Dévoir N°4 Statistiques en grande dimension

Exaucé LUWEH ADJIM NGARTI[1] & Yassine LAAMOUMRI[2]

[1] *CMI ISI exauce.luweh-adjim-ngarti@etu.u-bordeaux.fr*

[2] *CMI ISI yassine.laamoumri@etu.u-bordeaux.fr*

1 Introduction

Fin 2019, un nouveau virus responsable du syndrome respiratoire aigu sévère a été signalé à Wuhan, en Chine causant la pandémie COVID-19, responsable de millions de morts dans le monde. Tzampoglou et Loukidis (2020) (2) tente d'expliquer l'impact des facteurs climatiques et socio-démographiques sur le nombre de cas et de decès. Notre objectif dans cet article est de comparer des méthodes de sélection de variables adaptées à la grande dimension sur un critère de prédiction et sur un critère d'identification, et en tenant compte de la possible non linéarité des prédicteurs, ce qui n'est pas le cas dans Tzampoglou et Loukidis (2000) (2).

2 Méthodes

2.1 Sélection par la méthode exhaustive

La régression par la méthode exhaustives est une approche de sélection de modèle qui consiste à tester toutes les combinaisons possibles des prédicteurs, puis sélectionner le meilleur modèle en fonction de certains critères statistiques. Cette méthode est très coûteuse en temps de calcul et devient même impossible pour p très grand. Il est dit dans Hastie et al (2009) (1) qu'il est toujours possible de le faire du moment où $p \leq 40$, ce qui est notre cas ($p = 27$).

2.2 Sélection pas à pas ascendante et descendante

Comme nous l'avons susmentionner, la méthode précédente est impossible si $p >> 40$, nous avons deux autres méthodes similaire mais plus efficace :

- La **sélection pas à pas ascendante** , qui commence sans prédicteur dans le modèle, ajoute itérativement les prédicteurs les plus contributifs, et s'arrête lorsque l'amélioration n'est plus statistiquement significative.
- La **sélection pas à pas descendante**, qui commence avec tous les prédicteurs, élimine itérativement les prédicteurs les moins contributifs, et s'arrête lorsque tous les prédicteurs sont statistiquement significatifs.

2.3 Régression Ridge et Lasso

Les méthodes de sélection vues jusqu'à présent permettent de choisir un meilleur modèle. Toutefois ces méthodes sont très variables. Pour réduire cette variabilité on peut s'orienter vers

les modèles de régression pénalisés qui consistent à ajouter un terme de pénalité au modèle. On s'intéresse à deux variantes :

— **Régression Lasso** : consiste à ajouter une pénalité sous la forme $\lambda \sum_{j=1}^{p} |\beta_j|$. Cette variante permet de sélection certaines en varibles en forçant certains coefficients à passer à 0.
— **Régression Ridge** : consiste à ajouter une pénalité sous la forme $\lambda \sum_{j=1}^{p} \beta_j^2$. Cette variante est moins biaisée que la première mais ne permet pas une sélection de variable.

2.4 PCR, PLS, et sPLS

Les régressions sur composantes consistent à trouver de nouvelles composantes qui s'écrivent le plus souvent comme des combinaisons linéaires des p variables explicatives dans l'idée de diminuer le nombre de paramètres du modèle ou la dépendance entre les covariables. Il existe plusieurs façons de construire ces composantes, nous nous intéressons à :

— La **régression PCR** : il s'agit de faire simplement une ACP sur la matrice des variables explicatives.
— La **régression PLS** : la principale différence est qu'on ne cherche pas les composantes qui maximisent la variabilité des observations projetées mais celle qui maximisent la colinéarité avec la cible.
— **La régression sPLS** : consiste à faire une régression PLS tout en ajoutant une pénalité type Lasso.

2.5 Critères

Pour les méthodes de sélection pas à pas ascendante, descendante où la méthode exhaustive. Nous allons utiliser 3 critères pour le choix du meilleur modèle :

— **P valeur** : c'est le risque de conclure à tort que les résultats obtenus ne peuvent pas être dus au hasard.
— **AIC** : Le critère AIC représente donc un compromis entre le biais, et la parcimonie.
— **R2 ajusté** :R2 ajusté mesure le degré d'adéquation et pénalise le nombre de variables.
En fin, pour comparer nos différentes méthodes de sélection, nous allons tester le pouvoir prédictif des modèles choisis en calculant le **RMSE** qui indique à quel point les données sont concentrées autour de la ligne de meilleur ajustement.

3 Résultats

3.1 Données

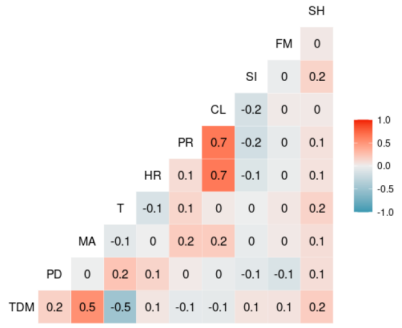
Nous avons 55 observations de 10 variables. Lire Tzampoglou et al 2020 (2) pour plus d'informations sur le collect des données.

La variable réponse
— TDM (le nombre total des décès par million d'habitants dus au COVID-19 entre le 1/3/2020 et le 31/5/2020),

Variables explicatives

— **T** la température moyenne (∘C)
— **HR** l'humidité relative moyenne (%),
— **PR** le cumul des précipitations (mm)
— **CL** la fraction de la couverture nuageuse
— **PD** la densité de population (personnes/km2)
— **MA** l'âge médian de la population,
— **SI** l'indice de "rigueur" (mesure composite basée sur 9 indicateurs)
— **FM** le délai entre le 1er cas et l'imposition des 1ères mesures (jours),
— **SH** le délai entre le 1er cas et l'ordre de rester à la maison (jours).

3.2 Analyse descriptive



(a) Corrélation linéaire

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -494.3869   376.9177  -1.312  0.19644
T            -10.3388     1.8535  -5.578 1.40e-06 ***
HR             0.2449     1.6268   0.151  0.88104
PR             0.1380     1.7311   0.080  0.93681
CL          -245.7273   327.4389  -0.750  0.45698
PD             0.8201     0.2868   2.860  0.00646 **
MA            16.2195     3.1112   5.213 4.75e-06 ***
SI             1.5211     2.3170   0.657  0.51492
FM             1.5278     1.1960   1.277  0.20815
SH             1.3655     0.6571   2.078  0.04357 *
ID            -0.4071     0.6105  -0.667  0.50834
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 69.09 on 44 degrees of freedom
Multiple R-squared:  0.6474,    Adjusted R-squared:  0.5672
F-statistic: 8.078 on 10 and 44 DF,  p-value: 3.179e-07
```

(b) Corrélation multi-linéaire

FIGURE 1 – Corrélation entre les variables

Nous remarquons grâce à la figure 1a qu'il y a très peu de corrélation entre les variables explicatives. Sauf la couverture nuageuse (**CL**) qui est corrélée avec l'humidité relative(**HR**) et le cumul de précipitations (**PR**) .

La variable à expliquer (**TDM**) est corrélée négativement avec la température (**T**) et positivement corrélée avec l'âge médian (**MA**) sinon très peu corrélée avec les autres variables. La table 1 nous permet d'affirmer que cette corrélation est statistiquement significative.

Test de corrélation avec TDM			
Méthode	Pearson	Kendall	Sperman
Estimation MA	0.53	0.35	0.48
P valeur	$2.94x10^{-5}$	$2.94x10^{-5}$	$2.94x10^{-5}$
Estimation T	-0.48	-0.52	-0,68
P valeur	$2.94x10^{-5}$	$2.94x10^{-5}$	$2.94x10^{-5}$

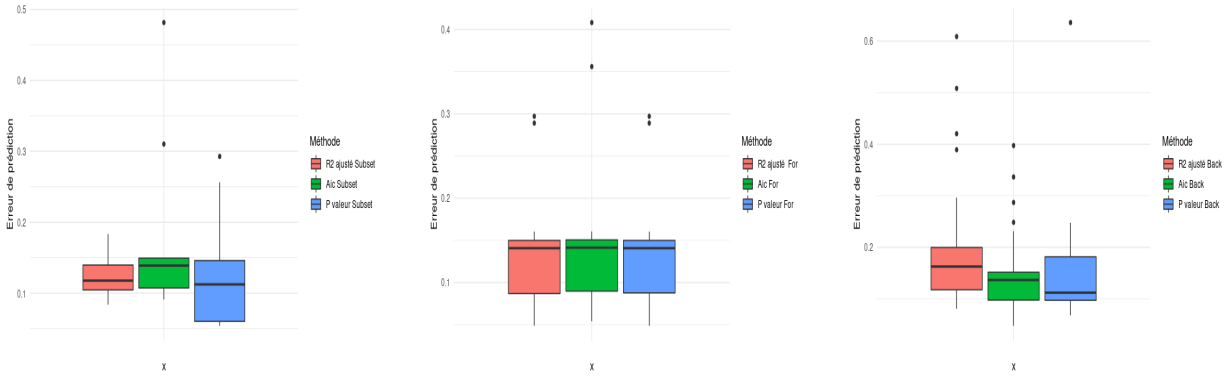
TABLE 1 – Test de corrélation de l'âge médian et de la température avec le taux de mortalité.

Même si la variable à expliquer est très peu liée aux variables explicatives. Il est possible qu'elle soit liée de manière multiple à certaines variables. Un test de corrélation multiple permet de desceller une telle liaison. La figure 1b confirme les remarques faites précédemment, il existe une corrélation linéaire faible entre les prédicteurs et la variable à expliquer. L'hypothèse de linéarité n'est pas admise, soit $\mathcal{X}$ l'ensemble des prédicteurs et n le nombre de prédicteurs, nous avons donc cette formulation du modèle :

$$TDM = \sum_{i=1}^{3} \sum_{j=1}^{n} \beta_{ij} X_i^j. \tag{1}$$

3.3 Comparaison de modèles

Nous allons comparer différentes méthodes de prédiction décrite dans la section 2 en considérant le modèle défini par 1. Dans tous les modèles les données ont été normalisées comme suit : pour chaque variable X, $X = \frac{X - X_{min}}{X_{max} - X_{min}}$ et nous avons effectué 30 répétitions pour chaque méthode.



(a) Sélection exhaustive (b) Sélection pas à pas ascendante (c) Sélection pas à pas descendante

FIGURE 2 – Erreur de prédiction, nombre de répétitions = 30

D'après la figure 2, pour la méthode de sélection exhaustive, le critère **R2 ajusté** présente un meilleur résultat. Et pour la méthode de sélection pas à pas ascendante les trois critères ont sensiblement la même performance. En revanche pour la méthode de sélection pas à pas descendante le critère **AIC** présente de meilleur résultat car l'erreur est basse avec une faible dispersion. Dans la suite nous allons comparer que les méthodes avec les critères avec de meilleurs résultats aux autres méthodes.

Si on considère la médiane (figure 3), la méthode Ridge est la plus efficace pour la prédiction. La méthode Lasso présente une performance très proche et avec parcimonie donc semble plus intéressante. Toutefois, malgré que la méthode exhaustive est plus coûteuse,son erreur de prédiction présente moins de dispersion.

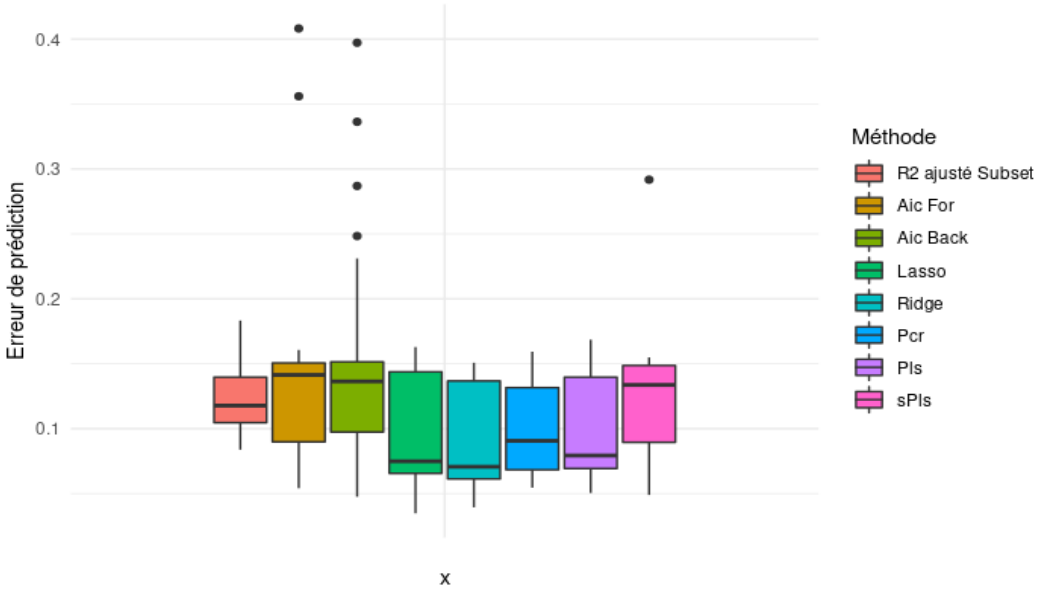


FIGURE 3 – Comparaison de méthodes de sélection, nombre de répétitions = 30

3.4 Importance des variables

A chaque répétition une variable X est considérée comme importante si et seulement si X, X^2 ou X^3 est considérée comme importante.

Importance des variables									
Méthode	T	HR	PR	CL	PD	MA	SI	FM	SH
Subset Aic	10	8	8	9	9	10	10	7	8
Subset R2	2	0	0	0	0	8	0	0	0
Subset pvalue	10	2	6	3	10	10	4	3	4
Forward Aic	30	8	11	17	30	30	24	23	16
Forward R2	30	1	5	8	30	30	18	13	8
Forward Pvalue	30	1	5	8	30	30	19	14	8
Backward Aic	30	28	25	29	30	30	29	27	29
Backward R2	30	12	16	15	27	30	17	16	15
Backward pvalue	7	9	6	8	5	18	2	5	5
Lasso	30	13	25	26	30	30	23	25	28
Pourcentage	**69.67**	**27.33**	**35.67**	**41**	**67**	**75.33**	**48.67**	**44.33**	**40.33**

TABLE 2 – Nombre de fois où les variables ont été retenues comme pertinentes par les méthodes

Nous remarquons que les 3 variables (T,MA et PD) considérées comme significatives par le modèle linéaire sont les variables le plus souvent retenues par le nouveau modèle.Toutefois il faut noter que la nouvelle formulation permet de sélectionner plus de variables.

3.5 Impact de la température sur le TDM

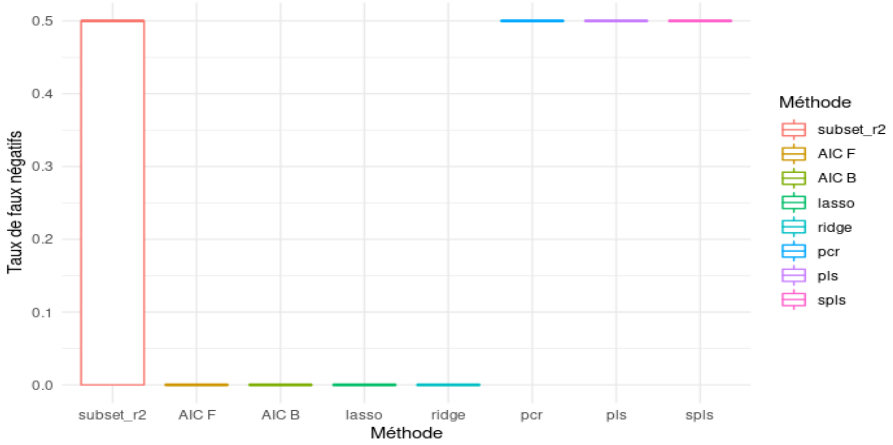


FIGURE 4 – Taux de faux négatifs, nombre de répétitions = 30

D'après la figure 4, les méthodes qui réussissent à retenir cette variable sont : sélection pas à pas ascendante et descendante avec critère AIC et les méthodes de régression Lasso et Ridge. Les méthodes PCR, PLS et sPLS ont un taux de négatifs de 0.5 mais en décomptant l'intercepte, on ne peut considérer ces méthodes.

4 Conclusion

En somme, la méthode Lasso est la plus efficace dans notre cas car avec une erreur relativement basse avec une faible dispersion et plus parcimonieuse. Pour une sélection de variables, les méthodes telles que PCR, PLS et sPLS sont en revanche très peu adaptées car l'analyse descriptive souligne une très faible corrélation entre les variables. Le test de corrélation multiple sans les variables polynomiales considèrent uniquement 3 variables explicatives comme importante alors qu'en les considérant toutes les variables ont été considérées comme importante dans au moins 20% des cas(table 2). Le travail effectué par Tzampoglou et Loukidis (2020) (2) peut-être amélioré en supposant une non linéarité des prédicteurs.

Bibliographie

[1] Hastie, T. ; Tibshirani, R. ; Friedman, J. The Elements of Statistical Learning : Data Mining, Inference, and Prediction ; Springer Science Business Media : Berlin/Heidelberg, Germany, 2009

[2] Tzampoglou P, Loukidis D. Investigation of the Importance of Climatic Factors in COVID-19 Worldwide Intensity. Published 2020 Oct 22

[3] Obilor, Esezi Isaac and Amadi, Eric.Test for Significance of Pearson's Correlation Coefficient.Published 2018 mars.