MRR2024 Proj127 Partie 2 : Mortalité Hospitalière et Caractéristiques des Patients

Bastien Mousse et Amine Ould Hocine

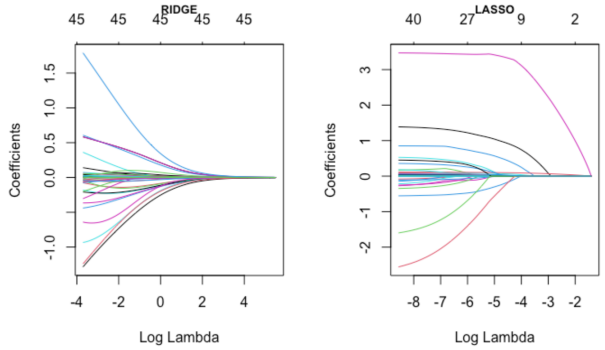
1 Modèle complet

Pour débuter, nous avons effectué une modélisation logistique incluant toutes les variables du jeu de données initialement détaillé dans le premier rapport. Rappelons que les variables prédictives issues d'autres modèles ont été exclues, tout comme les colonnes contenant plus de 20% de valeurs manquantes. Pour les autres variables, les valeurs manquantes ont été imputées : par la moyenne dans le cas de données numériques, et par la modalité la plus fréquente pour les variables catégoriques ou qualitatives. Cette stratégie permet d'exploiter pleinement les données disponibles et d'évaluer la pertinence des différentes variables dans le modèle. Certaines variables, notamment *age*, *dzgroup*, *avtisst* et *hday*, présentent des p-values inférieures à $\alpha = 0.5\%$, ce qui motive leur inclusion pour des analyses approfondies.

2 Problématique

Avant la modélisation, les données ont été réparties de manière aléatoire entre un ensemble d'apprentissage (80% des données) et un ensemble de test (20%). Cette stratégie vise à maintenir un équilibre dans les proportions des classes au sein des deux échantillons, ce qui est essentiel pour garantir une construction efficace des modèles. Dans l'ensemble d'apprentissage, environ 74,25% des patients appartiennent à la catégorie ayant survécu à l'hospitalisation, tandis que 25,75% représentent ceux décédés. De même, l'ensemble de test présente des proportions similaires, avec respectivement 73,44% et 26,56%. Ces répartitions équilibrées permettent d'assurer une modélisation robuste et une évaluation fiable.

3 Modélisation



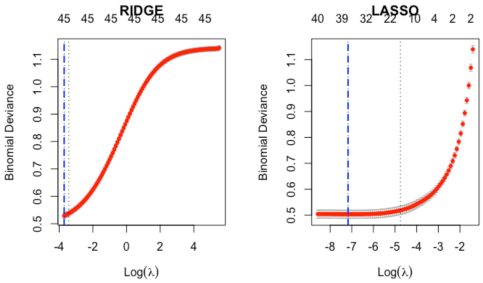
Les valeurs de $\log(\lambda)$ diffèrent selon les modélisations, avec des valeurs généralement plus élevées pour la méthode Ridge par rapport à Lasso.
Néanmoins, une tendance commune peut être observée : lorsque λ augmente, les coefficients tendent progressivement vers zéro.

Cette caractéristique particulière du paramètre λ souligne son rôle : à mesure que sa valeur augmente, le nombre de coefficients significatifs diminue. Une différence notable entre Ridge et Lasso se manifeste dans la convergence des coefficients. Avec Ridge, ces derniers convergent tous approximativement au même λ très élevé, alors que dans le cas de Lasso, ils convergent vers zéro pour des λ totalement différents.

4 Validation croisée

La validation croisée va permettre de partitionner les données en plusieurs sous-ensembles grâce à la méthode k-folds. Cette méthode divise les données en plis (folds) de taille égale. On utilisera les données d'apprentissage pour cela.

Les λ optimaux en fonction de la méthode ne sont pas du tout les mêmes :



- Pour Ridge, le λ optimal est d'environ 0.026, qui est également l'un des plus petits λ pris en compte dans la modélisation Ridge initiale.
- Pour Lasso, le λ optimal est d'environ 0.00077 et n'est pas totalement l'un des plus petits pris en compte dans la modélisation, comme on peut le voir sur le graphique.

5 Performances

Matrice de confusion

RIDGE

	Réel 0	Réel 1	Sum
Prédiction 0	1222	114	1336
Prédiction 1	75	355	430
Sum	1297	469	1766

- Vrais Positifs (VP) : 355 patients survivants correctement prédites.
- Vrais Négatifs (VN) : 1222 patients décédés bien prédits.
- Faux Positifs (FP) : 75 patients survivants mal prédits.
- Faux Négatifs (FN) : 114 patients décédés mal prédits.

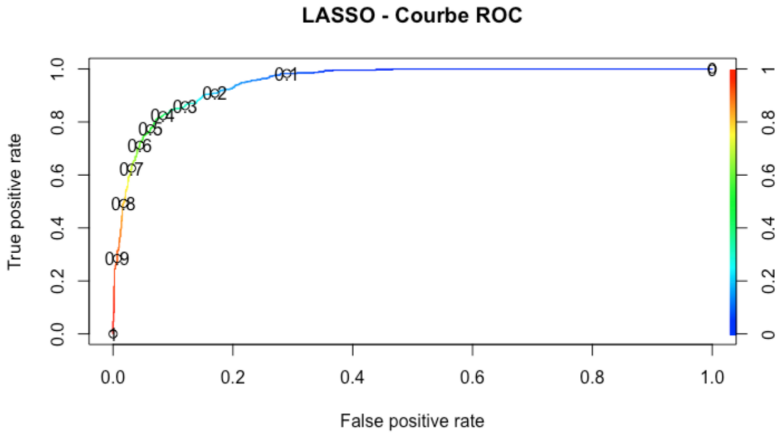
LASSO

	Réel 0	Réel 1	Sum
Prédiction 0	1216	106	1322
Prédiction 1	81	363	444
Sum	1297	469	1766

- Vrais Positifs (VP) : 1216 patients survivants correctement prédits.
- Vrais Négatifs (VN) : 363 patients décédés bien prédits.
- Faux Positifs (FP) : 81 patients survivants mal prédits.
- Faux Négatifs (FN) : 106 patients décédés mal prédits.

Courbe ROC et AUC

L'analyse comparative entre les modèles RIDGE et LASSO indique une nette supériorité de LASSO. L'AUC (Surface sous la courbe) plus élevée pour LASSO témoigne de sa meilleure capacité discriminante, indiquant une forte aptitude à distinguer les patients survivants à l'hospitalisation. En résumé, les performances globales et les indicateurs d'efficacité suggèrent que le modèle de régression LASSO est plus adapté pour prédire la variable cible dans notre jeu de données.



- RIDGE - AUC sur l'ensemble de test : 0.9481927 ;
- LASSO - AUC sur l'ensemble de test : 0.9503973.

Le seuil semblant le plus pertinent est de 0.5