Sélection de variables par le GLM-Lasso pour la prédiction du risque palustre.

Bienvenue Kouwayè [1,2], Noël Fonton [2] & Fabrice Rossi [3]

[1] *Uiversité paris 1, 90 rue de Tolbiac, kouwaye2000@yahoo.fr*
[2] *Université d'Abomey-Calavi, CIPMA 072 BP 50 Cotonou, hnfonton@gmail.com*
[3] *Uiversité paris 1, 90 rue de Tolbiac, Fabrice.Rossi@univ-paris1.fr*

Résumé. Nous étudions dans ce travail une méthode de sélection de variables basée sur le Lasso dans le contexte épidémiologique. L'un des objectifs est de construire automatiquement un modèle prédictif en limitant le recours aux experts médicaux qui opèrent des prétraitements sur les données collectées. Ces prétraitements consistent entre autres à recoder certaines variables en classe et à choisir manuellement certaines interactions en se basant sur la connaissance des données. L'approche proposée utilise toutes les variables explicatives sans traitement et génère automatiquement toutes les interactions entre les variables, ce qui nous conduit en grande dimension. Nous utilisons le Lasso qui est une méthode robuste de sélection de variables en grande dimension. Le nombre d'observations dans les études épidémiologiques étant faible, nous proposons une validation croisée à deux niveaux pour éviter le risque de sur apprentissage dans la phase de sélection de variables. Les estimateurs Lasso étant biaisés et la variable d'intérêt qu'est le nombre d'anophèles à prédire étant discret, nous utilisons un modèle GLM pour débiaiser les variables sélectionnées par le Lasso et faire de la prédiction. Les résultats montrent que quelques variables climatiques et environnementales seulement sont des facteurs principaux liés au risque d'exposition au paludisme.

Mots-clés. Lasso, validation croisée, sélection de variables, prédiction.

Abstract. In this study, we propose an automatic learning method for variables selection based on Lasso in epidemiology context. One of the aim of this approach is to overcome the pretreatment of experts in medicine and epidemiology on collected data. These pretreatment consist in recoding some variables and to choose some interactions based on expertise. The approach proposed uses all available explanatory variables without treatment and generate automatically all interactions between them. This lead to high dimension. We use Lasso, one of the robust methods of variable selection in high dimension. To avoid over fitting a two levels cross-validation is used. Because the target variable is account variable and the lasso estimators are biased, variables selected by lasso are debiased by a GLM and used to predict the distribution of the main vector of malaria which is Anopheles. Results show that only few climatic and environmental variables are the mains factors associated to the malaria risk exposure.

Keywords. Lasso, cross-validation, variable selection, prediction.

1 Introduction

Le paludisme est un problème de santé publique en Afrique surtout dans la zone sub-saharienne. Il constitue la première cause de mortalité pour des enfants de moins de cinq ans et frappe essentiellement les couches les plus vulnérables de la population : les femmes enceinte et les nouveau-nés. Des études de cohorte ont été conduites dans les zones endémiques pour étudier la mise en place et l'évolution du système immunitaire du nouveau-né face à cette maladie. Ces études ont aussi pour objectif d'étudier les déterminants liés à l'apparition des premières infections palustres chez le nouveau-né. Certaines études ont montré que la distribution du principal vecteur du paludisme qu'est l'anophèle ainsi que le risque d'exposition au paludisme présentent des dépendances à la fois spatiales et temporelles et non homogènes à une petite échelle (niveau maison) [2]. Dans l'analyse et le traitement des données issues de ces enquêtes, les experts opèrent des prétraitements qui consistent entre autre à recoder certaines variables en classes et à choisir manuellement des interactions de façon experte entre les variables explicatives. Ils utilisent ensuite des méthodes classiques de type *forward, backward* pour la sélection de variables [8]. L'objectif principal de ce travail est de s'affranchir de la phase de prétraitement des experts médicaux qui coûte en temps et qui présente un risque et de construire de façon automatique un modèle prédictif utilisant toutes les variables ainsi que toutes les interactions entres ces variables. Ce nombre élevé de variables nous conduit en grande dimension. Nous utilisons le Lasso, une méthode régularisante qui fait à la fois de la sélection est de l'estimation et qui est robuste pour la sélection de variables en grande dimension. Dans les enquêtes épidémiologiques, les observations sont peu nombreuses. Dans la sélection de variables, nous proposons une validation croisée à deux niveaux pour éviter le risque de sur apprentissage [7]. La variable d'intérêt est le risque d'exposition au paludisme, qui revient au nombre d'anophèles collectés dans les maisons donc discrète alors nous utilisons un modèle simple de type GLM avec un lien poisson. Ainsi le GLM-Lasso permet de faire la sélection de variables et le GLM permet de débiaiser les coefficients des variables sélectionnées par le Lasso pour la prédiction. Les résultats obtenus seront comparés à ceux de la méthode de référence (B-GLM) basée l'intervention des experts [2]. Ces résultats montrent que quelques variables climatiques et environnementales sont les facteurs principaux liés au risque d'exposition au paludisme.

2 Méthodologie

2.1 Collecte des données et variables utilisées

Les données utilisées dans ce travail proviennent d'une enquête épidémiologique conduite entre juillet 2007 et juillet 2009 dans la commune de Tori-Bossito au Bénin. Les données sont de deux types : climatiques et environnementales (saison, quantité de pluie, type de végétation, type de sol, etc), et des données entomologiques (nombre de moustiques,

nombre d'anophèles infectés ou non.).

2.2 Modèle d'étude

Le GLM-Lasso consiste à pénaliser la log-vraisemblance du GLM en ajoutant une pénalité L_1 [3,4,5]. Les coefficients des variables sont donnés par :

$$\hat{\beta} = Arg\max_{\beta} \left[l_{GLM}(\beta|Y) + \lambda \sum_{i=1}^{p} \beta_i \right] \quad \text{avec} \;\; \lambda \geq 0 \tag{1}$$

Le choix du paramètre λ se fait en minimisant le score. En pratique, l'équation (1) n'a pas de solution numérique exacte. On utilise l'approximation de Laplace, la méthode de Newton-Raphson ou la méthode du score de Fisher. Les coefficients du Lasso étant biaisés, on utilise le GLM pour les débiaiser et faire de la prédiction. Sous forme matricielle le GLM se présente comme suit :

$$g[E(Y|\beta)] = X\beta \tag{2}$$

où $(Y|\beta)$ suit une loi de Poisson de paramètre $E(Y|\beta)$, n est le nombre observations, X la matrice de dimensions $n \times (p+1)$ des co-variables (environnementales et climatiques), β est le vecteur de longueur $(p+1)$ des effects fixes y compris la constante, Y est le vecteur des observations de la variable d'intérêt. Ainsi

$$\mathbb{P}((Y = y_i|X = x)) = \frac{e^{(x\beta)^{y_i}}}{(y_i)!} \times e^{-e^{x\beta}} \tag{3}$$

Si on pose $Z_i = (Y = y_i|X = x)$ alors la vraisemblance des n observations peuvent être définie comme :

$$L(Z_1, \ldots, Z_n) = \prod_{i=1}^{n} \frac{e^{(x\beta)^{y_i}}}{(y_i)!} \times e^{-e^{x\beta}} \tag{4}$$

et la log-vraisemblance devient :

$$\mathcal{L}(Z_1, \ldots, Z_n) \;\; = \;\; \log\left(\prod_{i=1}^{n} \frac{e^{(x\beta)^{y_i}}}{(y_i)!} \times e^{-e^{x\beta}} \right) \tag{5}$$

$$\mathcal{L}(Z_1, \ldots, Z_n) = Cste + \sum_{i=1}^{n} y_i(x\beta) - e^{(x\beta)} \quad \text{où} \quad Cste = -\sum_{i=1}^{n} \log((y_i)!) \tag{6}$$

2.3 Algorithme LOLO-DCV

L'algorithme Leave one level out double cross-validation (LOLO-DCV) étudié dans ce travail est basé sur une validation croisée statifiée à deux niveaux. Le deuxième niveau de validation croisée permet d'éviter le risque de sur apprentissage dans la phase de sélection

Algorithme 2.1 LOLO-DCV

1. Les données sont divisées en N-blocs
2. A chaque étape du premier niveau de la validation-croisée
   (a) Les blocs sont regroupés en deux parties : E_A et E_T, E_A : l'ensemble d'apprentissage qui contient les observations de $(N-1)$-blocs,
   E_T : l'ensemble de test, contenant les observations du dernier bloc.
   (b) On met de côté E_T
   (c) deuxième niveau de validation croisée.
      i. On opère une validation-croisée complète sur E_A
      ii. les deux paramètres de régularisation $\lambda.min$ et $\lambda.1se$ sont récupérés.
      iii. Les coefficients des variables actives (variables à coefficient non nul) associés à ces deux paramètres sont récupérés et débiaisés.
      iv. On utilise un modèle GLM pour faire de la prédiction sur E_T
      v. La présence $\mathcal{P}(X_i)$ de chaque variable est déterminée
3. l'étape (2c) est répétée jusqu'à faire de la prédiction pour toutes les observations.

de variables parce que le nombre d'observations n'est pas élevé. L'algorithme se présente comme décrit dans (2.1). Il est basé sur le score de validation qui est la déviance du modèle définit comme :

$$Score(\lambda_i) = Deviance(\lambda_i) = 2 \times \left(\mathcal{L}_{(sat)} - \mathcal{L}_{(\lambda_i)}\right) \tag{7}$$

où $\mathcal{L}_{(sat)}$ est la log-vraisemblance du modèle complet qui ajuste parfaitement les données, et $\mathcal{L}_{(\lambda_i)}$ log-vraisemblance du modèle considéré.

$$Score(\lambda_{max}) = Deviance(NULL) = 2 \times \left(\mathcal{L}_{(sat)} - \mathcal{L}_{\lambda_{max}}\right) \tag{8}$$

Le modèle obtenu à $\lambda = \lambda_{max}$ on obtient modèle nul (le modèle contenant uniquement l'intercept). En posant

$$R = 1 - \frac{Score(\lambda_i)}{Score(\lambda_{max})} = 1 - r \tag{9}$$

on a : $Deviance(\lambda_i) = (1-R) \times Score(\lambda_{max})$. On sait que $\mathcal{L}_{(sat)} = 0$ et ainsi r devient le rapport de vraisemblance entre le modèle considéré et le modèle nul.

La valeur optimale $\lambda.min$ de λ est celle qui minimise la fonction $Score(.)$.

$$\lambda.min = Arg \min_{\lambda_i}[Score(\lambda_i)] \tag{10}$$

La valeur $\lambda.1se$ est telle que définie par T. Hastie et *al* qui minimise le score plus sa déviation standard [5]. Pour $\lambda.min$ et $\lambda.1se$, l'algorithme détermine les variables les plus

TABLE 1 – **Critères de qualité pour B-GLM**

Méthode	Deviance	W.Deviance	Pouvoir prédictif (%)
B-GLM	3101.68	3101.49	73.53

TABLE 2 – **Critères de qualité pour LOLO-DCV**

Méthode	Deviance	W.Deviance	Pouvoir prédictif (%)
LOLO DCV lamba_min	5573.98	5573.67	78.76
LOLO DCV lamba_1se	5573.98	5573.67	78.76
Var freq lamba_min	2860.75	2860.59	75.00
Var freq lamba_1se	3259.69	3259.53	76.80

fréquentes (Var_freq), variables qui apparaissent un certain nombre de fois au premier niveau de la validation-croisée selon un seuil fixé. Ces sous ensembles de variables fréquentes sont utilisée pour la prédiction via un GLM.

2.4 Pouvoir Prédictif et critère de qualité

Les critères de qualité utilisés pour la sélection sont : La déviance définie plus haut, la déviance pondérée *W.Deviance* définie par :

$$W.Deviance(\lambda_i) = \frac{\frac{1}{w_i} \times Deviance(\lambda_i)}{\sum_i \frac{1}{w_i}} \tag{11}$$

où le nombre d'observations de l'ensemble d'apprentissage
et le Pouvoir prédictif P_a défini par :

$$\begin{cases} P_a(\hat{Y}_i) = 1 & \text{si} \quad -0.5 \leq Y_i - \hat{Y}_i \leq 0.5 \\ P_a(\hat{Y}_i) = 0 & \text{sinon.} \end{cases}$$

où $\hat{Y}_i$ est la prédiction pour chaque observation Y_i.

3 Résultats et Conclusion

Le meilleur sous-ensemble optimal de variables pour chaque méthode est :
B-GLM : La saison, le nombre de jours de pluie, la quantité moyenne de pluie, l'utilisation de répulsif, la végétation, l'interaction entre saison et la végétation.

LOLO-DCV : La saison et l'interaction entre le nombre de jours de pluie et le village. Les résultats des tables (1, 2) montrent que les meilleurs prédictions sont obtenues par LOLO-DCV et le sous ensemble optimale pour la prédiction de LOLO-DCV est plus parcimonieux que celui obtenu par la méthode (B-GML). Ces résultats montrent que la machine peut remplacer les experts pour la sélection de variables et améliorer leurs résultats.

Bibliographie

[1] De Brabanter J., Pelckmans K., Suykens J.A.K., Vandewalle J. , (2002) Robust cross-validation score function for LS-SVM non-linear function estimation, *: Int. Conference on Artificial Neural Networks - ICANN*, pp 713-719.

[2] G. Cottrell, B. Kouwayè and *al*, (2012) Modeling the Influence of Local Environmental Factors on Malaria Transmission in Benin and Its Implications for Cohort Study, *PlosOne*, 7(8).

[3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. (2004),Least angle regression, *The Annals of statistics.* , 32 :407-499.

[4] R. Tibshirani, (1996) : Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Serie B (Methodological)*, 58 :267-288.

[5] H. Zou and T. Hastie, (2005) : Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society. Serie B*, 67 :301-320.

[6] M. Osborne, B. Presnell and B. Turlach, (2000) : A new approach to variable selection in least squares problems, *IMA Journal of Numerocal Analysis*, 20 :389-403.

[7] A. Y. Ng, (1997), Preventing "Overfitting" of Cross-Validation Data, *International Conference on Machine Learning*, pp 245-253.

[8] I. Guyon, B. Presnell and B. Turlach, An Introduction to Variable and Feature Selection, 3 :1157-1182, Journal of Machine Learning Research, 2003. [9] B Gianluca. : Structural feature selection for wrapper methods, editor, *proceedings of the* 13^{th} *European Symposium on Artificial Neural Networks* (ESANN 2005), d-side pub., pages 405-410, April 27-29, Bruges (Belgium), 2005.