Intervalles de confiance valides en présence de sélection de modèle

François Bachoc, Hannes Leeb et Benedikt M. Pötscher

Department of Statistics and Operations Research, University of Vienna,
francois.bachoc@univie.ac.at
hannes.leeb@univie.ac.at
benedikt.poetscher@univie.ac.at

Résumé. Dans le contexte de la régression linéaire, on considère l'inférence statistique en présence de sélection de modèle. Sur ce sujet, Berk et al. (Annals of Statistics, 2013) ont récemment introduit une nouvelle classe d'intervalles de confiance, appelés intervalles de confiance PoSI, qui couvrent une certaine quantité d'intérêt non-standard. Ces intervalles de confiance sont uniformément valides, quelle que soit la procédure de sélection de modèle sous-jacente. Dans cet article, nous généralisons les intervalles de confiance PoSI à la prédiction post-sélection de modèle. Nous définissions deux prédicteurs non-standards : le premier étant l'extension naturelle de la quantité d'intérêt de Berk et al., le second ayant des propriétés d'optimalité plus pertinentes. Pour ces deux prédicteurs, nous construisons des intervalles de confiance, qui étendent ceux de Berk et al., et nous donnons des résultats théoriques, exacts et asymptotiques, associés. Nous renforçons ces résultats par une étude de simulation.

Mots-clés. Inférence post-sélection de modèle, intervalles de confiance, prédicteurs optimaux post-sélection de modèle, quantités d'intérêt non standard, régression linéaire.

Abstract. We consider inference post-model-selection in linear regression. In this setting, Berk et al. (Annals of Statistics, 2013) recently introduced a class of confidence sets, the so-called PoSI intervals, that cover a certain non-standard quantity of interest with a user-specified minimal coverage probability, irrespective of the model selection procedure that is being used. In this paper, we generalize the PoSI intervals to post-model-selection predictors. We define two non-standard predictors: the first one being the natural extension of the quantity of interest of Berk et al., the second one having more relevant optimality properties. For these two predictors, we construct confidence intervals, extending those of Berk et al., and give corresponding algorithms and exact and asymptotic coverage properties. We reinforce these results by a simulation study.

Keywords. Inference post-model-selection, confidence intervals, optimal post-model-selection predictors, non-standard targets, linear regression.

1 Contexte : sélection de modèle et coefficients de régression non-standards

On considère le modèle de régression linéaire

$$Y = X\beta + U, \tag{1}$$

où X est la matrice $n \times p$ de variables explicatives, $U \sim N(0, \sigma^2 I_n)$ est le vecteur aléatoire de résidus et β est le paramètre de régression inconnu. Dans cette section, X est considérée fixe.

On note $\hat{\beta}$ l'estimateur par moindre carrés ordinaires de β, $\hat{\beta} = (X'X)^{-1}X'Y$ dans le modèle (1), que nous appelons modèle linéaire complet. On considère l' estimateur $\hat{\sigma}^2$ standard (non-biaisé) de σ^2. On note $P_{n,\beta,\sigma}$ la loi de Y, qui dépend des paramètres inconnus β et σ et de la taille d'échantillon n.

On considère des sous-modèles du modèle linéaire complet (1), qui sont obtenus en supprimant des colonnes de X. Ainsi, un sous-modèle est représenté par sous-ensemble de $\{1, ..., p\}$ correspondant aux indices des colonnes de X qui sont conservées. Pour un sous-modèle M (potentiellement vide), on note M^c son complémentaire dans $\{1, ..., p\}$ et $|M|$ son cardinal. Avec $m = |M| \geq 1$ et $M = \{j_1, ..., j_m\}$, on note, pour une matrice T de taille $l \times p$, $T[M]$ pour la matrice de taille $l \times m$ obtenus en ne conservant que les colonnes de T dont les indices sont dans M. Pour un vector v de taille $p \times 1$ on notera, par abus de notation, $v[M] = (v'[M])'$, i.e., $v[M] = (v_{j_1}, ..., v_{j_m})'$.

A l'aide des notations ci dessus, on peut définir, pour un sous-modèle M, l'estimateur par moindre carrés $\hat{\beta}_M$ correspondant, défini par

$$\hat{\beta}_M = (X'[M]X[M])^{-1} X'[M]Y, \tag{2}$$

avec la convention $\hat{\beta}_\varnothing = 0$.

Pour un sous-modèle M fixé, $\hat{\beta}_M$ est en fait un estimateur sans biais de

$$\beta_M^{(n)} = \beta[M] + (X'[M]X[M])^{-1} X'[M]X[M^c]\beta[M^c], \tag{3}$$

avec la convention $\beta_M^{(n)} = \beta$ pour $M = \{1, ..., p\}$ et $\beta_M^{(n)} = 0$ pour $M = \varnothing$.

Nous appelons les coefficients de $\beta_M^{(n)}$ les coefficients de régression non-standards, dans le modèle M. Ces coefficients permettent au sous-modèle M de fournir une représentation optimale du modèle linéaire complet (1), dans le sens que (pour $M \neq \varnothing$)

$$\beta_M^{(n)} = \underset{v \in \mathbb{R}^{|M|}}{\operatorname{argmin}} ||X\beta - X[M]v||^2. \tag{4}$$

On considère désormais une procédure de sélection de modèle $\hat{M}$ qui associe à chaque (X, Y) un sous-modèle $\hat{M} = \hat{M}(X, Y) \subseteq \{1, ..., p\}$. Alors, motivé par (4), [2] définissent

le vecteur aléatoire $\beta^{(n)}_{\hat{M}}$ en remplaçant M par la variable aléatoire $\hat{M}$ dans (3). Puis, ils construisent les intervalles de confiance dits PoSI (Post-Selection Inference) pour les coefficients de $\beta^{(n)}_{\hat{M}}$. Dans cet article, nous adoptons un point de vue différent, et nous nous concentrons sur la construction d'intervalles de confiance pour des prédicteurs linéaires post-sélection de modèle. Les sections 2 et 3 suivantes présentent donc deux prédicteurs d'intérêt et des intervalles de confiance associés. La section 4 présente leurs performances pratiques dans une étude de simulation. Ces trois sections sont issues du manuscrit [1], auquel le lecteur peut se référer pour davantage de détail.

2 Intervalles de confiance pour le prédicteur d'intérêt dépendant des données

On considère $x_0 \in \mathbb{R}^p$ fixé et on suppose vouloir prédire $y_0 \sim N(x_0'\beta, \sigma^2)$, indépendamment de Y. Alors, le prédicteur $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$ est le pendant naturel de $\beta^{(n)}_{\hat{M}}$. Nous appelons ce prédicteur le prédicteur d'intérêt dépendant des données, pour insister sur sa dépendance à X. Ce prédicteur est inconnu en pratique, par sa dépendance à β. Il possède la propriété d'intérêt suivante : lorsque x_0 est aléatoire, suit la distribution empirique obtenue par les n lignes de X, et est indépendant de Y, alors, pour y_0 suivant la loi donnée plus haut conditionnellement à x_0, nous avons

$$\mathbb{E}\left(\left[y_0 - x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}\right]^2\right) \leq \mathbb{E}\left(\left[y_0 - x_0'[\hat{M}]v(Y)\right]^2\right),$$

pour toute fonction $v(Y) \in \mathbb{R}^{|\hat{M}|}$.

Soit maintenant un niveau de confiance nominal $1-\alpha \in (0,1)$ fixé. On considère donc des intervalles de confiance pour $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$ de la forme

$$CI = x_0'[\hat{M}]\hat{\beta}_{\hat{M}} \pm K(x_0, \hat{M})||s_{\hat{M}}||\hat{\sigma}, \tag{5}$$

où $x_0'[\hat{M}]\hat{\beta}_{\hat{M}}$ est le prédicteur post-sélection de modèle de y_0 utilisé en pratique, où $\|\cdot\|$ est la norme Euclidienne, où

$$s_M' = x_0'[M]\left(X'[M]X[M]\right)^{-1}X'[M], \tag{6}$$

et où $K(x_0, M) = K(x_0, M, X, \alpha)$ est une constante positive, pouvant dépendre de x_0, M, X et α, mais pas de Y. On a utilisé la notation $a \pm b$ pour l'intervalle $[a-b, a+b]$ ($a \in \mathbb{R}$, $b \geq 0$). Nous considérons ces intervalles de confiance (et [2] en considèrent des similaires) car, pour M *fixé*, l'intervalle $x_0'[M]\hat{\beta}_M \pm q_{S,n-p,1-\alpha/2}||s_M||\hat{\sigma}$, avec $q_{S,r,1-\alpha/2}$ le quantile $(1-\alpha/2)$ de la loi de Student à $n-p$ degrés de liberté, couvre classiquement $x_0'[M]\beta^{(n)}_M$ avec probabilité $1-\alpha$, quelles que soit les valeurs des paramètres β et σ.

Ainsi, la constante $K_{naif} = q_{S,n-p,1-\alpha/2}$ donne l'intervalle de confiance dit naïf, qui est généralement trop court lorsque $\hat{M}$ est aléatoire [3].

Ainsi, nous proposons quatre constantes $K_1(x_0)$, $K_2(x_0[\hat{M}], \hat{M})$, $K_3(x_0[\hat{M}], \hat{M})$ et K_4, qui satisfont

$$K_{naif} \leq K_1 \leq K_2 \leq K_3 \leq K_4.$$

Nous ramenons le lecteur à [1] pour les définitions précises de $K_1, ..., K_4$ et les algorithmes d'évaluation associés. Pour résumer : K_1 est une extension directe de la constante PoSI de [2], K_4 correspond à une borne supérieure K_{univ} proposée dans une version antérieure de [2][1], et K_2 et K_3 sont spécifiques à cet article. Toutes ces constantes sont assez grandes, d'après la proposition suivante (voir [1]).

Proposition 2.1. *On considère une procédure de sélection de modèle $\hat{M}$ quelconque et $x_0 \in \mathbb{R}^p$ fixé arbitrairement. Alors les intervalles de confiance CI de la forme* (5)*, où $K(x_0, \hat{M})$ peut être remplacé par $K_1, ..., K_4$ satisfont*

$$\inf_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma} \left(x_0'[\hat{M}] \beta^{(n)}_{\hat{M}} \in CI \right) \geq 1 - \alpha. \tag{7}$$

Enfin, notons une différence importante entre K_1 et K_2, K_3, K_4. La constante K_1 est théoriquement préférable, car elle satisfait (7) tout en étant de valeur minimale. En revanche, elle dépend de x_0 dans son intégralité (voir [1]). Ainsi, elle n'est pas accessible dans la situation dans laquelle, après qu'un modèle $\hat{M}$ soit sélectionné, seul le vecteur correspondant $x_0[\hat{M}]$ est observé, pour des raisons de coût. Notons que qu'en effet, seul $x_0[\hat{M}]$ est nécessaire pour construire le prédicteur post-sélection de modèle $x_0'[\hat{M}]\hat{\beta}_{\hat{M}}$. Les constantes K_2, K_3, K_4 sont ainsi spécialement conçues pour cette situation, car elles ne dépendent pas de $x_0[\hat{M}^c]$ qui n'est pas accessible. Notons que cette situation motive particulièrement l'étude de prédicteurs non standards, tels que $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$, puisqu'alors le prédicteur classique $x_0'\beta$ ne pourrait pas être utilisé, même si β était connu.

3 Intervalles de confiance pour le prédicteur d'intérêt indépendant des données

Le prédicteur d'intérêt dépendant des données $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$, traité en section 2, a une limitation importante : il dépend de la matrice X, et ses propriétés d'optimalité aussi. Nous étudions donc dans cette section un autre prédicteur, dit prédicteur d'intérêt indépendant des données, qui résout ce problème.

Dans cette section, nous considérons X comme aléatoire, dont chaque ligne suit une distribution commune $\mathcal{L}$ sur $\mathbb{R}^p$, ayant une matrice des moments (non centrés) d'ordre

[1] qui est encore disponible sur `http://www-stat.wharton.upenn.edu/~lzhao/papers/MyPublication/24PoSI-submit.pdf`.

deux Σ, qui est finie et définie positive. Nous considérons de même que x_0' suit la loi $\mathcal{L}$, et que la distribution de Y, y_0, conditionnellement à X, x_0 est la même qu'en section 2. Ainsi, les résultat de cette dernière section restent valide dans cette section, conditionnellement à X et x_0.

Pour deux sous-ensembles non vides M_1 et M_2 de $\{1, ..., p\}$ et pour Q une matrice de taille $p \times p$, on note $Q[M_1, M_2]$ la matrice obtenue par Q en supprimant toutes les lignes i avec $i \notin M_1$ et toutes les colonnes j avec $j \notin M_2$.

Avec ces notations, nous définissons le prédicteur d'intérêt indépendant des données par

$$x_0[\hat{M}]'\beta_{\hat{M}}^{(\star)} = x_0[\hat{M}]'\beta[\hat{M}] + x_0[\hat{M}]'\left(\Sigma[\hat{M}, \hat{M}]\right)^{-1}\Sigma[\hat{M}, \hat{M}^c]\beta[\hat{M}^c],$$

avec la convention $x_0[M]'\beta_M^{(\star)} = 0$ pour $M = \varnothing$ et $x_0[M]'\beta_M^{(\star)} = \beta$ pour $M = \{1, ..., p\}$. Ce prédicteur est optimal post-sélection de modèle dans le sens que, lorsque $x_0 \sim \mathcal{L}$,

$$\mathbb{E}\left(\left[y_0 - x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}\right]^2\right) \leq \mathbb{E}\left(\left[y_0 - x_0'[\hat{M}]v(Y)\right]^2\right),$$

pour toute fonction $v(Y) \in \mathbb{R}^{|\hat{M}|}$. Ainsi, nous voyons l'intérêt de $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$ par rapport à $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$: le second prédicteur minimise un critère qui dépend de la matrice X, n'apportant ainsi pas nécessairement d'information pour la prédiction pour de nouveaux x_0, tandis que le premier prédicteur minimise justement un critère défini en fonction de la distribution des nouveaux x_0.

Nous utilisons les intervalles de confiance de la section 2 pour couvrir $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$. La propriété (7) n'est alors plus exactement garantie, mais reste asymptotiquement valide, pour p fixé et $n \to \infty$.

Theorem 3.1. *Sous des hypothèses peu restrictives sur la loi $\mathcal{L}$ et la procédure de sélection de modèle $\hat{M}$, les intervalles de confiance CI de la section 2 satisfont, pour tout x_0 fixé,*

$$\inf_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}\left(x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)} \in CI \middle| X\right) \geq (1 - \alpha) + o_p(1), \tag{8}$$

lorsque $n \to \infty$.

4 Etude de simulation

Nous évaluons par Monte Carlo, pour les intervalles de confiance CI obtenus par les constantes $K_{naif}, K_1, K_2, K_3, K_4$, les probabilités de couverture minimales pour les deux prédicteurs optimaux des sections 2 et 3,

$$\inf_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}\left(x_0'[\hat{M}]\beta_{\hat{M}}^{(n,\star)} \in CI \middle| X\right),$$

où x_0 et X sont fixés après avoir été échantillonnés selon une loi $\mathcal{L}$ Gaussienne, dont la matrice Σ est explicitée dans [1] (cas Watershed).

On considère $\alpha = 0.05$, $p = 10$, $n = 20$ et $n = 100$ et les procédures de sélection de modèle AIC, BIC et LASSO. Les estimations des probabilités de couverture minimales sont reportées dans le tableau suivant.

n	Sélection de modèle	Prédicteur d'intérêt							
		dépendant des données $x_0[\hat{M}]'\beta_{\hat{M}}^{(n)}$				indépendant des données $x_0[\hat{M}]'\beta_{\hat{M}}^{(\star)}$			
		K_{naif}	K_1	K_3	K_4	K_{naif}	K_1	K_3	K_4
20	AIC	0.84	0.99	1.00	1.00	0.79	0.97	0.99	0.99
20	BIC	0.84	0.99	1.00	1.00	0.74	0.96	0.98	0.98
20	LASSO	0.90	1.00	1.00	1.00	0.18	0.48	0.61	0.61
100	AIC	0.87	0.99	1.00	1.00	0.88	0.99	1.00	1.00
100	BIC	0.88	0.99	1.00	1.00	0.87	0.99	1.00	1.00
100	LASSO	0.88	0.99	1.00	1.00	0.87	0.99	1.00	1.00

Ainsi, le passage de $n = 20$ à $n = 100$ rend les intervalles de confiance obtenus par $K_1, .., K_4$ suffisamment grands, et leur confère une probabilité de couverture toujours supérieure au niveau nominal 0.95, ce qui illustre le théorème 3.1. Dans le cas $n = 20$, la probabilité de contenir $x_0[\hat{M}]'\beta_{\hat{M}}^{(n)}$ est toujours supérieure à 0.95 pour $K_1, ..., K_4$, comme l'indique la proposition 2.1. Pour $x_0[\hat{M}]'\beta_{\hat{M}}^{(\star)}$, c'est également le cas pour les procédures de sélection de modèle AIC et BIC, mais pas pour le LASSO. En effet, comme expliqué dans [1], la procédure LASSO que nous utilisons estime des modèles contenant particulièrement peu de variables. Enfin, nous voyons également que les intervalles de confiance naifs, obtenus par K_{naif}, sont trop courts : leurs probabilités de couverture minimales sont toujours inférieures à 0.95.

Bibliographie

[1] F. Bachoc, H. Leeb et B. Pötscher (2014), *Valid confidence intervals for post-model-selection predictors*, soumis, `http://arxiv.org/abs/1412.4605`.

[2] R. Berk, L. Brown, A. Buja, K. Zhang et L. Zhao (2013), *Valid post-selection inference*, Annals of Statistics (41) 802-837.

[3] H. Leeb, B. M. Pötscher et K. Ewald (2013). *On various confidence intervals post-model- selection.* Statistical Science, à venir.