DEVOIR NUM 04 DE L'UE STATISTIQUES EN GRANDE DIMENSION

Merveille Agbo
Master MSS & merveille.agbo@etu.u-bordeaux.fr

Introduction

Nous sommes tous témoins du nouveau virus responsable du syndrome respiratoire aigu sévère (SRAS-CoV-2) qui a été signalé à Wuhan, en Chine et qui s'est rapidement répandu dans le monde entier, entraînant la pandémie de COVID-19 en cours. Face à cela, plusieurs chercheurs ont tentés de comprendre ce virus. Parmi eux, nous nous intéressons à l'étude de Tzampoglou et Loukidis (2000). Investigation of the Importance of Climatic Factors in COVID-19 Worldwide Intensity. Int. J. Environ. Res. Public Health (2020) qui tente de faire la lumière sur le degré d'influence des facteurs climatiques, à savoir la température, l'humidité relative, les précipitations et la couverture nuageuse, sur l'intensité de l'épidémie de COVID-19 en examinant les données des cas et des décès rapportés à l'échelle mondiale pour la période **3/2020-5/2020** (première vague) tout en considérant simultanément d'autres facteurs importants (sociaux, économiques, réponse du gouvernement) qui devraient avoir une incidence sur l'intensité de l'épidémie.).

Cette étude avait pour but d'étudier la corrélation entre les taux de cas et de décès du COVID-19 et les facteurs climatologiques et sociodémographiques pouvant être à l'origine de cette corrélation en supposant la linéarité des prédicteurs. Notre objectif 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.

Méthodes

1. **La sélection de tous les sous-ensembles.**
   La sélection de tous les sous-ensembles ou régression par la méthode exhaustive est une approche de sélection de modèle qui consiste à d'abord tester toutes les combinaisons linéaires possibles des prédicteurs, et ensuite sélectionner le meilleur modèle en se basant sur des critères statistiques que nous décrivons plus bas. Le calcul de cette méthode peut s'avérer très long et parfois impossible surtout lorsque le nombre de prédicteurs est grand.

2. **La sélection pas à pas ascendante.**
   C'est une approche qui commence sans prédicteurs et ajoute progressivement les prédicteurs qui améliore le plus l'ajustement du modèle (grâce des critères statistiques). Elle s'arrête lorsque l'amélioration n'est plus statistiquement significative.

3. **La sélection pas à pas descendante.**
   Cette approche procède de la même manière que celle précédente mais l'ordre inverse. En effet, elle commence avec tous les prédicteurs, puis élimine progressivement les prédicteurs qui contribuent le moins à l'ajustement du modèle et s'arrête lorsque tous les prédicteurs sont statistiquement significatifs.

Les critères statistiques qui permettent la sélection du meilleur modèle dans les trois dernières approches sont :

- ✓ L'**AIC** représente le critère d'information d'Akaike, qui permet de pénaliser les modèles en fonction du nombre de paramètres afin de satisfaire le critère de parcimonie. On choisit alors le modèle avec l'AIC le plus faible (Cameron et Trivedi, 2005).
- ✓ Le **R²ajusté** représente de coefficient de détermination ajusté qui montre la proportion de variation expliqué par le modèle. Compris entre 0 et 1, plus il est proche de 1, plus le modèle est bon.
- ✓ La **P.value** représente la probabilité pour un modèle statistique donnée sous l'hypothèse nulle d'obtenir la même valeur ou une valeur encore plus extrême que celle observée (Wikipédia, 2021).

4. La régression Ridge et Lasso.

L'approche de la **régression Ridge** consiste à réduire les coefficients de régression en rajoutant une pénalité sur leur taille. Ainsi, le critère des moindres carrées avec pénalité L2 s'écrit :

$$\min_{\beta_1,\cdots,\beta_p} \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j z_{ij}\right)^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

λ (λ ≥ 0) est un paramètre (coefficient de pénalité) qui permet de contrôler l'impact de la pénalité : à fixer

Fonction de pénalité

En imposant une pénalité aux coefficients, on diminue les annulations possibles de deux prédicteurs entre eux lorsqu'il y a des prédicteurs étroitement corrélés dans un jeu de données.

La **régression lasso** quant à elle, est similaire à la régression Ridge mais avec quelque différence. Le critère des moindres carrées avec pénalité L1 s'écrit :

$$\min_{\beta_1,\cdots,\beta_p} \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j z_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

Lasso fait une sélection de variables en annulant certains coefficients : les variables associées aux coefficients nuls ne sont de facto exclues du modèle prédictif.

5. Les régressions PCR, PLS et sPLS.

Ce sont des approches basées sur des composantes principales. En effet, on effectue une régression linéaire sur les composantes retenues afin de diminuer le nombre de paramètres du modèle et (ou) la colinéarité entre les prédicteurs. Les composantes principales peuvent se construire comme suit :

- **La régression PCR** (sous Composantes principales) : Il s'agit de faire une Analyse en Composante Principale (ACP) sur la matrice des prédicteurs. On cherche donc les CP qui maximisent l'Inertie totale des observations
- **La régression PLS** (régression par les moindres carrés partiels) : La principale différence avec la PCR est qu'on ne recherche pas les composantes qui maximisent l'inertie totales des observations projetées mais ceux qui maximisent la liaison avec la variable réponse.
- **La régression sPLS** (approche parsemée des moindres carrés partiels) : Il s'agit de faire une régression PLS tout en ajoutant une pénalité Lasso.

Au final pour comparer les différentes méthodes de sélection, nous utilisons **le taux d'erreur** que nous allons estimer par **le taux d'erreur test qui est la racine carrée de la différence moyenne entre les valeurs observées et les valeurs prédites par le modèle choisir élevée au carrée**. Pour tenir de la dépendance de ce dernier à la partition aléatoire qui sera effectuée pour une stabilisation, on répète **30** fois la procédure.

Résultats

1. Description des données.

Nous disposons d'un jeu de données similaire à celles de l'étude Tzampoglou P and Loukidis D. Investigation of the Importance of Climatic Factors in COVID-19 Worldwide Intensity. Int. J. Environ. Res. Public Health (2020).

Dans ce jeu de données, nous avons **n = 55** pays et **p = 10** variables. Les variables sont tous de natures quantitatives. La variable réponse **Y = TDM** représente le nombre total des décès par million d'habitants dus au COVID-19 entre le 1/3/2020 et le 31/5/2020.

Les potentiels prédicteurs **X** sont récapitulés dans le tableau suivant :

Noms	Libellés	Description
T	Température moyenne	Indique la température moyenne (en °C)
HR	Humidité relative moyenne	Indique l'humidité relative moyenne (en %)
PR	Cumul des précipitations	Indique le cumul des précipitations (en mm)
CL	Couverture nuageuse moyenne	Indique la fraction de la couverture nuageuse moyenne
PD	Densité de population	Indique la densité de population (personnes/km2)
MA	Âge médian	Indique l'âge médian de la population
SI	Indice de "rigueur"	Indique l'indice de "rigueur" moyen qui est une mesure composite basée sur 9 indicateurs tels que les fermetures d'écoles, les fermetures de lieux de travail, les restrictions de voyage et les confinements à domicile
FM	Première mesure	Indique le délai entre le 1er cas et l'imposition des 1ères mesures (jours)
SH	Rester à la maison	Indique le délai entre le 1er cas et l'ordre de rester à la maison (jours)

Tableau 1 : Description des prédicteurs.

2. Analyse bivariée.

Dans cette section, nous cherchons à identifier les liaisons entre nos variables. Nous allons dans un premier temps utiliser la matrice de corrélation afin d'avoir une première idée de la relation des prédicteurs entre eux. Ensuite, nous traçons un nuage de points combinant la variable cible avec chacune des variables indépendantes afin de vérifier l'hypothèse de linéarité du modèle linéaire utilisé dans l'article.

2.1. Relation entre les prédicteurs.

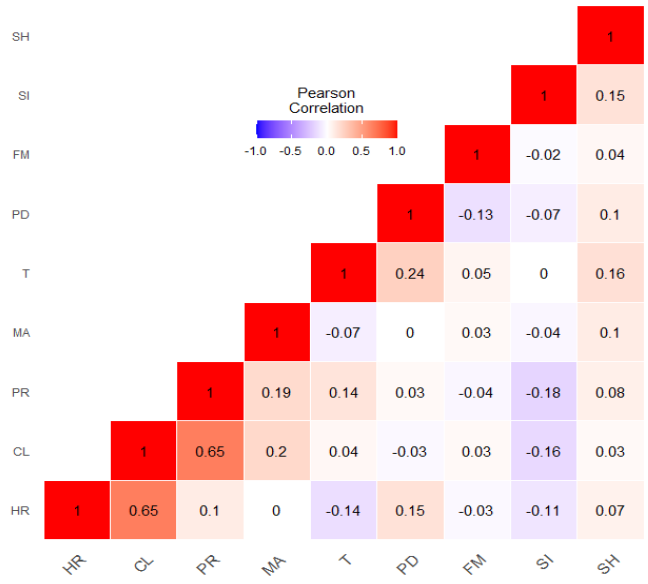


Figure 1 : Matrice de corrélation entre les prédicteurs.

Dans la **figure 1** ci-dessus, le coefficient de **corrélation de Pearson** entre deux variables identiques sont alignés dans les cases la deuxième diagonale. On parle de corrélation lorsque ce coefficient est supérieur ou égal à **0,5** en valeur absolue. Par conséquent, nous pouvons dire de manière générale qu'il y a corrélation linéaire positive entre les variables **CL**, **HR (0.65)** et les variables **PR** et **CL (0.65)**.

2.2. Lien entre chaque prédicteur et la variable réponse.

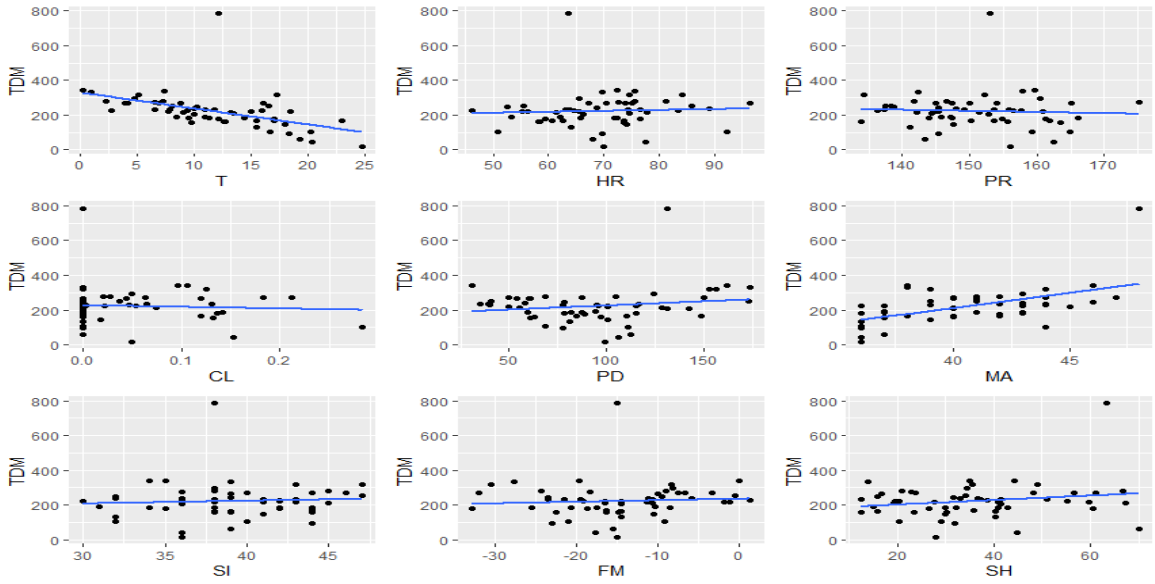
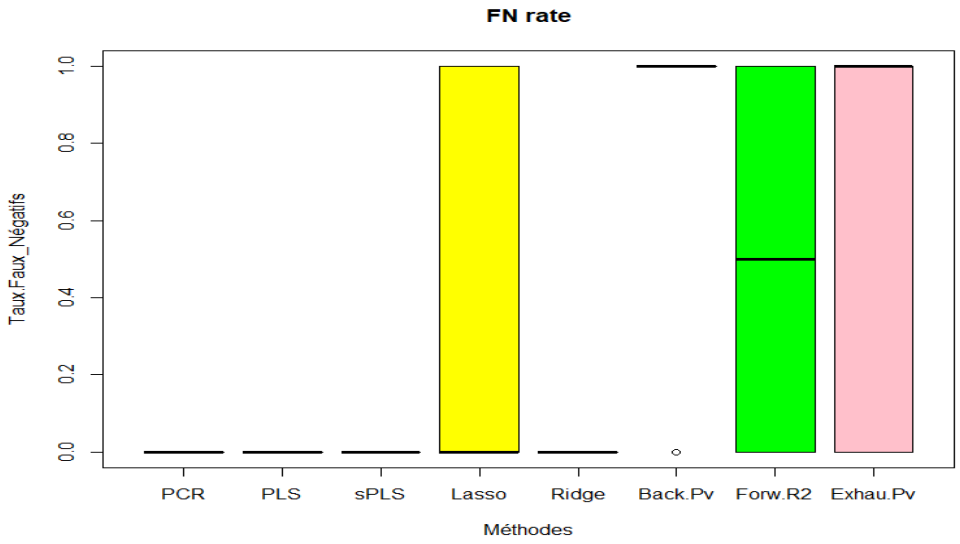


Figure 2 : Taux De Mortalité (TDM) en fonction de chaque prédicteurs.

L'analyse de la **figure 2** ci-dessus nous montre à première vue un pays très éloigné des autres dans chaque nuage. De plus, on remarque une relation affine (trait bleu) négative entre la variable **TDM** et le prédicteur **T** et une relation affine positive entre la variable **TDM** et le prédicteur **MA**. Les relations entre la variable TDM et les autres prédicteurs ne sont facilement détectables. Par conséquent, l'hypothèse de linéarité du modèle linéaire utilisé dans l'article n'est pas forcément vérifiée. Cela dit, nous admettons que la température

a un impact sur le nombre de décès de Covid-19. Le taux de faux négatifs en admettant comme vraie l'association avec la température est représentée dans la **figure 3** ci-dessous :



Figures 3 : Taux de faux négatifs en fonction des 8 méthodes de sélection.

L'analyse de ce dernier nous montre que les méthodes qui réussissent à retenir cette variable sont : **PCR, PLS, sPLS, Ridge et pas à pas descendante avec critère P. Value**.

3. Analyse multivariée.

3.1. Comparaison de critère pour les méthodes exhaustive, forward et backward.

	AIC			R^2ajusté			P.value		
	Exhau	Forward	Backward	Exhau	Forward	Backward	Exhau	Forward	Backward
Mean_err	**0,2**	**0,1**	**0,2**	**0,1**	**0,1**	**0,2**	**0,1**	**0,1**	**0,1**

Tableau 2 : Critères AIC, R^2Ajusté et P. value, et erreurs moyennes des méthodes associées.

D'après le **tableau 2** ci-dessus, on peut retenir soit le critère **R^2a**, soit le critère **P.v** pour la méthode de sélection de tous les sous-ensembles (Exhaustive) ; les trois critères donnent les mêmes erreurs moyennes pour la méthode de sélection pas à pas ascendante (Forward). Le choix final du critère de ces deux méthodes se fera sur la **figure 3** ci-dessous. La méthode de sélection pas à pas descendant (Backward) quant à lui sera associée au critère **P.v.**

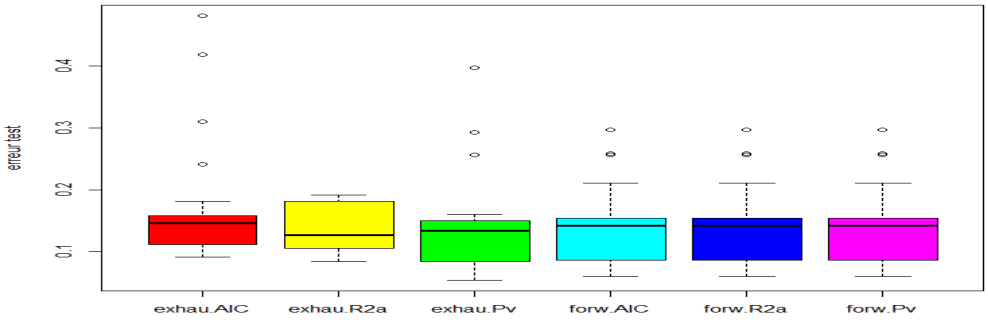


Figure 4 : Erreur test en fonction des méthodes exhaustive et forward.

L'analyse de la figure ci-dessus, nous permet de choisir le critère p.value **P.v** pour la méthode exhaustive. Cependant, les erreurs de la méthode pas à pas ascendante pour les trois critères sont distribuées de façon identique. On fait le choix du critère **R^2ajusté**.

3.2. Comparaisons des 8 méthodes de sélection de variables.

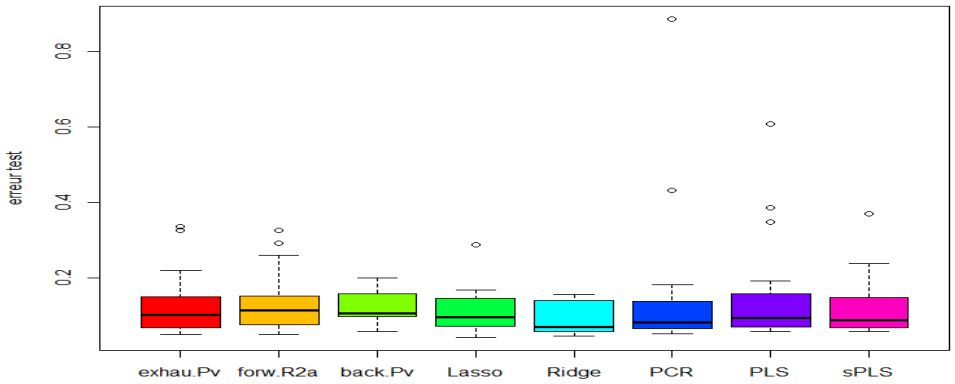


Figure 5 : Erreur test en fonction des 8 méthodes de sélection.

Conclusion

L'analyse de la **figure 4** ci-dessus, nous révèle que la **régression Ridge** est la meilleure méthode de sélection car la distribution de ces erreurs de prédiction notamment celle de la **médiane** est inférieure à celle des autres méthodes. Le fait qu'elle tienne compte de la colinéarité entre les prédicteurs (**figure 1**) renforce le choix de ce modèle. Ensuite, toutes les méthodes n'arrivent pas à sélectionner la variable température **T** (**figure. 3**) qui s'avère être liée au taux de mortalité **TDM** (**figure. 2**) ; ce qui prouve que toutes les méthodes comme **Lasso**, ne sont pas pertinentes. Enfin, l'utilisation de la régression Ridge et la prise en compte de la non linéarité pourrait améliorer le travail de Tzampoglou et Loukidis (2000).

Bibliographie

Hirotugu Akaike (1973), *Information theory and an extension of the maximum likelihood principle* dans *Second International Symposium on Information Theory*, 267-281 p.

Castro, Yohann (2011), *Constructions déterministes pour la régression parcimonieuse*, Toulouse 3, TEL.

Michy Alice (2016), *Principal Components Regression (PCR) in R*, Electrical Engineering student at Polytechnic University of Milan and graduate from University of Genoa, July 21.