Une méthode de sélection de prédicteurs sous contrainte de non multicolinéarité dans les modèles linéaires généralisés

Christian Derquenne

Chercheur indépendant - chris.emr@wanadoo.fr

Résumé. Cet article se place dans le cadre de la multicolinéarité entre les prédicteurs au sein des modèles linéaires généralisés. Le phénomène de multicolinéarité peut entrainer des incohérences sur les coefficients de régression et des oublis de prédicteurs, cela peut par conséquent poser des problèmes d'interprétation qui peuvent entrainer de mauvaises décision. Le critère proposé est une nouvelle méthode de sélection de variables dans le cadre des modèles linéaires généralisés qui permet de respecter la non-multicolinéarité entre les prédicteurs. Ce critère est constitué de plusieurs statistiques : l'indépendance globale et marginale pour tester la non-multicolinéarité, l'ajustement global du modèle aux données, les effets marginaux des coefficients de régression multiple, la cohérence des signes de ceux-ci avec les coefficients de régression simple. Le modèle sélectionné possède deux propriétés : interprétabilité et prévision.

Mots-clés. Modèles linéaires généralisés, Multicolinéarité, méthodes de sélection de prédicteurs

Abstract. This article is placed within the framework of the multicollinearity between the predictors within a Generalized linear models The phenomenon of multicollinearity can lead to inconsistencies in the regression coefficients and omissions of predictors this can therefore pose problems of interpretation which can lead to poor decisions. The proposed criterion is a new method for selecting variables within the framework of generalized linear models which makes it possible to respect non-multicollinearity between the predictors. This criterion is made up of several statistics: the overall and marginal independence to test non-multicollinearity, the overall fit of the model to the data, the marginal effects of the multiple regression coefficients, the consistency of the signs of these with the coefficients simple regression. The selected model has two properties: interpretability and prediction.

Keywords. Generalized Linear model, Multicollinearity, stepwise methods

1 Contexte - objectif

La qualité des résultats d'un modèle statistique est primordiale pour les experts métier dans leurs domaines d'applications. Par exemple, en management de l'énergie, majoritairement l'objectif des modèles mis en oeuvre est d'obtenir la meilleure prévision de consommation d'électricité offrant les plus faibles erreurs (RMSE, MAPE, ...) afin de fournir des résultats fiables à destination de la production et donc de garantir l'équilibre offre/demande. Le modèle est très performant mais cela peut être au détriment de la compréhension des résultats en termes d'interprétabilité du sens et de l'influence des prédicteurs par les commanditaires de l'étude. Dans d'autres cas, l'objectif peut être justement de construire des modèles dans les résultats

sont interprétables et explicables. Et dans ces conditions, l'aspect qualité de la prévision n'est plus obligatoirement recherché. L'explicativité/interprétabilité des modèles, ou du moins de leurs résultats, devient par conséquent une discipline à part entière. Certaines approches statistiques sont plutôt dédiées à la prévision (réseaux de neurones, par exemple) et pour lesquelles il est nécessaire a posteriori de faire appel à des indicateurs numériques (LIME, valeurs de Shapley, cartes de salience, ...) si l'on désire interpréter les résultats. A l'opposé, les approches focalisées sur l'interprétabilité construisent des modèles permettant d'exploiter immédiatement le sens et l'importance des prédicteurs à l'aide par exemple des méthodes de régularisation, telles que les régressions Ridge, Lars [Hoerl et al., 1970], Lasso [Tibshirani et al., 1996] ou Elastic-Net [Zou et al., 2005] qui répondent en partie à la question de prédicteurs liés entre eux, mais elles exigent des paramètres non analytiques, donc difficiles à évaluer a priori. Le recours à la validation croisée est donc nécessaire. La régression sur composantes principales, qui consiste à réaliser une ACP sur toutes les variables explicatives, permet de garder au moyen de différentes stratégies certaines composantes principales qui résument au mieux les prédicteurs initiaux. Cependant, cette méthode n'est pas toujours optimale à l'égard du sens des coefficients et de l'importance des prédicteurs. La régression PLS (Partial Least Squares) [Wold H. et al., 1984] permet de mieux tenir compte de ces deux propriétés. En effet, cette méthode repose sur le principe de l'algorithme NIPALS (Non linear Iterative Partial Least Squares) qui est très puissant en termes de robustesse. Enfin, un premier critère nommé MCC (MultiCollinearity Criterium) a été proposé en 2022 [Derquenne, 2022] pour prendre en compte la multicolinéarité, dès le départ de la construction d'un modèle linéaire gaussien de régression multiple. Le critère MCC permet de maximixer les trois propriétés suivantes : reconstitution de l'inertie expliquée de l'estimateur MCO, adéquation des signes et de l'ordre des coefficients de régression multiple par rapport aux coefficients de corrélation linéaire simple associés. Une quatrième propriété a été introduite telle que la correspondance de la significativité des paramètres de régression multiple et simple. Un second critère nommé MCGC (MultiCollinearity Generalized Criterium) a permis d'étendre cette approche aux modèles linéaires généralisés [Derquenne, 2023]. Les trois premières propriétés à respecter dans le critère MCC sont adaptées au type de modèle, par exemple en régression logistique booléenne, une quatrième est rajoutée, telle que la proportion de bien classés.

Par ailleurs, la propriété de parcimonie d'un modèle est également très recherchée que ce soit dans un objectif de prévision ou dans un objectif d'interprétabilité. Le rapport qualité/prix de ces modèles est mesuré à l'aide de nombreux indicateurs, par exemple, les critères d'information AIC ou BIC. Les méthodes de sélection de prédicteurs permettent de garder les variables les plus influentes significativement dans le modèle, mais pas toujours de façon efficace face à la multicolinéarité qui met en défaut l'interprétabilité des résultats. En effet, comme déjà indiqué, les signes des coefficients de régression multiple peuvent être opposés à ceux des coefficients de corrélation linéaire simple, ou encore certains qui étaient fortement liés à la réponse peuvent être éliminés au détriment d'autres prédicteurs qui avaient à l'origine peu d'influence. Le modèle choisi peut alors être inintéressant sous l'aspect métier, voire incohérent et même dangereux à appliquer.

Le statisticien est alors face à un choix cornélien : soit construire un modèle statistique interprétable en éliminant la multicolinéarité quitte à garder des prédicteurs non significatifs ce

qui peut mettre en défaut la qualité d'ajustement sachant le prix payé par le nombre de variables explicatives, soit obtenir un modèle statistique parcimonieux et puissant avec des prédicteurs significatifs mais pénalisé par la multicolinéarité, donc inefficace en termes d'interprétabilité. Ce choix n'est heureusement pas binaire, en effet une solution potentielle serait un modèle puissant mais sans multicolinéarité. Dans ces conditions les résultats issus de ce modèle de régression multiple, notamment le sens de l'influence des prédicteurs et de leur importance pourraient être interprétés "en toute sécurité" car l'espace des variables candidates à l'explication sera structuré en séparant l'effet mutuel de chacune d'elles.

Dans ce article, nous proposons un nouveau critère permettant de construire un tel modèle statistique. La section 2 introduit ce nouveau critère nommé MCSRC (MultiCollinearity Stepwise Regression Criterium), puis il est appliqué à un exemple simulé utilisé dans [Derquenne, 2022] et comparé avec d'autres méthodes de sélection de prédicteurs. Dand la section 3, nous concluons sur les apports et les faiblesses de l'approche proposée, et nous fournissons quelques voies futures en termes d'amélioration et de nouveaux développements.

2 Un critère de sélection de prédicteurs en régression tenant compte de la multicolinéarité (MCSRC)

Rappelons que la multicolinéarité entre les prédicteurs d'un modèle de régression multiple peut entrainer d'une part, des signes contraires des coefficients de régression par rapport aux corrélations linéaires simples et d'autre part, des résultats de tests marginaux des coefficients (test t) en contradiction avec ce qui aurait pu être attendu d'après les tests sur les coefficients de régression simple. Pour pallier ces problèmes des solutions ont été proposées, comme indiqué dans l'introduction.

2.1 Le cas du modèle linéaire gaussien

Illustrons ce problème sur un jeu de données simulé. Ses caractéristiques sont les suivantes : $X_1 \rightarrow \mathcal{N}(0,1)$, $X_2 = X_1 + \mathcal{N}(0,1.96)$, $X_3 = 2X_2 + 3 + \mathcal{N}(0,0.04)$, $X_4 \rightarrow \mathcal{N}(0,2.25)$, $X_5 \rightarrow \mathcal{N}(0,1)$, $X_6 = X_5 + \mathcal{N}(0,0.04)$ et $Y = -1.5X_1 + 2X_2 + 0.5X_3 - 0.5X_4 + 0.1X_5 - 0.1X_6 + 4 + \mathcal{N}(0,0.25)$. Nous avons appliqué, (i) le critère classique des MCO, (ii) la régression sur composantes principales avec une sélection pas à pas de celles-ci reposant sur des tests statistiques de nullité des coefficients entrant et sortant, (iii) la régression sur premières composantes principales (RFPC) fondées sur une étape préalable de classification des prédicteurs [Derquenne et al., 2002], (iv) la régression PLS, (v) la sélection pas à pas des prédicteurs initiaux à l'aide des p-valeurs entrante et sortante (Stepwise Regression), (vi), la sélection ascendante de prédicteurs (Forward Regression), (vii) la méthode Incremental Forward Stagewise Regression [Hastie et al., 2007], (viii) le critère Lars, (ix) le critère Lasso, (x) le critère ElasticNet et (xi) le critère MCC proposé dans [Derquenne, 2022].

La table 1 fournit les coefficients de régression standardisés pour chacune des méthodes, ainsi que les corrélations simples. Premièrement, seuls RFPC et le multi-critère MCC fournissent des signes cohérents pour les six prédicteurs. Les forces de liaison sont relativement bien respectées pour la sélection pas à pas par p-valeurs, pour la régression PLS, avec un petit avantage pour le critère MCC. Seuls RFPC et MCC sont en parfaite adéquation avec l'ordre des coefficients

de régression et des corrélations simples. Globalement, les R^2 ajustés pour la plupart des méthodes sont supérieurs à 0,96, sauf pour RFPC (=0,66) et le multi-critère (=0,84). Ce dernier résultat est logique car MCC n'optimise pas seulement le critère des MCO. Enfin, *Eval* représente l'évaluation [Derquenne, 2022] de chaque méthode pa rapport à MCC. Ce dernier obtient "logiquement" la plus grande valeur (*Eval*=0,96), puis viennent Lasso, Lars, RFPC, ElasticNet, la régression sur composantes principales et PLS ($Eval \in [0,83; 0,85]$), enfin, MCO et les trois méthodes de sélection de prédicteurs ($Eval \in [0,73; 0,77]$). Par conséquent dans cet exemple, seul le critère MCC parait efficace face à la multicolinéarité en termes de reconstitution de force et de sens des liaisons tout en préservant la qualité du modèle. Ce résultat a été corroboré sur de nombreuses applications réelles et des simulations.

	$\hat{\beta}^{MCO}$	$\hat{\beta}^{PCR_1}$	$\hat{\beta}^{RFPC}$	$\hat{\beta}^{PLS}$	$\hat{\beta}^{Spval}$	$\hat{\beta}^{Forw}$	$\hat{\beta}^{Swise}$	$\hat{\beta}^{Lars}$	$\hat{\beta}^{Lasso}$	$\hat{\beta}^{ENet}_{\lambda=0,5}$	$\hat{\beta}^{New}$	r_{yX}
X_1	-0.379	-0.381	0.246	-0.323	-0.383	-0.383	-0.380	-0.263	-0.263	-0.209	0.083	0.171
X_2	0.708	0.560	0.332	0.549	0.695	0.695	0.000	0.393	0.393	0.589	0.863	0.910
X_3	0.419	0.569	0.332	0.557	0.435	0.435	1.130	0.608	0.608	0.594	0.866	0.913
X_4	-0.195	-0.194	-0.204	-0.263	-0.194	-0.194	-0.189	-0.125	-0.125	-0.167	-0.192	-0.084
X_5	0.099	0.064	0.041	0.051	0.051	0.051	0.000	0.000	0.000	0.071	0.046	0.136
X_6	0.050	0.013	-0.041	0.021	0.000	0.000	0.000	0.000	0.000	0.000	-0.008	-0.083
R^2_{adj}	0.987	0.987	0.663	0.980	0.987	0.987	0.981	0.962	0.962	0.953	0.840	n.a
Eval	0.786	0.830	0.850	0.827	0.774	0.774	0.727	0.854	0.854	0.830	0.958	1.000

Table 1: Coefficients de régression des méthodes

Intéressons-nous plus précisément aux résultats obtenus par les trois méthodes de sélection de prédicteurs : ascendante, pas à pas, et stagewise. La méthode de sélection ascendante consiste à entrer dans le modèle le prédicteur qui possède la plus petite p-valeur du coefficient de régression à condition qu'elle soit inférieure à un seuil de première espèce α_e fixé par le statiscien, par exemple 0,05, puis un deuxième prédicteur est ajouté au premier avec la condition précédente. L'algorithme se poursuit tant que la $p - valeur \leq \alpha_e$. Le problème de cette méthode est la non remise en cause des prédicteurs déjà entrés dans le modèle. La méthode de sélection pas à pas débute de la même façon que la précédente, mais elle a besoin de deux seuils α's, l'un pour les préditeurs entrants (α_e) ; l'autre pour les prédicteurs sortants (α_s). En effet, lors de l'ajout d'une variable ($p - valeur \leq \alpha_e$), si la p-valeur d'une autre déjà entrée dans le modèle dépasse le seuil α_s, alors cette dernière est éliminée. L'algorithme continue tant que les deux seuils α's sont respectés. Enfin, la méthode stagewise ascendante crée un profil de coefficients comme suit : à chaque étape, il incrémente le coefficient de la variable la plus corrélée aux résidus courants d'une quantité $\pm\epsilon$, de signe déterminé par le signe de la corrélation. Efron et al. (2004) ont en fait considéré la version limitée de cet algorithme, avec $\epsilon \downarrow 0$, qui possède également des chemins de coefficients linéaires par morceau.

Comme nous pouvons le constater (cf. table 1) les méthodes de sélection ascendante et pas à pas au moyen du test de Student sur les coefficients fournissent les mêmes prédicteurs X_1, X_2, X_3, X_4, X_5, avec les p-valeurs respectives $< 0,0001$, 0,0002, 0,0154, $< 0,0001$ et $< 0,0001$, alors que l'approche stagewise sélectionne X_1, X_3, X_4 dont les trois p-valeurs sont toutes $< 0,0001$. Les

résultats montrent que les coefficients associés à X_1 sont négatifs pour les trois méthodes, alors que le coefficient de corrélation linéaire simple est positif. Sa p-valeur est égale à 0,0891, alors que le coefficient de régression multiple est très significatif. Comme indiqué, ce problème est typique de la multicolinéarité entre ces prédicteurs. Leur matrice de corrélations montre que certains sont fortement liés.

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.000 (na)	0.514 ($< 0,001$)	0.506 ($< 0,001$)	0.114 (0,261)	-0.031 (0.761)	-0.021 (0.826)
X_2	0.514 ($< 0,001$)	1.000 (na)	0.998 ($< 0,001$)	0.136 (0,178)	0.075 (0.460)	-0.044 (0.638)
X_3	0.506 ($< 0,001$)	0.998 ($< 0,001$)	1.000 (na)	0.131 (0.195)	0.069 (0.496)	-0.036 (0.692)
X_4	0.114 (0.261)	0.136 (0.178)	0.131 (0.195)	1.000 (NA)	0.047 (0.641)	-0.026 (0.755)
X_5	-0.031 (0.761)	0.075 (0.460)	0.069 (0.496)	0.047 (0.641)	1.000 (na)	-0.970 ($< 0,001$)
X_6	-0.021 (0.826)	-0.044 (0.638)	-0.036 (0.692)	-0.026 (0.755)	-0.970 ($< 0,001$)	1.000 (na)

Table 2: Matrice des coefficients de corrélation linéaire simple et p-valeurs associées

Le critère de sélection de prédicteurs proposé va tenir compte conjointement de la multicolinéarité, de la force de liaison entre les prédicteurs et la réponse, et de la cohérence des signes des coefficients de régression multiple avec les coefficients de corrélation simple.

2.1.1 Prise en compte de la multicolinéarité

Nous utilisons deux tests d'indépendance. Le premier est global ; le second est marginal.

Test global d'indépendance : Soient $X_1, ..., X_j, ..., X_p$, p variables gaussiennes de taille n et soit $\mathbf{R}$, la matrice de corrélations linéaires de Pearson associée. Soit le jeu d'hypothèses H_0 : Indépendance linéaire entre les variables de $\mathbf{R}$ vs H_1 : Il existe au moins une dépendance linéaire dans $\mathbf{R}$. La statistique de test est de la forme : $D = -(n-1)\log|\mathbf{R}|$ où $|\mathbf{R}|$ désigne le déterminant de $\mathbf{R}$, alors sous H_0, $D \to \chi^2_{p(p-1)/2}$

Test marginal d'indépendance : Soient $X_1, ..., X_j, ..., X_p$, p variables gaussiennes de taille n et soit R_j^2, le coefficient de détermination du modèle linéaire : $X_j^* = \sum_{k \neq j} \gamma_k X_k + \epsilon$. R_j^2 sera d'autant plus élevé que X_j sera corrélé avec des variables X_k's. Soit le jeu d'hypothèses H_0 : X_j est indépendante linéairement des $p-1$ autres variables vs H_1 : Il existe au moins une variable parmi les $p-1$ autres liée linéairement avec X_j. La statistique de test est de la forme : $F_j = [R_j^2(n-p-2)]/[(1-R_j^2)(p-1)]$, alors sous H_0 $F_j \to \mathcal{F}(p-1, n-p-2)$

2.1.2 Force de liaison des prédicteurs avec la réponse

Nous considérons le modèle de régression linéaire multiple suivant : $Y = \mathbf{X}\beta + \epsilon$ où $\epsilon \to \mathcal{N}(0, \sigma^2)$

Comme pour la multicolinéarité, nous utilisons deux tests global et marginal.

Test global d'ajustement : Soit Y, la réponse et soient $X_1, ..., X_j, ..., X_p$, les p prédicteurs numériques de taille n. Soit le jeu d'hypothèses H_0 : $\beta_1 = ... = \beta_j = ... = \beta_p = 0$ vs H_1 : Il existe un $\beta_j \neq 0$. La statistique de test est de la forme : $F = [R^2(n-p-1)]/[(1-R^2)p]$, où R^2 est le coefficient de détermination global du modèle, alors sous H_0 $F \to \mathcal{F}(p, n-p-1)$.

Test marginal d'ajustement : Soit Y, la réponse et soient $X_1, ..., X_j, ..., X_p$, les p prédicteurs de taille n. Soit le jeu d'hypothèses $H_0 : \beta_j = 0$ *vs* $H_1 : \beta_j \neq 0$. La statistique de test est de la forme : $t(X_j) = \beta_j/\sigma_{\beta_j}$, où σ_{β_j} est l'écart-type du coefficient β_j, alors sous H_0, $t(X_j) \to t_{n-p-1}$.

2.1.3 Forme du critère MCSRC

Le critère MCSRC est composé des résultats des quatre tests précédents et, de la cohérence des signes entre les coefficients de régression et les coefficients de corrélation simple. Afin de décider si un modèle peut être sélectionné, il est nécessaire de fixer des seuils d'acceptation. Ceux-ci sont obtenus à l'aide des risques de première espèce des tests proposés précédemment comme dans les méthodes classiques de sélection ascendante, descendante et pas à pas.

Pour les deux tests d'indépendance, garder l'hypothèse nulle correspondra à la non-multicolinéarité et donc l'objectif recherché, alors que pour les deux tests sur la force des liaisons des prédicteurs, le but recherché sera le rejet de l'hypothèse nulle. Formalisons ces voeux.

Soient α_D, α_r, α_F et α_t, les risques de première espèce choisis pour les tests d'indépendance globale, marginale, d'ajustement global et marginal et soient $pval_D$, $pval_r(j)$, $pval_F$ et $pval_t(j)$, les p-valeurs obtenues sur ces tests. Signalons que $pval_r(j)$ et $pval_t(j)$ correspondent aux p-valeurs spécifiques à la variable X_j. Pour les deux tests d'indépendance, l'acceptation correspond à $pval_D \geq \alpha_D$ et $pval_r(j) \geq \alpha_r$, alors que pour les deux tests d'ajustement nous avons : $pval_F \leq \alpha_F$ et $pval_t(j) \leq \alpha_t$. Signalons que le test marginal d'indépendance n'est pas utilisé dans le cas $p = 2$.

Le critère MSC pour le modèle linéaire gaussien prend la forme suivante :

$$A_1 = \left[\left(\frac{pval_D - \alpha_D}{1-\alpha_I} + \frac{1}{p} \sum_{j=1}^{p} \frac{pval_r(j) - \alpha_r}{1-\alpha_r} + \frac{\alpha_F - pval_F}{\alpha_F} + \frac{1}{p} \sum_{j=1}^{p} \frac{\alpha_t - pval_t(j)}{\alpha_t} \right) \Big/ 4 \right] \quad (1)$$

$$A_2 = 1_{[pval_D \geq \alpha_D]} \times \prod_{j=1}^{p} 1_{[pval_r(j) \geq \alpha_r]} \times 1_{[pval_F \leq \alpha_F]} \times \prod_{j=1}^{p} 1_{[pval_t(j) \leq \alpha_t]} \times \prod_{j=1}^{p} 1_{[\beta_j \times r_{yX_j} \geq 0]} \quad (2)$$

Enfin $C_{MCSRC}^{(MLG)} = A_1 \times A_2 \in [0, 1]$. Pour que le critère soit strictement positif, il est nécessaire que les $3p + 2$ contraintes dans A_2 soient respectées. L'objet A_1 permet de moduler la qualité des modèles qui respectent la multicolinéarité conjointement et la force d'ajustement du modèle. Par conséquent le vecteur des coefficients de régression du modèle sélectionné sera :

$$\tilde{\beta} = \arg \max_{m \in \mathcal{M}} C_{MCSRC}^{(MLG)}(m) \quad (3)$$

où m est un modèle appartenant à $\mathcal{M}$, l'ensemble des $2^p - 1$ modèles possibles.

Un autre intérêt de ce critère est de pouvoir calculer la contribution de chaque prédicteur à la part d'inertie expliquée par le modèle. Celle-ci est fournie par le coefficient de détermination R^2 qui peut être mis sous la forme suivante : $\sum_{j=1}^{p} \tilde{\beta}_j r_{yX_j}$, où $\tilde{\beta}_j$ est le coefficient standardisé de régression multiple associé à X_j, alors le pourcentage de contribution de chaque prédicteur est donnée par :

$$CTR(X_j) = \frac{\tilde{\beta}_j r_{yX_j}}{\sum_{j=1}^{p} \tilde{\beta}_j r_{yX_j}} \times 100\% \quad (4)$$

Remarque : Cet indicateur est seulement utilisable si les signes du coefficient de régression multiple et de corrélation linéaire simple sont identiques (possible absence de multicolinéarité).

Appliquons le critère proposé à l'exemple simulé. Nous avons choisi les risques de première espèce suivants : $\alpha_D = 0,01$, $\alpha_F = 0,01$, $\alpha_r = 0,01$ et $\alpha_t = 0,025$. Les prédicteurs sélectionnés sont X_3, X_4 et X_5. Nous pouvons constater sur la table 3 que les coefficients estimés de X_1 par les méthodes de sélection pas à pas, ascendante et stagewise sont négatifs et significatifs, alors que la corrélation simple est positive et non significative. Le critère MCSRC ne retient pas X_1. X_2 n'est pas sélectionné par stagewise et MCSRC, alors qu'elle l'est par les deux autres méthodes. Par contre X_3 est tout le temps choisi, mais avec une plus faible p-valeur pour stagewise et MCSRC. Ces deux derniers résultats sont logiques car X_2 et X_3 sont très fortement liés (cf. table 2). La variable X_4 est retenue par les quatre méthodes avec de très faibles p-valeurs alors que la corrélation simple avec la réponse est non significative. X_5 est toujours sélectionnée sauf par stagewise, ce qui est logique car la corrélation simple n'est pas significative. Et X_6 n'est jamais choisie, ce qui est logique car la corrélation simple n'est pas significative, non plus.

Seul le critère MCSRC respecte la non multicolinéarité, la valeur du critère MCSRC est de 0,856, alors qu'elle est nulle pour tous les autres. Remarquons également que le $R^2_{adj} = 0,879$ pour MCSRC est relativement proche de ceux des autres critères (entre 0,981 et 0,987) : peu de perte de qualité d'ajustement et respect de la non-multicolinéarité. De plus, la valeur obtenue pour le critère MCC [Derquenne, 2022] construit pour le modèle linéaire gaussien qui vaut 0,833 (ligne *Eval* du tableau 3) est supérieure à celles des autres méthodes. Rappelons que ce critère est dans $[0, 1]$, plus une valeur est proche de 1, plus le modèle obtenu est optimal face à la multicolinéarité. Lors de la phase de régularisation de la matrice de corrélations entre les prédicteurs, le critère MCC fournit une valeur optimale $\tilde{\delta}$ pour éliminer la multicolinéarité. Dans le cas du modèle sélectionné par le critère MCSRC (X_3,X_4,X_5), nous obtenons $\tilde{\delta} = 0,131$. Cela signifie que les corrélations entre ces trois prédicteurs sont inférieures à 0,131, en valeur absolue.

Signalons enfin que la méthode proposée peut fournir aussi un ensemble de modèles candidats acceptables (la valeur du critère est stictement positive). Dans cet exemple, les modèles potentiels sont par ordre décroissant du critère : $\{X_2, X_4, X_5\}$ (MSC=0,835,$R^2_{adj} = 0,873$) ; $\{X_3, X_4\}$ (MSC=0,796,$R^2_{adj} = 0,873$) ; $\{X_2, X_4\}$ (MSC=0,792,$R^2_{adj} = 0,869$) ; $\{X_3, X_5\}$ (MSC=0,775,$R^2_{adj} = 0,836$). Tous ces modèles respectent la cohérence des signes de coefficients de régression multiple et de corrélation simple.

	$\hat{\beta}^{MCO}$	$\hat{\beta}^{Spval}$	$\hat{\beta}^{Forw}$	$\hat{\beta}^{Swise}$	$\hat{\beta}^{MSC}$	r_{yX}
X_1	-0.379 (< 0.001)	-0.383 < 0.001)	-0.383 (< 0.001)	-0.380 (< 0.001)	0.000 (na)	0.171 (0.089)
X_2	0.708 (< 0.001)	0.695 (< 0.001)	0.695 (< 0.001)	0.000 (na)	0.000 (na)	0.910 (< 0.001)
X_3	0.419 (0.020)	0.435 (0.015)	0.435 (0.015)	1.130(< 0.001)	0.935 (< 0.001)	0.913 (< 0.001)
X_4	-0.195 (< 0.001)	-0.194 (< 0.001)	-0.194 (< 0.001)	-0.189 (< 0.001)	-0.210 (< 0.001)	-0.084 (0.406)
X_5	0.099 (0.049)	0.051 (< 0.001)	0.051 (< 0.001)	0.000 (na)	0.046 (0.023)	0.081 (0.179)
X_6	0.050 (0.320)	0.000 (na)	0.000 (na)	0.000 (na)	0.000 (na)	-0.083 (0.414)
R^2_{adj}	0.987	0.987	0.987	0.981	0.879	n.a
Eval	0.786	0.774	0.774	0.727	0.833	1.000

Table 3: Coefficients de régression des méthodes de sélection de prédicteurs

2.2 Adaptation du critère MCSRC aux modèles linéaires généralisés

Nous nous restreindrons au modèle logit booléen et au modèle logit polytomique ordonné. Pour chacun d'eux, le critère est modifié par l'ajout de nouvelles conditions. Dans le premier modèle, la statistique de Fisher est remplacée par le rapport de vraisemblances estimé et par la déviance permettant de juger l'adéquation du modèle aux données ; le second tient compte de ces deux éléments, ainsi que de la propriété du rapport de côtes proportionnelles.

2.2.1 Modèle logit booléen

Dans le cadre de la régression logistique binaire [Derquenne, 2023], l'étude de la multicolinéarité a montré que celle-ci entrainait les mêmes problèmes pour les signes des coefficients et pour les tests marginaux associés. Les méthodes de sélection de prédicteurs subissent par conséquent les mêmes écueils.

Soit $Y \in \{0;1\}$, la réponse booléenne et soient $X = (X_1, ..., X_j, ..., X_p)$, les p prédicteurs numériques de taille n. Nous considérons le modèle de régression logistique multiple dichotomique : $Pr[Y = 1/X] = \frac{exp^{\beta_0 + X\beta}}{1+exp^{\beta_0+X\beta}}$, où β est le vecteur des p coefficients. Les tests sur la force de liaison des prédicteurs et sur l'adéquation du modèle aux données sont les suivants.

Test du rapport de vraisemblances, soit le jeu d'hypothèses $H_0 : \beta_1 = ... = \beta_j = ... = \beta_p = 0$ vs H_1 : Il existe un $\beta_j \neq 0$. La statistique de test est de la forme : $LR = -2(l_{H_0} - l_{est})$, où l_{H_0} et l_{est} sont respectivement la log-vraisemblance sous l'hypothèse nulle et la log-vraisemblance estimée, alors sous H_0, $LR \rightarrow \chi^2_p$ où p est le nombre de prédicteurs. α_{LR} et $pval_{LR}$ sont respectivement le risque de première espèce et la p-valeur du test.

Test de la déviance, soit le jeu d'hypothèses H_0 : adéquation du modèle aux données vs H_1 : non adéquation. La statistique de test est de la forme $Dev = -2(l_{est} - l_{sat})$, où l_{sat} est la log-vraisemblance saturée, alors sous H_0, $Dev \rightarrow \chi^2_{n-p-1}$ où n est le nombre d'observations. α_{Dev} et $pval_{Dev}$ sont respectivement le risque de première espèce et la p-valeur du test.

Les deux statistiques pour évaluer la multicolinéarité sont celles utilisées pour le modèle linéaire gausienne, cela se justifie par le fait que la matrice de corrélations $\mathbf{R}$ entre les prédicteurs numériques se retrouve dans l'estimateur du maximum de vraisemblance et influence par conséquent les coefficients et les tests associés à ceux-ci [Derquenne, 2023]. Enfin, les coefficients de régression sont testés à l'aide de la statistique de Wald, analogue à celle de Student du modèle linéaire gaussien. α_w et $pval_w(j)$ sont respectivement le risque de première espèce et la p-valeur pour le prédicteur X_j. Signalons que le test marginal d'indépendance n'est pas utilisé dans le cas $p = 2$.

Le critère MCSRC pour le modèle logit dichotomique est de la forme suivante :

$$B_1 = \left[\left(\frac{pval_D - \alpha_D}{1-\alpha_D} + \left(\frac{1}{p} \sum_{j=1}^{p} \frac{pval_r(j) - \alpha_r}{1-\alpha_r} + \frac{\alpha_w - pval_w(j)}{\alpha_w} \right) + \frac{\alpha_{LR} - pval_{LR}}{\alpha_{LR}} + \frac{pval_{Dev} - \alpha_{Dev}}{1-\alpha_{Dev}} \right) \Big/ 5 \right] \quad (5)$$

$$B_2 = 1_{[pval_D \geq \alpha_D]} \times \prod_{j=1}^{p} 1_{[pval_r(j) \geq \alpha_r]} \times 1_{[pval_{LR} \leq \alpha_{LR}]} \times \prod_{j=1}^{p} 1_{[pval_w(j) \leq \alpha_w]} \times 1_{[pval_{Dev} \geq \alpha_{Dev}]} \times \prod_{j=1}^{p} 1_{[\beta_j^{mul} \times \beta_j^{uni} \geq 0]} \quad (6)$$

où β_j^{mul} et β_j^{uni} sont respectivement les coefficients de régression logistique multiple et simple. Enfin $C_{MCSRC}^{(logit(0,1))} = B_1 \times B_2 \in [0, 1]$. Pour que le critère soit strictement positif, il est nécessaire

que les $3p+3$ contraintes dans B_2 soient respectées. L'objet B_1 permet de moduler la qualité des modèles qui respectent la non-multicolinéarité conjointement à la force d'ajustement du modèle et l'adéquation de celui-ci aux données. Par conséquent le vecteur des coefficients de régression du modèle sélectionné sera :

$$\tilde{\beta} = \arg\max_{m \in \mathcal{M}} C_{MCSRC}^{(logit(0,1))}(m) \tag{7}$$

2.2.2 Modèle logit ordinal

Soit $Y \in \{1; ...; k; ...; ...; r\}$, la réponse ordinale à r catégories ordonnées et soient $X = (X_1, ..., X_j, ..., X_p)$, les p prédicteurs numériques de taille n. Nous considérons le modèle de régression logistique multiple ordinale : $Pr[Y \leq 1/X] = \frac{exp^{\theta_1 + X\beta_1}}{1+exp^{\theta_1+X\beta_1}}$, ..., $Pr[Y \leq k/X] = \frac{exp^{\theta_k+X\beta_k}}{1+exp^{\theta_k+X\beta_k}}$, ..., $Pr[Y \leq r-1/X] = \frac{exp^{\theta_{r-1}+X\beta_{r-1}}}{1+exp^{\theta_{r-1}+X\beta_{r-1}}}$. Ce modèle est nommé modèle logit à rapport de côtes proportionnelles, si les $r-1$ vecteurs de coefficients β_k, pour $k = 1, 2, ..., r-1$ sont égaux. En d'autres termes, $\forall, j = 1, ..., p; \beta_{1,j} = ... = \beta_{k,j} =, ... = \beta_{r-1,j}$. Cette condition représente l'hypothèse nulle du test du rapport de côtes proportionnelles (proportionnal odds ratio).

Deux types de tests peuvent être réalisés : le premier est global car il compare les $r-1$ vecteurs de coefficients, cela revient à tester si les $r-1$ pentes sont parallèles, il y a dans ce cas $(r-2)p$ égalités à vérifier ; le second teste marginalement pour chaque prédicteur X_j, si $\beta_{1,j} = ... = \beta_{k,j} = , ... = \beta_{r-1,j}$, il n'y a dans ce cas que $r-2$ égalités à respecter.

Le test global du rapport de côtes proportionnelles peut être effectué à l'aide des statistiques du rapport de vraisemblances, du score et de Wald. Par exemple, la statistique du rapport de vraisemblances prend la forme suivante : $LR_{OR} = -2(l_{orp} - l_{norp})$, où l_{orp} et l_{norp} sont respectivement la log-vraisemblance du modèle à rapport de côtes proportionnelles et la log-vraisemblance du modèle général. Sous H_0, $LR_{OR} \rightarrow \chi^2_{(r-2)p}$. α_{OR} et $pval_{OR}$ sont respectivement le risque de première espèce et la p-valeur du test.

Le test marginal peut être effectué notamment à l'aide du test de Brant [Brant, 1990]. Pour chaque prédicteur, Brant teste sous H_0, l'égalité des coefficients adjacents, tel que : $\beta_{\leq k,j} = \beta_{>k:r,j}$ pour $k = 1, r-1$ où $\beta_{\leq k,j}$ et $\beta_{>k:r,j}$ sont respectivement les coefficients univariés des modèles : logit$(Y \leq k \ vs \ Y > k)$ et logit$(Y \leq k+1 \ vs \ Y > k+1)$. La statistique de Wald prend la forme suivante : $w_{jk} = (\beta_{\leq k,j} - \beta_{>k:r,j})^2 / (\sigma^2_{\beta_{\leq k,j}} + \sigma^2_{\beta_{>k:r,j}})$. Sous H_0, $w_{jk} \rightarrow \chi^2_1$, lorsque $r = 3$. Si $r > 3$, alors cette statistique est composée de $r-2$ éléments et elle suivra un χ^2_{r-2}. α_B et $pval_{B(j)}$ sont respectivement le risque de première espèce et la p-valeur du test. Signalons que ce test marginal n'est pas utilisé dans le cas $p = 2$.

Le critère MCSRC pour le modèle logit ordinal est de la forme suivante :

$$C_1 = \left[\left(5B_1 + \frac{pval_{OR} - \alpha_{OR}}{1 - \alpha_{OR}} + \frac{1}{p} \sum_{j=1}^{p} \frac{pval_{B(j)} - \alpha_B}{1 - \alpha_B} \right) \Big/ 7 \right] \tag{8}$$

$$C_2 = B_1 \times 1_{[pval_{OR} \geq \alpha_{OR}]} \times \prod_{j=1}^{p} 1_{[pval_{B(j)} \geq \alpha_B]} \tag{9}$$

Enfin $C_{MCSRC}^{(logit(ord))} = C_1 \times C_2 \in [0, 1]$. Pour que le critère soit strictement positif, il est nécessaire que les $4p+4$ contraintes dans C_2 soient respectées. L'objet C_1 permet de moduler la qualité des modèles qui respectent la non-multicolinéarité, la force d'ajustement du modèle, l'adéquation

de celui-ci aux données et le rapport de côtes proportionnelles. Le vecteur des coefficients de régression du modèle sélectionné sera :

$$\tilde{\beta} = \arg\max_{m \in \mathcal{M}} C_{MCSRC}^{(logit(ord))}(m) \tag{10}$$

3 Apports, applications et voies futures

Le critère MCSRC proposé est une nouvelle méthode de sélection de variables dans le cadre des modèles linéaires généralisés qui permet de respecter la non-multicolinéarité entre les prédicteurs. Ce critère est constitué de plusieurs statistiques : l'indépendance globale et marginale pour tester la non-multicolinéarité, l'ajustement global du modèle aux données, les effets marginaux des coefficients de régression multiple, la cohérence des signes de ceux-ci avec les coefficients de régression simple, et dans le cas du modèle logit ordinal, le rapport de côtes proportionnelles. L'avantage de MCSRC sur les autres méthodes de sélection de variables réside en particulier sur la possibilité d'interpréter les résultats du modèle en "toute sécurité" car l'effet néfaste de la multicolinéarité ne peut pas apparaitre dans le modèle sélectionné. Cet avantage est renforcé par la préservation de la qualité d'ajustement du modèle pour expliquer la réponse qui reste comparable aux autres méthodes. En d'autres termes, le modèle sélectionné possède deux propriétés : interprétabilité et prévision. Le critère introduit dans cet article a été évalué sur différents cas simulés (pas de corrélation entre prédicteurs, corrélations modérées, très fortes corrélations), ainsi que sur d'autres exemples issus de la littérature, et des cas réels d'applications dans les domaines marketing, médical et management de l'énergie. Les voies futures vont consister à établir les propriétés mathématiques du critère proposé, l'étendre à la prévision de chroniques à l'aide de plusieurs prédicteurs temporels et à la modélisation multivariée de réponses.

Bibliographie

Bastien Ph., Esposito Vinzi E. & Tenenhaus M., (2005), PLS generalized linear regression, *Computational Statistics & Data Analysis*, **48**(1), 17-46.

Brant R., (1990), Assessing proportionality in the proportional odds model for ordinal logistic regression, *Biometrics* **46**, 1171-1178.

Derquenne Ch., (2022), Un multi-critère pour contrôler la multicolinéarité dans les modèles linéaires de régression multiple, *53ièmes Journées de Statistique*, Lyon, France, 538-543.

Derquenne Ch., (2023), Un multi-critère pour traiter la multicolinéarité dans les modèles linéaires généralisés, *54ièmes Journées de Statistique*, Bruxelles, Belgique.

Hoerl, A.E. & Kennard R.W., (1970), Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 42 (1)*, 80-86.

Tibshirani, R., (1996), Regression Shrinkage and Selection via the Lasso, *J. R. Statist. Soc. B*, **58**, *No.* 1, 267-288.

Wold S., Ruhe A., Wold H. & Dunn III W.J., (1984), The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM J. Sci. Stat. Comput.*, **5**, n°3, 735-743.

Zou H. & Hastie T., (2005), Regularization and Variable Selection via the Elastic Net, *J. R. Statist. Soc. B*, **67**, *No.* 2, 301-320.