B A
S E

Biotechnol. Agron. Soc. Environ. 2005 **9** (1), 11–18

Conséquences de la sélection de variables sur l'interprétation des résultats en régression linéaire multiple

Arcadius Yves Justin Akossou, Rodolphe Palm

Unité de Statistique et Informatique. Faculté universitaire des Sciences agronomiques de Gembloux. Avenue de la Faculté d'Agronomie, 8. B-5030 Gembloux (Belgique). E-mail : palm.r@fsagx.ac.be

Reçu le 30 mars 2004, accepté le 14 juillet 2004.

La sélection de variables, *a priori* ou *a posteriori*, est une pratique très courante en régression multiple. L'utilisateur n'est cependant pas toujours conscient des conséquences sur les résultats qu'induit cette sélection. Dans cette note, les notions de biais d'omission et de biais de sélection des variables sont illustrées à partir d'un exemple simulé. Les conséquences de la sélection des variables sur l'estimation des paramètres et sur les prédictions sont ensuite examinées. L'attention de l'utilisateur est attirée sur les risques d'interprétation abusive des coefficients de régression, particulièrement après sélection de variables. Par contre, les conséquences de la sélection des variables sur la qualité des valeurs prédites de la variable à expliquer sont assez limitées, du moins pour l'exemple examiné.
Mots-clés. Régression, sélection de variables, biais d'omission, biais de sélection, simulation, méthode statistique.

Consequences of variable selection on the interpretation of the results in multiple linear regression. *A priori* or *a posteriori* variable selection is a common practise in multiple linear regression. The user is however not always aware of the consequences on the results due to this variable selection. In this note, the presence of omission bias and selection bias is explained by means of a Monte Carlo experiment. The consequences of variable selection on the regression coefficients and on the predicted values are then analysed. The user's attention is drawn to the risk of misinterpretation of the regression coefficients, specially after variable selection. On the other hand, the consequences of variable selection on the predicted values of the response variable are rather limited, at least for the given example.
Keywords. Regression, variable selection, omission bias, selection bias, simulation, statistical method.

1. INTRODUCTION

La régression linéaire multiple est une méthode statistique largement utilisée pour modéliser la relation existant entre une variable y, appelée variable à expliquer, et plusieurs variables x_1, x_2, ..., x_j, ..., x_p, appelées variables explicatives. Cette modélisation est effectuée soit parce qu'on s'intéresse directement à la quantification de l'effet de chacune des variables x_j sur la variable y, soit parce qu'on souhaite disposer d'un modèle en vue de la prédiction des valeurs de la variable y à partir des valeurs des variables explicatives. Dans le premier cas, l'intérêt se porte sur les valeurs estimées des coefficients de régression, alors que dans le second cas, on ne cherche pas nécessairement à interpréter les valeurs de ces paramètres.

La construction d'un modèle commence par le choix des variables explicatives potentielles. Ce choix se fait sur la base de la connaissance que l'on a du problème mais il est souvent aussi lié à la disponibilité des informations. Quant aux données elles-mêmes, elles peuvent provenir soit d'un dispositif expérimental, soit d'observations ou de mesures réalisées dans un environnement non contrôlé.

Ces deux situations sont fondamentalement différentes car, dans le premier cas, l'expérimentateur définit lui-même les valeurs des variables explicatives et peut donc concevoir, s'il le souhaite, un dispositif conduisant à des variables explicatives orthogonales, c'est-à-dire non corrélées. Dans le second cas, par contre, les variables explicatives sont généralement corrélées surtout lorsque le nombre de variables explicatives est élevé. On se trouve alors confronté au problème de la multicolinéarité qui, dans le cas extrême (multicolinéarité exacte), conduit à une indétermination des paramètres, et, dans les cas moins extrêmes (multicolinéarité approximative), à des paramètres estimés avec une très faible précision.

Les conséquences de la multicolinéarité et différentes solutions alternatives à la régression classique mises au point pour en atténuer les effets sont présentées par Palm et Iemma (1995), notamment.

La solution généralement adoptée est cependant la réduction de la multicolinéarité par l'élimination d'une ou de plusieurs variables explicatives. Se pose alors le problème du choix des variables à éliminer de l'équation. Des synthèses relatives à ce sujet sont données par Hocking (1976), Miller (1990) et Thompson (1978a ; b).

L'objectif de cette note est d'attirer l'attention de l'utilisateur sur l'influence de l'élimination de variables sur les résultats de la régression, d'une part lorsque cette élimination est réalisée *a priori*, c'est-à-dire sans tenir compte des observations collectées et, d'autre part, lorsqu'elle est faite *a posteriori*, par la mise en œuvre d'un algorithme de sélection de variables. Nous mettrons notamment en évidence l'existence du biais d'omission et du biais de sélection. Nous examinerons également les conséquences de la sélection de variables sur la précision des estimations des coefficients de régression et sur la qualité des prédictions.

Bien que plusieurs résultats puissent être obtenus de manière analytique, nous avons eu recours à la simulation de données, afin de mieux illustrer les problèmes.

Nous nous plaçons dans le cas de la régression linéaire multiple avec des variables explicatives corrélées et nous considérons que les hypothèses classiques de la régression sont vérifiées.

Nous examinons successivement le modèle théorique, la génération des données et les méthodes de régression comparées.

2. MODÈLE THÉORIQUE ET NOTATIONS

On considère le modèle théorique classique suivant :

$$y = \boldsymbol{x}\,\boldsymbol{\beta} + \varepsilon,$$

où y est la variable dépendante, $\boldsymbol{x}$ est le vecteur-ligne relatif aux p variables explicatives, $\boldsymbol{\beta}$ est le vecteur-colonne des coefficients de régression théoriques et ε est le résidu.

Soit un échantillon aléatoire et simple de n individus provenant de la population décrite par le modèle ci-dessus. Pour ces n individus, le modèle théorique s'écrit :

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \varepsilon,$$

$\boldsymbol{y}$ étant le vecteur-colonne des n réalisations de la variable à expliquer, $\boldsymbol{X}$ la matrice, de dimensions $n \times p$, des valeurs observées des variables explicatives, $\boldsymbol{\beta}$ le vecteur des p paramètres et ε le vecteur des n résidus théoriques. On considère que ces n résidus sont des réalisations indépendantes d'une variable aléatoire normale, de moyenne nulle et d'écart-type égal à σ.

Sous les conditions énoncées ci-dessus, l'estimateur des moindres carrés :

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\,\boldsymbol{X})^{-1}\,\boldsymbol{X}'\,\boldsymbol{y},$$

est un estimateur non biaisé de $\boldsymbol{\beta}$, de matrice de variances et covariances égale à :

$$\boldsymbol{V}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

La variance d'un coefficient de régression particulier β_j ($j = 1, \ldots, p$) est l'élément jj de cette matrice, que nous désignons par :

$$V\left(\hat{\beta}_j\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj}.$$

La variance de la prédiction $\hat{y}_i$ correspondant à un vecteur $\boldsymbol{x}_i$ fixé est donnée par la relation :

$$V\left(\hat{y}_i\right) = \sigma^2 \boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_i'$$

et la variance de l'erreur de prédiction est égale à :

$$V\left(y_i - \hat{y}_i\right) = \sigma^2 \left[1 + \boldsymbol{x}_i (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_i'\right].$$

Des estimations non biaisées de ces différentes quantités sont obtenues en remplaçant la variance résiduelle théorique σ^2 par son estimation non biaisée :

$$\hat{\sigma}^2 = \left(\boldsymbol{e}'\boldsymbol{e}\right) \big/ (n - p),$$

avec

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}}.$$

Par la suite, nous serons à plusieurs reprises amenés à considérer deux groupes de variables explicatives : celles qui figurent dans l'équation de régression calculée et celles qui n'y figurent pas. La matrice $\boldsymbol{X}$ et le vecteur $\boldsymbol{\beta}$ sont alors partitionnés :

$$\boldsymbol{X} = \left(\boldsymbol{X}_A \boldsymbol{X}_B\right) \text{ et } \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_A \\ \boldsymbol{\beta}_B \end{bmatrix},$$

et le modèle théorique s'écrit :

$$\boldsymbol{y} = \boldsymbol{X}_A \boldsymbol{\beta}_A + \boldsymbol{X}_B \boldsymbol{\beta}_B + \boldsymbol{\varepsilon}.$$

Dans le modèle théorique, nous n'avons pas fait explicitement apparaître le terme indépendant. Si on souhaite qu'un terme indépendant figure dans le modèle, il suffit de considérer qu'une colonne de $\boldsymbol{X}$ est une variable artificielle, dont toutes les observations sont égales à l'unité. Par la suite, nous considérons que le terme indépendant, noté β_0, correspond au coefficient de la variable artificielle x_0 et que les coefficients β_j ($j = 1, \ldots, p$) sont les coefficients des variables explicatives x_j. Dans le cas du modèle partitionné, la variable artificielle se situe dans la matrice $\boldsymbol{X}_A$.

3. GÉNÉRATION DES DONNÉES

Comme nous l'avons signalé dans l'introduction, nous nous proposons de mettre en évidence de manière empirique les conséquences de la sélection des variables sur les résultats de la régression. Cette approche repose sur la réalisation d'un grand nombre de répétitions des différentes procédures de régression, à partir de données répondant à un même modèle théorique connu. Dans la mesure où des données permettant de mettre en œuvre cette approche n'existent pas, le modèle théorique n'étant jamais connu, le recours à la méthode de Monte-Carlo, basée sur la génération par ordinateur des données selon un modèle théorique, constitue la seule alternative envisageable.

Ce modèle théorique peut être déduit de la modélisation d'un jeu de données réelles ou, à l'autre extrême, être purement artificiel.

La solution retenue dans la présente étude est intermédiaire. Le modèle est artificiel mais ses caractéristiques ont été choisies de manière à ce qu'il corresponde à une situation analogue à celle que l'utilisateur est susceptible de rencontrer en pratique et qui ne soit pas particulièrement extrême. Le nombre de variables explicatives, à l'exclusion de la variable artificielle, est de cinq. Ces variables explicatives ont des distributions normales réduites. L'indice de colinéarité, défini comme la moyenne des inverses des valeurs propres, a été fixé à 20. Cette valeur est considérée comme intermédiaire par plusieurs auteurs (Hoerl *et al.*, 1986 ; Marquardt, Snee, 1975 ; Rencher, Pun, 1980). À titre de comparaison, cet indice est pratiquement identique à la valeur observée dans le cas d'un exemple réel traité par Palm (1988) et concernant l'étude de l'accroissement en grosseur d'épicéas en fonction de leur circonférence et de cinq caractéristiques du peuplement dont ils proviennent.

Les variables explicatives ont été générées à partir des procédures proposées par Hoerl *et al.* (1986) et Bendel, Afifi (1977).

Le vecteur $\boldsymbol{y}$ est généré à partir du modèle théorique. Le vecteur $\boldsymbol{\varepsilon}$ est constitué de réalisations indépendantes de variables aléatoires normales réduites ce qui signifie que la variance résiduelle est égale à l'unité et, pour donner une importance croissante aux variables explicatives, on a considéré que le vecteur $\boldsymbol{\beta}$ est tel que $\beta_i = i\,\beta_1$ ($i = 2, ..., 5$) . La valeur de β_1 a été fixée à 0,27 afin que le coefficient de détermination multiple R^2 de la régression de y sur l'ensemble des variables soit théoriquement égal à 0,80, et l'ordonnée à l'origine β_0 a été fixée à zéro.

Enfin, on a considéré que la taille de l'échantillon est de 50.

À partir de ce modèle théorique, 52.000 individus ont été générés. Les 50.000 premiers individus constituent les données de 1.000 répétitions de 50 individus et les 2.000 derniers individus servent à la validation.

Nous nous plaçons donc dans le cas du modèle aléatoire, pour lequel la matrice des variables explicatives résulte de l'échantillonnage et est différente d'un échantillon à l'autre. L'étude a également été réalisée dans le cas du modèle fixe, pour lequel la matrice des variables explicatives est identique pour chaque répétition. Les résultats obtenus pour le modèle fixe, qui ne sont cependant pas repris pour ne pas alourdir le texte, sont dans l'ensemble comparables à ceux obtenus pour le modèle aléatoire.

Pour l'ensemble des données simulées, la matrice de corrélation des cinq variables explicatives x_i ($i = 1$, ..., 5) est la suivante :

1,00	– 0,83	– 0,85	0,77	0,29
– 0,83	1,00	0,96	– 0,60	– 0,30
– 0,85	0,96	1,00	– 0,68	– 0,24
0,77	– 0,60	– 0,68	1,00	0,69
0,29	– 0,30	– 0,24	0,69	1,00

et les corrélations de la variable à expliquer y avec les variables explicatives sont respectivement égales à 0,16, 0,02, 0,02, 0,59 et 0,81.

Rappelons que ces différentes corrélations ne sont pas directement contrôlées mais résultent du contrôle des autres facteurs. Les corrélations entre les variables explicatives varient, en valeur absolue, de 0,24 à 0,96. Les corrélations les plus importantes concernent les trois premières variables qui sont aussi précisément les variables les moins corrélées à la variable à expliquer. L'examen de ces différentes corrélations montre clairement qu'on ne se trouve pas en présence d'une situation extrême.

4. MÉTHODES DE RÉGRESSION COMPARÉES

On se propose de comparer les résultats de régressions calculées de trois manières différentes.

La première méthode consiste à établir l'équation de régression en prenant en compte les cinq variables (modèle complet). Cette méthode correspond à une situation idéale puisqu'on considère que l'utilisateur connaît *a priori* et dispose de toutes les variables qui interviennent dans le modèle théorique.

Dans la deuxième méthode, l'équation de régression est calculée sur un sous-ensemble de variables sélectionnées *a priori*. On se place ainsi dans le cas d'un utilisateur qui, par manque de connaissance du modèle théorique ou pour des raisons pratiques, n'aurait pas collecté l'information concernant une ou plusieurs variables qui sont pourtant présentes dans le modèle théorique.

Enfin, dans la troisième méthode, la sélection des variables se fait *a posteriori*, c'est-à-dire sur la base de données observées, l'algorithme retenu étant la sélection mixte pas à pas (*stepwise*) avec une valeur *F* pour l'entrée et la sortie d'une variable égale à l'unité. Cette méthode correspond à une procédure très couramment utilisée en pratique : ne connaissant pas le modèle théorique, l'utilisateur procède à la recherche du modèle et exclut les variables qu'il juge non pertinentes, au vu des données dont il dispose.

Pour les modèles avec sélection, toutes les combinaisons de variables ont été étudiées, mais nous nous limiterons principalement aux résultats obtenus pour le sous-ensemble constitué des variables x_3, x_4 et x_5, cette combinaison étant celle qui a été retenue le plus grand nombre de fois par la méthode de sélection (267 sélections pour les 1.000 régressions).

Pour chaque équation de régression, les caractéristiques suivantes ont été notées :
- les valeurs estimées des coefficients de régression ;
- les erreurs-standards estimées des coefficients de régression ;
- l'écart-type résiduel estimé et le coefficient de détermination multiple ;
- l'erreur quadratique moyenne de prédiction ;
- le coefficient noté R_p^2 défini ci-dessous.

L'erreur quadratique moyenne de prédiction EQMP est calculée sur les 2.000 individus réservés à la validation par la formule :

$$\text{EQMP}=\left(\frac{1}{2000}\sum_{i=1}^{2000}(y_i-\hat{y}_i)^2\right)^{1/2},$$

et le coefficient R_p^2 est une mesure analogue au coefficient de détermination multiple mais il est calculé, lui aussi, à partir des données de validation. Il est défini de la manière suivante :

$$R_p^2=1-\sum_{i=1}^{2000}(y_i-\hat{y}_i)^2 \Big/ \sum_{i=1}^{2000}(y_i-\bar{y})^2,$$

la moyenne $\bar{y}$ étant la moyenne arithmétique des valeurs de la variable à expliquer de l'échantillon qui a servi au calcul de la régression. Ce paramètre mesure le gain de précision, lors de la prédiction, lié à l'utilisation de l'équation de régression, plutôt que la moyenne.

Les moyennes et les écarts-types des caractéristiques ci-dessus ont ensuite été calculés pour les 1.000 équations de régression faisant intervenir l'ensemble des variables (modèle complet), pour les 1.000 équations faisant intervenir les variables x_3, x_4 et x_5 (sélection *a priori*) et enfin pour le sous-groupe des 267 répétitions pour lesquelles la procédure de sélection pas à pas a conduit à la sélection des variables x_3, x_4 et x_5 (sélection *a posteriori*).

5. RÉSULTATS ET DISCUSSION

5.1. Biais d'omission

Le **tableau 1** reprend, pour les différents cas, les moyennes observées des coefficients de régression. Celles-ci sont exprimées en pour cent des valeurs théoriques β_i correspondantes. Le tableau donne également les valeurs des différents types de biais, qui sont définis dans la suite. Bien que les équations aient été déterminées avec un terme indépendant, nous n'avons pas repris les valeurs obtenues pour ce paramètre dans le tableau, celles-ci étant, dans l'ensemble, très proches de zéro et ne présentant pas de différences notables pour les différents types de régression.

Pour le modèle complet, on constate que les moyennes des paramètres sont proches des valeurs théoriques. Cette bonne concordance était attendue puisque la méthode des moindres carrés est une méthode non biaisée lorsque le modèle ajusté correspond au modèle théorique : en moyenne, elle doit donc conduire aux valeurs théoriques des coefficients.

Par contre, pour les régressions avec sélection, les moyennes sont fort différentes de celles obtenues pour les modèles complets. Ainsi, pour la sélection *a priori* et pour le coefficient β_3 par exemple, l'écart entre les deux moyennes est de 160,4 - 97,9 = 62,5.

Cet écart correspond au biais sur β_3, lié à l'élimination des variables x_1 et x_2 ; il est appelé biais d'omission. Les différentes valeurs de ce biais sont reprises dans le **tableau 1**.

On constate donc que la suppression *a priori* des variables x_1 et x_2 conduit, lors de l'estimation des coefficients des variables x_3, x_4 et x_5, à un biais variant, en valeur absolue, de 18 à 63 %, selon le paramètre considéré.

Les variables x_3, x_4 et x_5 sont les variables qui figurent le plus souvent dans les équations à la suite de la

Tableau 1. Moyennes et biais des coefficients de régression estimés (en pour cent des valeurs théoriques) — *Means and bias of the estimated regression coefficients (in percent of theoretical values)*.

Modèles Caractéristiques	β_1	β_2	β_3	β_4	β_5
Modèle complet					
Moyenne	115,6	108,1	97,9	96,4	101,8
Sélection *a priori*					
Moyenne			160,4	133,1	84,2
Biais d'omission			62,5	36,7	-17,6
Sélection *a posteriori*					
Moyenne			166,3	140,8	81,0
Biais global			68,4	44,4	-20,8
Biais de sélection			5,9	7,7	-3,2

sélection pas à pas et qui ont les coefficients de régression théoriques les plus grands. Il s'agit indiscutablement d'un bon sous-ensemble de variables. Si au contraire, on sélectionne, par exemple, les variables x_1, x_2 et x_3 et qu'on élimine les variables x_4 et x_5, le biais d'omission sera bien plus important. En effet, les calculs, dont les résultats ne sont pas repris ici, ont montré que, pour β_1, le biais dépasse, dans ce cas, 420 %.

Lorsque le modèle théorique est connu, le biais d'omission peut être calculé, du moins pour le modèle fixe (Draper, Smith, 1998 ; Miller, 1990). La formule de calcul montre que ce biais est d'autant plus important que les variables omises sont importantes (coefficient de régression important par rapport à l'écart-type de la variable explicative) et que les variables omises sont corrélées aux variables retenues. Enfin, ce biais n'est pas lié à l'effectif de l'échantillon. Dans la pratique, le modèle théorique est évidemment inconnu et le biais d'omission ne peut pas être calculé.

5.2. Biais de sélection

Si on ne tient compte que des 267 répétitions pour lesquelles la procédure de sélection des variables pas à pas a conduit à la sélection des variables x_3, x_4 et x_5 (sélection *a posteriori*), on obtient des valeurs moyennes pour les coefficients qui s'écartent plus des valeurs obtenues pour le modèle complet que lorsque l'on considère l'ensemble des 1.000 répétitions.

À titre d'exemple, si on reprend le coefficient β_3, on a vu que, pour la sélection *a priori*, l'écart est de 62,5 alors que pour la sélection *a posteriori*, il est de 166,3 - 97,9 = 68,4 .

Cet écart correspond au biais global, dont les valeurs pour les différents paramètres sont données dans le **tableau 1**.

Pour le coefficient β_3, la différence entre le biais global et le biais d'omission est égale à 5,9. Cette valeur est aussi égale à la différence entre la moyenne des coefficients obtenus après sélection *a posteriori* (moyenne de 267 valeurs) et la moyenne des coefficients obtenus après sélection *a priori* (moyenne de 1.000 valeurs). Cet écart ne doit pas être attribué au hasard. Si on augmentait le nombre de simulations (en considérant par exemple 10.000 répétitions au lieu de 1.000), on obtiendrait un résultat du même ordre de grandeur. L'écart provient du fait que la sélection des variables et l'estimation des paramètres des variables sélectionnées se font sur la base des mêmes observations. Il est appelé biais de sélection et il s'ajoute au biais d'omission. Pour l'exemple considéré, ce biais est relativement faible par rapport au biais d'omission (**Tableau 1**).

Il n'existe pas de formule permettant de quantifier le biais de sélection, qui dépend de la méthode de sélection retenue. Des techniques d'estimation de ce biais, basées sur l'utilisation de méthodes de Monte-Carlo ont été proposées. Une discussion de ce problème peut être trouvée dans Miller (1990).

5.3. Erreurs-standards des coefficients de régression

Pour chaque type de modèles, le **tableau 2** donne l'écart-type des coefficients de régression estimés lors des répétitions et la moyenne des erreurs-standards estimées pour chaque équation de régression par la formule suivante :

$$\text{erreur–standard} = \left(\hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{jj}\right)^{1/2},$$

$\hat{\sigma}^2$ étant la variance résiduelle estimée et $\boldsymbol{X}$ la matrice des variables explicatives figurant dans l'équation (paragraphe 2). Pour les modèles avec sélection, la matrice $\boldsymbol{X}$ est remplacée par $\boldsymbol{X}_A$. Comme pour le **tableau 1**, les résultats ont été exprimés en pour cent du coefficient de régression théorique.

De façon globale, il y a peu de différences entre l'écart-type des coefficients de régression et la moyenne des erreurs-standards estimées au sein d'un même type de modèles. Cette concordance est liée au caractère non biaisé de l'estimateur de l'erreur-standard d'un coefficient de régression, même en présence de sélection de variables.

Par contre, pour un coefficient donné, on observe des résultats très différents pour le modèle complet et pour le modèle avec sélection. Si la sélection (*a priori* ou *a posteriori*) fait apparaître un biais, elle permet cependant de réduire, et parfois de manière importante, la variabilité des estimateurs.

Tableau 2. Écarts-types des coefficients de régression estimés et moyennes des erreurs-standards (en pour cent des valeurs théoriques) — *Standard deviations of the estimated coefficients of regression and means of the standard errors (in percent of the theoretical values)*.

Modèles / **Caractéristiques**	β_1	β_2	β_3	β_4	β_5
Modèle complet					
Écart-type	166,9	163,4	113,4	55,3	29,3
Moyenne	164,7	162,9	113,5	56,6	29,7
Sélection *a priori*					
Écart-type			28,9	28,6	17,6
Moyenne			28,3	28,6	17,2
Sélection *a posteriori*					
Écart-type			28,3	26,7	16,4
Moyenne			27,8	27,8	16,7

Pour le cas considéré, l'erreur-standard de β_3, par exemple, est divisée par un facteur de l'ordre de quatre.

La variabilité d'un paramètre est également un peu plus faible dans le cas de la sélection *a posteriori*, par rapport à la sélection *a priori*. Comme signalé antérieurement, les régressions obtenues après sélection *a posteriori* sont un sous-ensemble des régressions calculées dans le cas de la sélection *a priori*. Ce sous-ensemble n'est pas obtenu par un choix au hasard parmi les 1.000 régressions calculées, mais il est plus homogène que l'ensemble des 1.000 régressions, puisque les cas retenus sont précisément tous ceux qui ont conduit, par la méthode de sélection, au choix des variables x_3, x_4 et x_5.

Dans la mesure où, à la suite de la sélection, les estimateurs sont biaisés, la comparaison entre le modèle complet et le modèle avec sélection doit se faire sur l'erreur quadratique plutôt que sur l'erreur-standard. Celle-ci est égale à la racine carrée de la variance augmentée du carré du biais. Pour la sélection *a priori* on obtient, pour β_3 par exemple :

$$(28{,}3^2 + 62{,}5^2)^{1/2} = 68{,}6$$

De la même manière, on trouve 47 pour β_4 et 25 pour β_5. Pour cet exemple, les estimateurs biaisés basés sur la sélection *a priori* sont donc caractérisés par un écart quadratique moyen plus faible que les estimateurs du modèle complet. Il ne s'agit cependant pas là d'une règle générale et, pour d'autres sous-ensembles de variables, on pourrait obtenir la situation inverse.

5.4. Interprétation des coefficients de régression

Le **tableau 2** a montré que les écarts-types des coefficients de régression sont importants surtout dans le cas du modèle complet. Cela signifie que, en fonction des hasards de l'échantillonnage, les estimations des coefficients peuvent être très différentes.

Ainsi, pour le modèle complet et pour le coefficient β_5, l'examen de la distribution des 1000 valeurs estimées du coefficient, exprimées en pour cent de la valeur théorique, a montré que celles-ci varient de 3 à 232. Les pourcentiles 0,025 et 0,975 sont respectivement égaux à 46 et 158. Ces pourcentiles correspondent donc aux limites de l'intervalle qui contient les 95 % d'estimations les plus centrales. Pour le coefficient β_1, cet intervalle est beaucoup plus grand : il s'étend de -226 à 451. Pour les autres paramètres, on observe des situations intermédiaires.

Si on considère le modèle avec sélection *a priori* des variables x_3, x_4 et x_5, les intervalles s'étendent de 102 à 215 pour β_3, de 77 à 188 pour β_4 et de 50 à 118 pour β_5. Par rapport au modèle complet, les intervalles sont donc nettement plus réduits.

Dans le cas de la sélection *a posteriori*, si on considère uniquement les 267 répétitions ayant conduit à la sélection des variables x_3, x_4 et x_5, les intervalles sont très légèrement plus faibles que ceux obtenus pour la sélection *a priori*. Par contre, si on prend en compte les 1.000 répétitions et qu'on fixe à zéro la valeur estimée du paramètre lorsque la variable n'a pas été sélectionnée, la situation est totalement différente. Pour β_5, la distribution des 1.000 valeurs estimées est semblable à la distribution des valeurs obtenues dans le cas du modèle complet. On notera que la variable x_5 a été sélectionnée 998 fois sur les 1.000 répétitions. Pour β_3 et β_4, les distributions sont caractérisées par la présence d'un grand nombre de valeurs nulles correspondant aux répétitions pour lesquelles les variables x_3 ou x_4 n'ont pas été sélectionnées (551 valeurs nulles pour x_3 et 243 valeurs nulles pour x_4). Globalement les intervalles contenant les 95 % des estimations les plus centrales sont du même ordre de grandeur que les intervalles observés dans le cas du modèle complet.

À titre d'illustration, la **figure 1** donne la distribution des valeurs estimées pour le coefficient β_4 pour quatre situations différentes. Ce paramètre a été choisi en raison de son comportement intermédiaire. Le premier graphique donne la distribution des 1.000 coefficients pour le modèle complet ; le deuxième graphique donne la distribution des 1.000 coefficients dans le cas de la sélection *a priori* du modèle comportant les variables x_3, x_4 et x_5 ; le troisième

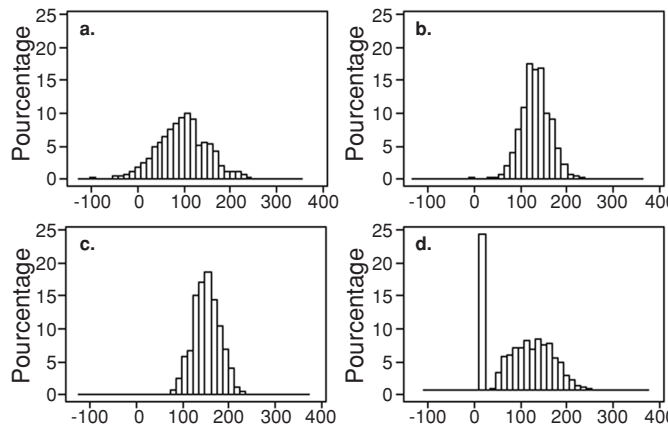

Figure 1. Distribution des coefficients de régression estimés pour la variable x_4, exprimés en pour cent de la valeur théorique β_4 : modèle complet (a), sélection *a priori* (b), sélection *a posteriori* et 267 répétitions (c), sélection *a posteriori* et 1.000 répétitions (d) — *Distribution of the estimated regression coefficients for variable x_4, in percent of the theoretical value of β_4: model without variable selection* (a), a priori *variable selection* (b), a posteriori *variable selection for 267 samples* (c), a posteriori *variable selection for 1.000 samples* (d).

graphique donne la distribution des 267 coefficients pour les répétitions ayant conduit à la sélection des variables x_3, x_4 et x_5. Enfin, le quatrième graphique donne la distribution des 1.000 valeurs dans le cas de la sélection *a posteriori*, la valeur nulle étant attribuée au coefficient de la variable x_4 pour les 243 répétitions pour lesquelles la variable x_4 n'a pas été sélectionnée. Les distributions (a), (b) et (c) sont les distributions dont on a repris les moyennes dans le **tableau 1** et les écarts-types dans le **tableau 2**, dans la colonne intitulée β_4.

L'importance du domaine de variation des paramètres estimés dans le cas du modèle complet rend assez illusoire l'interprétation d'un coefficient de régression. Pour la sélection *a priori*, la variabilité des estimations est nettement plus faible, mais compte tenu de l'existence du biais d'omission, la situation n'est pas nécessairement meilleure. Au contraire, le fait de négliger le biais dans l'interprétation, puisqu'il est inconnu en pratique, peut conduire l'utilisateur à accorder une confiance bien trop grande dans ses résultats. Si, pour la sélection *a posteriori*, on tient compte de l'effet du hasard lors de la sélection, la variabilité des estimations des paramètres est analogue à celle observée pour le modèle complet. Si par contre on ne tient compte que de la variabilité liée à l'estimation, les risques d'une interprétation erronée du coefficient de régression sont encore plus grands que pour la sélection *a priori*, puisqu'au biais d'omission, il faut encore ajouter le biais de sélection.

En définitive, quelle que soit la méthode de régression retenue, les valeurs obtenues pour les paramètres sont donc pratiquement non interprétables.

5.5. Qualité de l'ajustement et des prédictions

Le **tableau 3** donne les valeurs moyennes et les écarts-types, d'une part, pour les deux paramètres liés à la qualité de l'ajustement, $\hat{\sigma}$ et R^2, et, d'autre part, pour les deux paramètres liés à la qualité des prédictions, EQMP et R_p^2 . Rappelons que, pour chaque répétition, les deux premiers paramètres ont été déterminés sur les 50 individus utilisés pour le calcul de la régression et différents d'une répétition à l'autre et que les deux derniers ont été déterminés à partir des 2.000 individus de validation, identiques d'une répétition à l'autre (paragraphe 4). Dans le cas de la sélection, les résultats concernent le modèle comportant les variables x_3, x_4 et x_5.

On n'observe pas de différences importantes entre les régressions sans sélection et les régressions avec sélection de variables. Pour les critères d'ajustement, la sélection *a priori* semble très légèrement inférieure à la sélection *a posteriori* (écart-type résiduel plus grand et R^2 plus faible), ce qui peut s'expliquer par le fait que la sélection *a posteriori* vise précisément une certaine optimisation de l'ajustement : les trois variables ne sont en effet pas sélectionnées lorsqu'elles conduisent à un ajustement moins satisfaisant que d'autres ensembles de variables.

Tableau 3. Moyennes et écarts-types des paramètres $\hat{\sigma}$, R^2, EQMP et R_p^2 — *Means and standard deviations of parameters* $\hat{\sigma}$, R^2, EQMP and R_p^2.

Modèles / **Caractéristiques**	$\hat{\sigma}$	R^2	EQMP	R_p^2
Modèle complet				
Moyenne	0,996	0,814	1,061	0,780
Écart-type	0,106	0,049	0,042	0,019
Sélection *a priori*				
Moyenne	1,003	0,803	1,039	0,789
Écart-type	0,106	0,051	0,033	0,015
Sélection *a posteriori*				
Moyenne	0,988	0,811	1,037	0,790
Écart-type	0,100	0,048	0,030	0,014

On remarque également que les EQMP sont toujours supérieurs à l'écart-type résiduel moyen ; les erreurs de prédiction, en valeur absolue, sont donc en moyenne plus grandes que les erreurs d'ajustement. Les EQMP sont aussi légèrement inférieurs lorsqu'il y a sélection de variables.

Globalement, le **tableau 3** montre que les conséquences de la sélection des variables explicatives sur la qualité des prédictions sont assez limitées, l'erreur quadratique moyenne n'étant que de 4 à 6 % supérieure à l'écart-type théorique des résidus. On note aussi que des modèles différents (modèles à cinq variables et modèles à trois variables) conduisent à des prédictions de qualité en moyenne à peu près équivalente.

6. CONCLUSIONS

L'objectif de ces simulations était de mettre en évidence quelques problèmes liés à la sélection de variables.

Le modèle théorique retenu pour ces simulations décrit évidemment un cas bien particulier et les résultats ne peuvent pas être généralisés à toutes les situations rencontrées en pratique. On peut considérer qu'il correspond à une situation qu'on pourrait qualifier d'intermédiaire : le nombre de variables disponibles n'est pas très élevé, la colinéarité n'est pas excessive, le nombre d'observations est raisonnable, la corrélation entre la variable à expliquer et les variables explicatives est bonne, sans être excellente.

D'autre part, nous avons vu que, partant d'un modèle théorique, certains résultats peuvent être obtenus par voie analytique alors que d'autres, par contre, ne peuvent être établis que par simulations.

Les conséquences de la sélection des variables ont été examinées pour la sélection *a priori* et pour la sélection *a posteriori*. Ces deux situations se rencontrent très fréquemment en pratique. En effet, la liste des variables devant figurer dans le modèle théorique n'est, en pratique, jamais connue et l'utilisateur observe le plus souvent une série de variables qu'il suppose pertinentes. Il n'a cependant pas la certitude que des variables devant figurer dans le modèle théorique n'ont pas été omises lors de la collecte des données. L'omission d'une transformation non linéaire d'une variable, telle que le carré par exemple, pour la prise en compte d'un phénomène de non-linéarité est également analogue à une sélection *a priori*.

Quant à la sélection *a posteriori*, c'est-à-dire sur la base des données observées, elle constitue plutôt la règle que l'exception. En effet, l'utilisateur se trouve fréquemment en présence d'un nombre important de variables explicatives dont il ne connaît pas, à l'avance, l'impact sur la variable à expliquer et une sélection de variables lui paraît indispensable de manière à obtenir un modèle final d'une complexité raisonnable.

La sélection des variables peut également être vue comme la recherche d'un compromis entre la variance et le biais des estimateurs : plus on élimine des variables qui devraient figurer dans le modèle, plus la variance des estimateurs se réduit mais plus leur biais augmente. L'écart quadratique moyen, qui prend en compte, à la fois, la dispersion et le biais de l'estimateur peut, en effet, être plus petit après sélection qu'en l'absence de sélection.

En plus du biais d'omission, la sélection des variables sur base des observations introduit un biais de sélection dont l'importance, pour un sous-ensemble donné de variables, dépend de l'intensité de la compétition exercée par d'autres sous-ensembles concurrents. Pour l'exemple traité, ce biais est faible par rapport au biais d'omission. Il peut être éliminé si on divise l'échantillon initial en deux parties, le premier sous-échantillon étant utilisé pour la sélection des variables et l'autre sous-échantillon servant à l'estimation des paramètres des variables retenues. Cette solution n'est cependant pas à recommander, car l'élimination du biais de sélection se fait au détriment de l'erreur-standard des paramètres estimés. En effet, la division de l'échantillon initial en deux sous-échantillons de même effectif conduit par exemple à multiplier l'erreur-standard des coefficients de régression par un facteur égal à $\sqrt{2}$, soit 1,4.

L'existence dans le modèle théorique de plusieurs variables corrélées a comme effet, nous venons de le rappeler, de conduire à des coefficients de régression présentant une erreur quadratique moyenne importante, que celle-ci soit le fait d'une erreur-standard importante ou de la combinaison d'une erreur-standard et d'un biais. Concrètement, cela signifie une très grande incertitude sur la valeur des paramètres qui entraîne des difficultés dans l'interprétation de ceux-ci.

L'utilisateur doit donc être mis en garde contre les dangers d'une interprétation de la valeur ponctuelle d'un coefficient estimé. Selon les hasards de l'échantillonnage et selon les autres variables considérées dans le modèle, la valeur estimée du coefficient d'une variable particulière peut varier dans des proportions très importantes et des changements de signe ne sont pas exceptionnels. Le risque d'une interprétation abusive est évidemment plus grand lorsqu'il y a sélection de variables *a priori* ou *a posteriori*, puisque, le biais d'omission est inconnu.

Lorsque l'objectif de la modélisation est la prédiction, le choix d'une méthode de construction du modèle semble par contre avoir moins d'influence, car des modèles fort différents peuvent donner lieu à des prédictions de qualité très comparable, du moins lorsque le rapport de la taille de l'échantillon au nombre de variables explicatives est de l'ordre de la dizaine ou plus.

Bibliographie

Bendel RB., Afifi AA. (1977). Comparison of stopping rules in forward stepwise regression. *J. Am. Stat. Assoc.* **72**, p. 46–53.

Draper NNR., Smith H. (1998). *Applied linear regression*. New York: Wiley, 706 p.

Hocking R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, p. 1–49.

Hoerl AE., Kennard RW., Schuenemeyer JH. (1986). A simulation of biased estimation and subset selection regression technic. *Technometrics* **28**, p. 369–380.

Marquardt DW., Snee RD. (1975). Ridge regression in practice. *Amer. Stat.* **29**, p. 3–19.

Miller AJ. (1990). *Subset selection in regression*. New York: Chapman and Hall, 227 p.

Palm R. (1988). Les critères de validation des équation de régression linéaire. *Notes Stat. Inform. (Gembloux)* **88**/1, 27 p.

Palm R., Iemma AF. (1995). Quelques alternatives à la régression classique dans le cas de la colinéarité. *Rev. Stat. Appl.* **43** (2), p. 5–33.

Rencher AC., Pun FC. (1980). Inflation of R^2 in best subset regression. *Technometrics* **22**, p. 49–53.

Thompson M. (1978a). Selection of variables in multiple regression. Part I: a review and evaluation. *Int. Stat. Rev.* **46**, p. 1–19.

Thompson M. (1978b). Selection of variables in multiple regression. Part II: chosen procedures, computation and examples. *Int. Stat. Rev.* **46**, p. 129–146.

(11 réf.)