%'  *  2E *
      WRITE (3,99995)'% CORRECTS', REAL(QLTBT(2))/(NREP*NBOOT) * 100,'%'  *  2E *
      WRITE (3,99995)'% TROP GRANDS',REAL(QLTBT(3))/(NREP*NBOOT)*100,'%'  *  2E *
      WRITE (3,*) '% PAS DE PLEIN RANG AVANT ', CTRANK                    *  2E *
      WRITE (3,*) '% PAS DE PLEIN RANG APRES', CTRK2                      *  2E *
```

```
      WRITE (3,*) ' '                                                    *  2E *
      WRITE (3,*) '% QUALITE DE COUVERTURE'                              *  2E *
      WRITE (3,*) '% IDC EN FORMAT LATEX'                                *  2E *
      WRITE (3,*)                                                        *  2E *
      WRITE (3,*) '\\begin{center}'                                      *  2E *
      WRITE (3,*) '\\begin{tabular}{|l|l|rrrrrrrr|} \\hline'             *  2E *
      WRITE (3,99989) '\\multicolumn{10}{|c|}{Matrice 1, $\\sigma =$',   *  2E *
     *                STDDEV, ', N=', NOBS, ', b=', NSUB,               *  2E *
     *               '} \\\\ \\hline \\hline'                           *  2E *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Coefficients} & $\\beta_0$ &',  *  2E *
     *            '$\\beta_1$&$\\beta_2$&$\\beta_3$&$\\beta_4$&$',      *  2E *
     *            '\\beta_5$&$\\beta_6$ & $\\beta_7$ \\\\'               *  2E *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Valeurs exactes}',              *  2E *
     *            '& 1,0&2,0&1,5&1,5&0,5&0&0&0\\\\ \\hline \\hline'      *  2E *
      WRITE (3,*) '\\multicolumn{10}{|c|}{',SELECT,'}\\\\ \\hline'       *  2E *
      WRITE (3,*) '\\multicolumn{2}{|c|}{subsample} & & & & & && &\\\\',*  2E *
     *            ' \\hline'                                             *  2E *
                                                                         *  2E *
      DO 4990 K = 1,4                                                    *  2E *
         DO 5000 J = 1, 3                                                *  2E *
            IF (J.EQ.1.AND.K.EQ.1) THEN                                  *  2E *
               WRITE(3,99994) 'classique & Uni. gauche 2,5\\%',          *  2E *
     *           ('&', REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'   *  2E *
            ELSE IF (J.EQ.1.AND.K.EQ.2) THEN                             *  2E *
               WRITE(3,99994) 'pivotal & Uni. gauche 2,5\\%',            *  2E *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'    *  2E *
            ELSE IF (J.EQ.1.AND.K.EQ.3) THEN                             *  2E *
               WRITE(3,99994) 'percentile & Uni. gauche 2,5\\%',         *  2E *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'    *  2E *
            ELSE IF (J.EQ.1.AND.K.EQ.4) THEN                             *  2E *
               WRITE(3,99994) 'bootstrap-t & Uni. gauche 2,5\\%',        *  2E *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'    *  2E *
            ELSE IF (J.EQ.2) THEN                                        *  2E *
               WRITE(3,99994) '& Uni. droite 2,5\\%',                    *  2E *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),'\\\\'    *  2E *
            ELSE                                                         *  2E *
               WRITE(3,99994) '& Bilateral 95\\%',                       *  2E *
     *           ('&',REAL(CPTCVT(K,I,J))/NREP*100,I=1,NVAR-1),          *  2E *
     *           '\\\\ \\hline'                                          *  2E *
                                                                         *  2E *
            END IF                                                       *  2E *
                                                                         *  2E *
5000     CONTINUE                                                        *  2E *
4990  CONTINUE                                                           *  2E *
      WRITE (3,*) '\\end{tabular}'                                       *  2E *
      WRITE (3,*) '\\end{center}'                                        *  2E *
```

```
                                                                      *  2E *
      WRITE (3,*)                                                     *  2E *
      WRITE (3,*) '% LONGUEUR DES INTERVALLES DE CONFIANCE CLASSIQUE'  *  2E *
      WRITE (3,*) '%  EN FORMAT LATEX'                                *  2E *
      WRITE (3,*)                                                     *  2E *
      WRITE (3,*) '\\begin{center}'                                   *  2E *
      WRITE (3,*) '\\begin{tabular}{|l|l|rrrrrrrr|} \\hline'          *  2E *
      WRITE (3,99989) '\\multicolumn{10}{|c|}{Matrice 1, $\\sigma=$ ',*  2E *
     *                STDDEV, ', N=', NOBS, ', b=', NSUB,            *  2E *
     *               '} \\\\ \\hline \\hline'                        *  2E *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Coefficients} & $\\beta_0$ &',*  2E *
     *            '$\\beta_1$&$\\beta_2$&$\\beta_3$&$\\beta_4$&$',   *  2E *
     *            '\\beta_5$&$\\beta_6$ & $\\beta_7$ \\\\'            *  2E *
      WRITE (3,*) '\\multicolumn{2}{|c|}{Valeurs exactes}',           *  2E *
     *            '& 1,0&2,0&1,5&1,5&0,5&0&0&0\\\\ \\hline \\hline'   *  2E *
      WRITE (3,*) '\\multicolumn{10}{|c|}{',SELECT,'}\\\\ \\hline'    *  2E *
      WRITE (3,*) '\\multicolumn{2}{|c|}{subsample} & & & & & && &\\\\',*  2E *
     *            ' \\hline'                                          *  2E *
                                                                      *  2E *
      DO 5040 K=1,4                                                   *  2E *
         IF (K.EQ.1) THEN                                             *  2E *
            WRITE (3,99990) 'classique & moyenne ',                   *  2E *
     *           ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'              *  2E *
            WRITE (3,99990) '          & ecart-type ',                 *  2E *
     *           ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'        *  2E *
         ELSE IF (K.EQ.2) THEN                                        *  2E *
            WRITE (3,99990) 'pivotal & moyenne ',                     *  2E *
     *           ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'              *  2E *
            WRITE (3,99990) '          & ecart-type ',                 *  2E *
     *           ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'        *  2E *
         ELSE IF (K.EQ.3) THEN                                        *  2E *
            WRITE (3,99990) 'percentile & moyenne ',                  *  2E *
     *           ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'              *  2E *
            WRITE (3,99990) '          & ecart-type ',                 *  2E *
     *           ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'        *  2E *
         ELSE                                                         *  2E *
            WRITE (3,99990) 'bootstrap-t & moyenne ',                 *  2E *
     *           ('&', MOYIDC(I,K), I=1, NVAR-1), '\\\\'              *  2E *
            WRITE (3,99990) '          & ecart-type ',                 *  2E *
     *           ('&', SEIDC(I,K), I=1,NVAR-1), '\\\\ \\hline'        *  2E *
         END IF                                                       *  2E *
                                                                      *  2E *
5040  CONTINUE                                                        *  2E *
      WRITE (3,*) '\\end{tabular}'                                    *  2E *
      WRITE (3,*) '\\end{center}'                                     *  2E *
      WRITE (3,*) '\\end{document}'                                   *  2E *
```

```
                                                                        *  2E *
99989 FORMAT (A, F5.2,A, I5,A, I5, A)                                   *  2E *
C*************************************************************************
```

BIBLIOGRAPHIE

BERK, K. (1978), Comparing subset regression procedures, *Technometrics* **20**, 1–6.

CARIGNAN, M. (1996), *Intervalles de confiance bootstrap suite à la sélection d'un modèle en régression linéaire multiple*, Mémoire de maîtrise, Département de mathématiques et de statistique, Université de Montréal.

DUCHARME, G. (1997), Consistent selection of the actual model in regression analysis, *Journal of Applied Statistics* **24**, 549–558.

EFRON, B. (1979), Bootstrap methods: Another look at the jackknife, *Annals of Statistics* **7**, 1–26.

EFRON, B. ET TIBSHIRANI, R.J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

FREEDMAN, D.A. (1981), Bootstrapping regression models, *Annals of Statistics* **9**, 1218–1228.

HURVICH, C. ET TSAI, C.L. (1990), The impact of model selection on inference in linear regression, *The American Statistician* **44**, 214–217.

L'ECUYER, P. (1995), Combined multiple recursive random number generators, *Operations Research* **44**, 816–822.

MALLOWS, C.L. (1973), Some comments on C_p, *Technometrics* **15**, 661–675.

MARSAGLIA, G. ET TSANG, W.W. (1984), A fast, easily implemented method for sampling from decreasing or symmetric unimodal density functions, *SIAM Journal of Scientific and Statistical Computing* **5**, 349–359.

MILLER, A.J. (1990), *Subset Selection in Regression*, Chapman and Hall, New York.

NETER, J., KUTNER, M.H., NACHTSHEIM, C.J. ET WASSERMAN, W. (1996), *Applied Linear Statistical Models*, $4^{\text{ième}}$ édition, Irwin, Chicago.

NISHII, R. (1984), Asymptotic properties of criteria for selection of variables in multiple regression, *The Annals of Statistics* **12**, 758–765.

POLITIS, D.N., ROMANO, J.P. ET WOLF, M. (1999), *Subsampling*, Springer-Verlag, New York.

POLITIS, D.N. ET ROMANO, J.P. (1994), Large sample confidence regions based on subsamples under minimal assumptions, *Annals of Statistics* **22**, 2031–2050.

SAMPSON, A.R. (1974), A tale of two regressions, *Journal of the American Statistical Association* **69**, 682–689.

SCHWARZ, G. (1978), Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

SHIBATA, R. (1984), Approximate efficiency of a selection procedure for the number of regression variables, *Biometrika* **71**, 43–49.

SMITH, D.M. (1991), All possible subset regressions using the QR decomposition, *Applied Statistics* **40**, 502–513.

THOMPSON, M.L. (1978a), Selection of variables in multiple regression: Part I. A review and evaluation. *International Statistical Review* **46**, 1–19.

THOMPSON, M.L. (1978b), Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples, *International Statistical Review* **46**, 129-146.

ZHANG, P. (1992), On the distribution properties of model selection criteria, *Journal of the American Statistical Association* **87**, 732–737.