Intervalles de confiance uniformément valides en présence de sélection de modèle

François Bachoc [1], David Preinerstorfer [2] & Lukas Steinberger [3]

[1] *Institut de Mathématiques de Toulouse, Université Paul Sabatier, Francois.bachoc@math.univ-toulouse.fr*
[2] *CREATES, Aarhus University, david.preinerstorfer@econ.au.dk*
[3] *Department of Statistics and Operations Research, University of Vienna, lukas.steinberger@univie.ac.at*

Résumé. Nous proposons des méthodes générales permettant de contruire des intervalles de confiance post-sélection de modèle qui sont asymptotiquement valides. Les constructions sont basées sur des principes proposés récemment par Berk et al. (2013). En particulier, les modèles candidats utilisés peuvent être mal spécifiés, la quantité d'intérêt est spécifique au modèle sélectionné, et la couverture est garantie pour toute procédure de sélection de modèle. Dans un premier temps, nous développons une théorie générale. Dans un second temps, nous appliquons cette théorie générale aux situations pratiques importantes où les modèles considérés sont des modèles linéaires, homoscédastiques ou hétéroscédastiques, ou des modèles de régression binaire avec des fonctions de lien générales.

Mots-clés. Sélection de modèle, intervalles de confiance asymptotiquement valides, régression linéaire, régression binaire.

Abstract. We suggest general methods to construct asymptotically uniformly valid confidence intervals post-model-selection. The constructions are based on principles recently proposed by Berk et al. (2013). In particular the candidate models used can be misspecified, the target of inference is model-specific, and coverage is guaranteed for any data-driven model selection procedure. After developing a general theory we apply our methods to practically important situations where the candidate set of models, from which a working model is selected, consists of fixed design homoskedastic or heteroskedastic linear models, or of binary regression models with general link functions.

Keywords. Model selection, asymptotically valid confidence intervals, linear regression, binary regression.

1 Contexte

Considérons une situation dans laquelle nous observons n vecteurs aléatoires de taille $1 \times \ell$, $y_{1,n}, \ldots, y_{n,n}$, indépendants et définis sur un espace de probabilité commun $(\Omega, \mathcal{A}, \mathbb{P})$.

On notera la distribution de $y_n = (y'_{1,n}, \ldots, y'_{n,n})'$, sur les ensembles Boréliens de $\mathbb{R}^{n\times\ell}$, par $\mathbb{P}_n = \bigotimes_{i=1}^n \mathbb{P}_{i,n}$, où $\mathbb{P}_{i,n}$ est la distribution de $y_{i,n}$. Nous autorisons $\{\mathbb{P}_{i,n}\}$ à être un tableau triangulaire.

On considère maintenant un ensemble M_n, composé de d modèles, c'est à dire d ensembles non vides de distributions $\mathbb{M}_{1,n}, \ldots, \mathbb{M}_{d,n}$ sur les ensembles Boréliens de $\mathbb{R}^{n\times\ell}$, où d ne dépend pas de n. Nous ne *supposons pas* que la distribution $\mathbb{P}_n$ des observations est contenu dans l'un de ces ensembles ; c'est à dire que l'ensemble des modèles M_n peut-être *mal spécifié*. Nous définissons pour chaque modèle $\mathbb{M} \in \mathsf{M}_n$, une quantité d'intérêt $\theta^*_{\mathbb{M},n} = \theta^*_{\mathbb{M},n}(\mathbb{P}_n)$, que nous supposerons fixée et de dimension finie $m(\mathbb{M})$, où $m(\mathbb{M})$ ne dépend pas de n.

Nous renvoyons aux sections 3 et 4 pour des exemples spécifiques d'ensembles de modèles M_n. Typiquement, $\mathbb{M} \in \mathsf{M}_n$ est un ensemble de distributions paramétré par $\theta \in \mathbb{R}^{m(\mathbb{M})}$, et $\theta^*_{\mathbb{M},n}$ est le paramètre qui correspond à la projection de $\mathbb{P}_n$ sur $\mathbb{M}$, pour une certaine distance. On insiste donc sur le fait que la quantité d'intérêt $\theta^*_{\mathbb{M},n}$ est spécifique au modèle $\mathbb{M}$.

On suppose ensuite disposer pour tout $\mathbb{M} \in \mathsf{M}_n$ d'un estimateur $\hat{\theta}_{\mathbb{M},n}$ de la quantité d'intérêt $\theta^*_{\mathbb{M},n}$. L'estimateur $\hat{\theta}_{\mathbb{M},n}$ est une fonction mesurable de $\mathbb{R}^{n\times\ell}$ dans $\mathbb{R}^{m(\mathbb{M})}$. De plus, nous considérons une procédure de sélection de modèle, c'est à dire une fonction mesurable $\hat{\mathbb{M}}_n : \mathbb{R}^{n\times\ell} \to \mathsf{M}_n$. C'est donc finalement pour la quantité d'intérêt $\theta^*_{\hat{\mathbb{M}}_n,n}$ (qui est aléatoire) que nous souhaitons construire des intervalles de confiance.

La situation considérée ici, dans laquelle un modèle est sélectionné parmi un ensemble de modèles candidats, et dans laquelle on construit des intervalles de confiance postérieurement à la sélection du modèle, est appelée *inférence post-sélection de modèle*. L'inférence post-sélection de modèle fait actuellement l'objet de nombreux travaux, parmi lesquels [1-8]. Enfin, cet article est issu du manuscrit [3], accessible en ligne, et pour lequel nous renvoyons le lecteur pour davantage de contenu et de détails, et pour les preuves des résultats énoncés ici.

2 Théorie générale pour la construction d'intervalles de confiance

On considère disposer, pour tout modèle $\mathbb{M} \in \mathsf{M}_n$, d'un estimateur $\hat{\theta}_{\mathbb{M},n}$ de la quantité d'intérêt $\theta^*_{\mathbb{M},n}$. Dans l'article [5], qui traite le cas de modèles de régression linéaire homoscédastique, les auteurs observent que, dans cette situation, le vecteur $\{\hat{\theta}_{\mathbb{M},n} - \theta^*_{\mathbb{M},n}\}_{\mathbb{M}\in\mathsf{M}_n}$ est un vecteur gaussien. En exploitant cela, ils utilisent une approche de type *pire cas* (pour le modèle sélectionné), et obtiennent un intervalle de confiance $\mathrm{CI}^{(j)}_{1-\alpha,\hat{\mathbb{M}}_n}$, qui est fonction du modèle sélectionné, pour le composant j de $\theta^*_{\hat{\mathbb{M}}_n,n}$, qui vérifie

$$\mathbb{P}_n\left(\theta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j)}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ pour tout } j = 1, \ldots, m(\hat{\mathbb{M}}_n)\right) \geq 1 - \alpha. \tag{1}$$

La relation ci-dessus est vraie pour n'importe quelle procédure de sélection de modèle $\hat{\mathbb{M}}_n$, ce qui amène les auteurs de [5] à parler d'intervalles de confiance universellement valides. Cette universalité est particulièrement intéressante lorsque le statisticien a peu de contrôle sur la procédure de sélection de modèle, par exemple lorque le modèle est choisi par le praticien, de manière informelle, ou en prenant en compte des critères économiques, de type coûts d'observations de variables.

Le principe de cette section est d'utiliser une normalité asymptotique de $\hat{\theta}_{\mathbb{M},n} - \theta^*_{\mathbb{M},n}$, pour tout modèle $\mathbb{M} \in \mathsf{M}_n$ fixé, afin d'utiliser la construction de [5] avec plus de généralité, et de montrer une version asymptotique de (1). Nous notons donc $\hat{\theta}_n = (\hat{\theta}'_{\mathbb{M}_1,n}, \ldots, \hat{\theta}'_{\mathbb{M}_d,n})'$ et $\theta^*_n = (\theta^{*'}_{\mathbb{M}_1,n}, \ldots, \theta^{*'}_{\mathbb{M}_d,n})'$. De plus, on écris $k = \sum_{j=1}^d m(\mathbb{M}_{j,n})$, pour la dimension de $\hat{\theta}_n$ (cette dimension de dépend pas de n). On écris aussi $\mathbb{E}_n$, $\mathbb{V}_n$, and $\mathbb{VC}_n$, pour l'espérance, la variance, et la matrice de covariance selon la loi $\mathbb{P}_n$. On défini de même $\mathbb{E}_{i,n}$, $\mathbb{V}_{i,n}$, and $\mathbb{VC}_{i,n}$ pour la loi $\mathbb{P}_{i,n}$ et $\mathbb{E}$, $\mathbb{V}$, and $\mathbb{VC}$ pour la loi $\mathbb{P}$. On fait alors l'hypothèse suivante.

Condition 1 *Il existe des fonctions Borel-mesurables $g_{i,n} : \mathbb{R}^{1\times\ell} \to \mathbb{R}^k$ pour $i = 1, \ldots, n$, dépendant possiblement de θ^*_n, telles que*

$$\hat{\theta}_n(y_n) - \theta^*_n \quad = \quad \sum_{i=1}^n g_{i,n}(y_{i,n}) + \Delta_n(y_n),$$

où, en écrivant $r_n(y_n) := \sum_{i=1}^n g_{i,n}(y_{i,n})$, on a pour tout $i \in \{1, \ldots, n\}$ et tout $j \in \{1, \ldots, k\}$ que

$$\mathbb{E}_{i,n}\left(g^{(j)}_{i,n}\right) = 0 \quad \text{and} \quad 0 < \mathbb{V}_n\left(r^{(j)}_n\right) < \infty,$$

et pour toute coordonnée $j \in \{1, \ldots, k\}$ on a, avec $\{.\}$ la fonction indcatrice,

$$\mathbb{V}_n^{-1}\left(r^{(j)}_n\right) \sum_{i=1}^n \int_{\mathbb{R}^{1\times\ell}} \left[g^{(j)}_{i,n}\right]^2 \left\{|g^{(j)}_{i,n}| \geq \varepsilon \mathbb{V}_n^{\frac{1}{2}}(r^{(j)}_n)\right\} d\mathbb{P}_{i,n} \to 0 \text{ pour tout } \varepsilon > 0,$$

et

$$\mathbb{P}_n\left(\left|\mathbb{V}_n^{-1/2}\left(r^{(j)}_n\right)\Delta^{(j)}_n\right| \geq \varepsilon\right) \to 0 \text{ pour tout } \varepsilon > 0.$$

On pose

$$S_n(y_n) := \sum_{i=1}^n g_{i,n}(y_{i,n}) g'_{i,n}(y_{i,n}).$$

On note $A^\dagger$ l'inverse de Moore-Penrose d'une matrice carré A, on note $A^{1/2}$ la racine carré d'une matrice semi-définie positive A, et on écris $A^{\dagger/2}$ pour $[A^\dagger]^{1/2}$. On écris d_w pour une distance qui génére la topologie de la convergence faible sur l'espace des distributions de probabilités sur un espace euclidien. On note $N(\mu, \Sigma)$ pour la loi normale de vecteur moyenne μ et de matrice de covariance Σ. On écris aussi $\text{corr}(\Sigma) = \text{diag}(\Sigma)^{\dagger/2}\Sigma\,\text{diag}(\Sigma)^{\dagger/2}$, où $\text{diag}(\Sigma)$ est la matrice diagonale obtenue en remplaçant les composants hors-diagonale de Σ par 0. Le lemme suivant donne le résultat de normalité asymptotique que nous utilisons pour construire nos intervalles de confiance.

Lemma 2.1 *Sous la condition 1, pour $\varepsilon > 0$ on a, avec $\mathbb{P}_n \circ f$ la mesure image d'une fonction f sous $\mathbb{P}_n$,*

$$\mathbb{P}_n \left(d_w \left(\mathbb{P}_n \circ \left[\mathrm{diag}(S_n)^{\dagger/2} \left(\hat{\theta}_n - \theta_n^* \right) \right], N(0, \mathrm{corr}(S_n)) \right) \geq \varepsilon \right) \to 0,$$

et l'énoncé reste vrai si S_n est remplacé par $\mathbb{VC}_n(r_n)$.

Écrivons maintenant, pour $\alpha \in (0,1)$ et pour une matrice de covariance Γ, $K_{1-\alpha}(\Gamma)$ pour le quantile $1-\alpha$ de $\|Z\|_\infty$ pour $Z \sim N(0,\Gamma)$. Pour $\mathbb{M} = \mathbb{M}_{i,n} \in \mathsf{M}_n$ et $j \in \{1, \ldots, m(\mathbb{M})\}$ on note

$$j \star \mathbb{M} \quad := \quad \sum_{l=1}^{i-1} m(\mathbb{M}_{l,n}) + j,$$

où les sommes sur un ensemble vide d'indices valent 0. Par exemple, $\theta_n^{*(j \star \mathbb{M})} = \theta_{\mathbb{M},n}^{*\,(j)}$.

La proposition suivante donne une construction d'intervalles de confiance basée sur l'existence d'estimateurs consistants de la matrice de covariance $\mathbb{VC}_n(r_n)$.

Theorem 2.2 *Soit $\alpha \in (0,1)$ et supposons que la Condition 1 soit vraie. Soit $\hat{S}_n : \mathbb{R}^{n \times \ell} \to \mathbb{R}^{k \times k}$ une suite de fonctions Borel-mesurables telles que pour tout $\varepsilon > 0$, avec $||.||$ la plus grande valeur singulière de A,*

$$\mathbb{P}_n \left(\| \mathrm{corr}(\hat{S}_n) - \mathrm{corr}\left(\mathbb{VC}_n(r_n) \right) \| + \| \mathrm{diag}(\mathbb{VC}_n(r_n))^{-1} \mathrm{diag}(\hat{S}_n) - I_k \| \geq \varepsilon \right)$$

tende vers 0. Posons pour $\mathbb{M} \in \mathsf{M}_n$ et $j = 1, \ldots, m(\mathbb{M})$ l'intervalle de confiance

$$\mathrm{CI}_{1-\alpha,\mathbb{M}}^{(j),\mathrm{est}} \quad = \quad \hat{\theta}_{\mathbb{M},n}^{(j)} \pm \sqrt{[\hat{S}_n]_{j \star \mathbb{M}}}\; K_{1-\alpha} \left(\mathrm{corr}(\hat{S}_n) \right).$$

Alors, $\mathbb{P}_n \left(\theta_{\mathbb{M},n}^{*(j)} \in \mathrm{CI}_{1-\alpha,\mathbb{M}}^{(j),\mathrm{est}} \text{ pour tout } \mathbb{M} \in \mathsf{M}_n \text{ et } j = 1, \ldots, m(\mathbb{M}) \right)$ tend vers $1-\alpha$ lorsque $n \to \infty$. En particulier, pour toute procédure de sélection de modèle $\hat{\mathbb{M}}_n$, on a

$$\liminf_{n \to \infty} \mathbb{P}_n \left(\theta_{\hat{\mathbb{M}}_n,n}^{*(j)} \in \mathrm{CI}_{1-\alpha,\hat{\mathbb{M}}_n}^{(j),\mathrm{est}} \text{ pour tout } j = 1, \ldots, m(\hat{\mathbb{M}}_n) \right) \geq 1 - \alpha.$$

Le théorème précédent donne des intervalles de confiance qui sont, dans un certain sens, de taille minimale, puisqu'ils ont une probabilité exactement $1-\alpha$ de contenir simultanément toutes les quantités d'intérêts $\theta_{\mathbb{M},n}^{*(j)}$ pour tout $\mathbb{M} \in \mathsf{M}_n$ et $j = 1, \ldots, m(\mathbb{M})$. Ces intervalles de confiance nécessitent d'avoir un estimateur consistent de $\mathbb{VC}_n(r_n)$. Cette hypothèse peut-être trop contraignante lorsque l'ensemble de modèles M_n est mal spécifié, car $\mathbb{VC}_n(r_n)$ dépend de la loi inconnue des observations $\mathbb{P}_n$. Nous renvoyons alors à [3] pour des méthodes générales de construction d'intervalles de confiance plus conservatifs, et plus largement employables. Le principe de la construction de ces intervalles de confiance est de construire des estimateurs conservatifs pour les éléments diagonaux de $\mathbb{VC}_n(r_n)$, et d'utiliser des majorants calculables de $K_{1-\alpha}\left(\mathrm{corr}(\mathbb{VC}_n(r_n))\right)$.

3 Application à la régression linéaire

Dans cette section et la suivante, nous énonçons des résultats uniformes, par rapport à un ensemble $\mathbf{P}_n$ de lois potentielles $\mathbb{P}_n$ des observations. Pour T un sous ensemble Borélien de $\mathbb{R}$, on note $M(T^n)$ l'ensemble des mesures de probabilité sur $T^n = \bigtimes_{i=1}^n T \subseteq \mathbb{R}^n$. Le vecteur moyenne d'un élément Q de $M(T^n)$ est noté $\mu(Q)$. Pour un élément $Q \in M(T^1)$ et pour $0 < q < \infty$, on écris $m_q(Q)$ pour le q-ème moment absolu centré de Q, en supposant que $\mu(Q)$ existe. On écris $\bigotimes_{i=1}^n M(T)$ pour l'ensemble des mesures produits sur $M(T^n)$. Pour $Q \in \bigotimes_{i=1}^n M(T)$ on écris $Q = \bigotimes_{i=1}^n Q_i$.

On considère un ensemble de distributions potentielles $\mathbb{P}_n$, qui dépend de deux paramètres $\delta > 0$ et $\tau > 1$ et qui est défini par

$$\mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau) := \left\{ Q \in \bigotimes_{i=1}^n M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_1) = \ldots = m_2(Q_n) < \infty \\ \frac{\max_{i=1,\ldots,n} m_{2+\delta}(Q_i)^{\frac{2}{2+\delta}}}{m_2(Q_1)} \leq \tau \end{array} \right\}.$$

Nous nous intéressons alors aux modèles linéaires homoscédastiques. Dans ces modèles, on suppose que le vecteur moyenne $\mu(\mathbb{P}_n)$ appartient à $\mathrm{span}(X_n)$, l'espace vectoriel engendré par les colonnes d'une matrice $X_n \in \mathbb{R}^{n \times p}$, avec p fixé et ne dépendant pas de n. La matrice X_n est fixée et connue. On suppose aussi qu'il est *connu* que les observations ont la même variance. L'ensemble des modèles $\mathcal{I} = \{M_1, \ldots, M_d\}$ est alors un ensemble de sous ensembles non vides de $\{1, 2, \ldots, p\}$. Pour $M \in \mathcal{I}$, on note $X_n[M]$ la matrice obtenue en supprimant toutes les colonnes de X_n qui n'appartiennent pas à M. On considère alors pour tout $j \in \{1, \ldots, d\}$ les ensembles

$$\mathbb{M}_{j,n} = \left\{ Q \in \bigotimes_{i=1}^n M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_1) = \ldots = m_2(Q_n) < \infty \\ \mu(Q) \in \mathrm{span}(X_n[M_j]) \end{array} \right\},$$

et notre ensemble de modèles est donné par $\mathsf{M}_n = \{\mathbb{M}_{j,n} : j = 1, \ldots, d\}$.

La quantité d'intérêt est donnée par, pour $\mathbb{M} \in \mathsf{M}_n$ avec un ensemble d'indices M,

$$\beta^*_{\mathbb{M},n} = \beta^*_{\mathbb{M},n}(\mathbb{P}_n) = (X_n[M]' X_n[M])^{-1} X_n[M]' \mu(\mathbb{P}_n),$$

c'est à dire que $\beta^*_{\mathbb{M},n}$ est le vecteur de coefficients de la projection orthogonale de $\mu(\mathbb{P}_n)$ sur $\mathrm{span}(X_n[M])$.

En appliquant les méthodes générales données dans la Section [2], et en définissant un estimateur conservatif de la variance des observations, nous pouvons construire des intervalles de confiance qui fournissent les mêmes garanties que dans le Théorème 2.2.

Dans le cas de la régression hétéroscédastiques, le principe est le même, avec l'ensemble de distributions potentielles

$$\mathbf{P}_n^{(\mathrm{het})}(\delta, \tau) := \left\{ Q \in \bigotimes_{i=1}^n M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_i) < \infty \text{ for } i = 1, \ldots, n \\ \frac{\max_{i=1,\ldots,n} m_{2+\delta}(Q_i)^{\frac{2}{2+\delta}}}{\min_{i=1,\ldots,n} m_2(Q_i)} \leq \tau \end{array} \right\}$$

et l'ensemble de modèles $\mathsf{M}_n = \{\mathbb{M}_{j,n} : j = 1, \ldots, d\}$ est donné par

$$\mathbb{M}_{j,n} = \left\{ Q \in \bigotimes_{i=1}^{n} M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_i) < \infty \text{ for } i = 1, \ldots, n \\ \mu(Q) \in \text{span}(X_n[M_j]) \end{array} \right\}.$$

4 Application à la régression binaire

En régression binaire, l'ensemble des distributions potentielles pour les observations dépend du paramètre $\tau > 0$ et est défini par

$$\mathbf{P}_n^{(\text{bin})}(\tau) := \left\{ Q \in \bigotimes_{i=1}^{n} M(\{0,1\}) : Q_i(\{0\})Q_i(\{1\}) \geq \tau \ \forall i = 1, \ldots, n \right\}.$$

On s'intéresse alors à un ensemble de modèles $\mathsf{M}_n = \left\{\mathbb{M}_{(j_1,j_2),n} : j_1 \in \{1, \ldots, d_1\}, j_2 \in \{1, \ldots, d_2\}\right\}$, défini par des fonctions de réponses $h_1, ..., h_{d_1} : \mathbb{R} \to [0,1]$ et des sous ensembles $M_1, ..., M_{d_2}$ de $\{1, ..., p\}$. Les modèles sont définis par, avec $X_{i,n}$ la ligne i de X_n,

$$\mathbb{M}_{(j_1,j_2),n} = \left\{ Q \in \bigotimes_{i=1}^{n} M(\{0,1\}) : \begin{array}{l} \exists \beta \in \mathbb{R}^{|M_{j_2}|} : \forall i = 1, \ldots, n : \\ Q_i(\{1\}) = h_{j_1}(X_{i,n}[M_{j_2}]\beta) \end{array} \right\}.$$

On peut alors obtenir des résultats de couverture asymptotique identiques à ceux du théorème 2.2. Notons que dans le cas ou toutes les fonctions de lien sont *canoniques*, les intervalles de confiance peuvent être significativement plus courts (à garantie de couverture égale). Nous renvoyons à [3] pour plus de détails.

Bibliographie

[1] Bachoc, F., Ehler, M. et Gräf, M. (2017), Optimal configurations of lines and a statistical application, Advances in Computational Mathematics, 43(1), 113-126.

[2] Bachoc, F., Leeb, H. et Pötscher, B. M. (2014), Valid confidence intervals for post-model-selection predictors, arXiv :1412.4605.

[3] Bachoc, F., Preinerstorfer, D. et Steinberger, L. M. (2016), Uniformly valid confidence intervals post-model-selection, arXiv :1611.01043.

[4] Leeb, H. and Pötscher, B. M. (2005), Model selection and inference : Facts and fiction, Econometric Theory, 21, 21–59.

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

[6] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), Exact post-selection inference, with application to the lasso., Annals of Statistics, 44, 907–927.

[7] Van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014), On asymptotically optimal confidence regions and tests for high-dimensional models, Annals of Statistics, 42, 1166–1202.