Régression logistiquemultiple et prédiction des facteurs de risque concernant la survie en soins intensifs.

changer le titre

Julia Guerra [1], Maxime Jaunatre [2], Ellie Tideswell [3] | Master 2 BEE Grenoble
Mail [1], Mail [2] Mail [3]| 25 novembre 2019

Todo list

Présentation des données

Les données sont issues d'un échantillons de 200 patients d'hopitaux états-uniens, extrait d'une étude portant sur la survie des patients à l'issue d'un séjour en service de soins intensifs. L'étude en question propose 20 variables mesurées pour 200 patients. Ces variables sont très diverses et comprennent la survie, divers paramètres régissant leur entrée dans le service et d'autres paramètres physiologiques. Toutes les variables sont discrètes à l'exeption de l'âge, la pression systolique (mm Hg) et le rythme cardiaque (battement/min.) à l'admission. Aucune valeur n'est manquante et les variables continues ne présentent pas de valeurs abérrantes. Il est cependant remarquable que de nombreuses variables qualitatives sont composées de 2 classes déséquilibrées (Figure 1).

Alors, avant la réalisation d'un possible modèle expliquant les observations, il sera néccessaire de limiter les biais introduits par ces variables. On peut donc dans un premier temps supprimer la variable 'RACE', car l'assignation des individus ne repose pas sur une mesure précise. Afin d'écarter d'autres variables qualitatives, leurs corrélations ont été évalués au moyen d'un test de χ^2 par paires de variables.

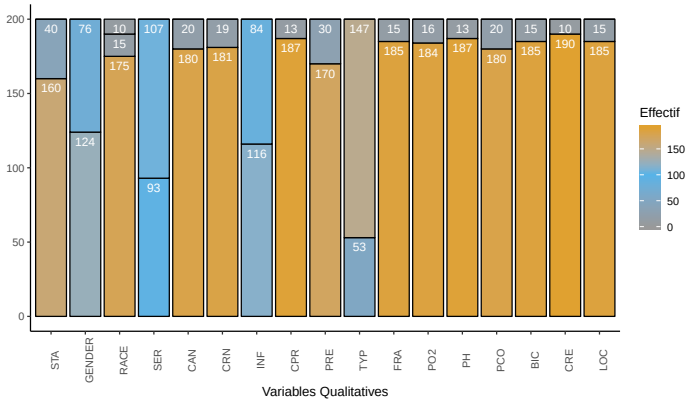

FIGURE 1 – **Répartition des effectifs dans les classes de variables qualitatives.**

Il apparaît qu'effectuer ces tests de χ^2 sur un plus petit jeu de données entraîne une dimition du nombre de corrélations. Pour les corrélations restantes après un échantillonnage, la variable 'SER' est corrélée au plus d'autres variables. De même 'PCO' est corrélée à 4 autres variables (Table 1). Il existe d'autres variables corrélées, mais il semble préférable de limiter le nombre de variables explicatives et de favoriser le choix de variables corrélées à plusieurs autres et dont les classes sont les moins déséquilibrées possibles (Figure 1). Parmis les variables qualitatives seront donc conservés : 'SER', 'INF', 'PRE', 'TYP'

	GENDER	SER	CAN	CRN	INF	CPR	PRE	TYP	FRA	PO2	PH	PCO	BIC	CRE	LOC
GENDER	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
SER	-	NA	0.0035	0.05	0.003	0.002	NA	5e-04	0.0145	0.003	0.001	0.027	5e-04	0.044	NA
CAN	-	-	NA	NA	NA	NA	NA	5e-04	NA	NA	NA	NA	NA	NA	NA
CRN	-	-	-	NA	NA	0.0235	NA	NA	NA	NA	0.023	NA	0.0355	0.002	0.034
INF	-	-	-	-	NA	0.041	0.0155	0.0225	NA	0.007	0.01	0.034	0.0135	NA	NA
CPR	-	-	-	-	-	NA	NA	0.0405	NA	NA	NA	NA	NA	NA	0.001
PRE	-	-	-	-	-	-	NA	NA	NA	NA	NA	NA	NA	NA	NA
TYP	-	-	-	-	-	-	-	NA	NA	NA	NA	NA	0.035	NA	NA
FRA	-	-	-	-	-	-	-	-	NA	NA	NA	NA	NA	NA	NA
PO2	-	-	-	-	-	-	-	-	-	NA	0.002	5e-04	0.026	NA	NA
PH	-	-	-	-	-	-	-	-	-	-	NA	5e-04	0.002	NA	NA
PCO	-	-	-	-	-	-	-	-	-	-	-	NA	NA	NA	NA
BIC	-	-	-	-	-	-	-	-	-	-	-	-	NA	0.003	NA
CRE	-	-	-	-	-	-	-	-	-	-	-	-	-	NA	NA
LOC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	NA

TABLE 1 – **P-valeurs des tests de χ^2 par paires de variables qualitatives sur 70% du jeu de donnée.** Les cases avec 'NA' indiquent des valeurs > 0.05.

et 'PCO'.

Afin de limiter le nombre de variables, une analyse de corrélations entre les variables quantitatives est aussi réalisé. Les deux variables 'HRA' (Rythme cardiaque) et 'SYS' (Pression systolique) ne présentent pas des répartitions suivant une loi normale, avec une P-valeur au test de Shapiro de respectivement 3.1 et 3.1, permettant de rejeter H0. Un test de corrélation de Kendall ne permet pas de souligner une corrélation entre les deux variables, avec une p-valeur de 3.1.

À vue des coïncidences entre variables, les modèles considerés partiront de sept variables pour expliquer la variable binomiale 'STA', qui indique la survie ou non du patient à l'issue de son séjour en service de soins intensifs. Ces variables sont : le sexe du patient ('GENDER'), son âge ('AGE'), la raison de son admission -médicale / chirurgicale- ('SER'), la présence d'une infection lors de l'admission ('INF'), une précédente admission dans les derniers 6 mois ('PRE'), la nature de l'admission -prévue / urgence- ('TYP') et la pression en dioxide de carbone dans le sang à l'admission ('PCO'). Le choix du meilleur modèle sera d'abord effectué sur un sample du 70% du jeu de données, ensuite généralisés à la totalité des données.

Modélisation

La selection de modèle avec plusieurs variables se fait par selection stepwise, au moyen de la fonction `stepAIC` du package `MASS`. Cette analyse est portée sur un jeu de donnée échantillonné ainsi que sur l'ensemble du jeu de données. Dans les deux cas, les variables choisies comme celles les plus discriminantes pour la survie sont l'âge de l'individu, la pression systolique lors de l'admission et le type d'admission.

Des analyses successifs du test de Wald et du test de rapport de vraisemblance confirment que les variables 'AGE', 'SYS' ou 'TYP' séparémment ne servent pas à expliquer assez de déviation des résidus : le modèle avec les trois variables ensemble est significativement meilleur que les trois modèles n'incluant que l'une d'elles à la fois, avec des moindres valeurs de déviation des résidus pour toutes les trois.

	AGE	SYS	TYP
Dév. modèle 1 variables	133.19	132.73	123.11
Dév. modèle 3 variables	104.31	104.31	104.31

TABLE 2 – Déviation des résidus selon les modèles utilisés pour chacune des variables dans le cas du jeu de donnée échantillonné.

De cette manière, on peut conclure que le modèle doit intégrer ces trois variables pour décrire la

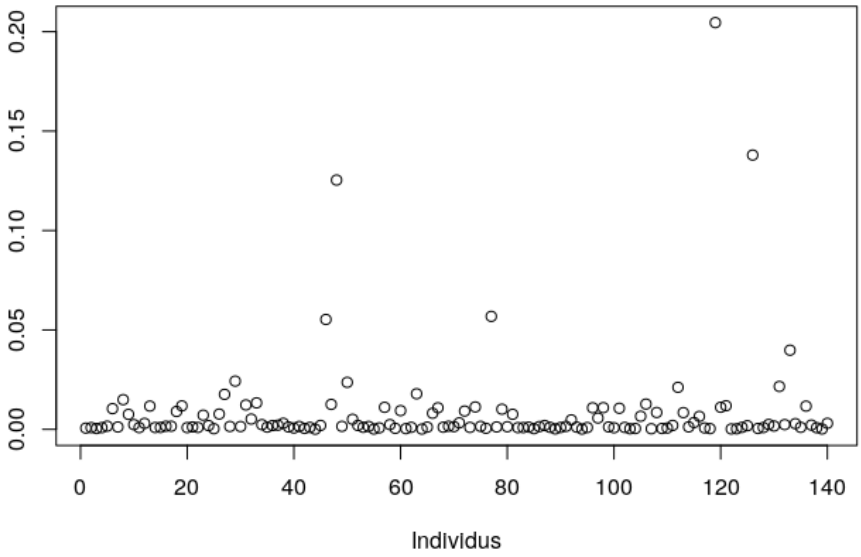

FIGURE 2 – Répartitions des distances de Cook pour le jeu de données échantillonné.

plupart des données. Le meilleur modèle à aborder est donc : `STA ~ TYP + AGE * SYS` .

Validation du modèle

Le test de Hosmer & Lemeshow soutient l'hypothèse que le modèle est compatible avec les données du sample 70%, même si la p-valeur est proche de 0.05 (3.1). La distance de Cook ne montre pas de points influents, ce qui nous fait accepter les observations du dataset.

Performance du modèle

La surface en dessous de la courbe ROC (dîte valeur AUC) est de 3.1. Ce modèle ne semble pas être très performant, vu la grande surface en dessous de la courbe et la valeur de sensibilité de 3.1, laquelle fait preuve du grand nombre de faux positifs. Les valeurs de AUC et de sensibilité s'améliorent relativement pour la totalité du jeu de données, avec un AUC de 3.1 et une sensibilité de 3.1.

Le choix du modèle était bien basé vu que, pour le sample 70%, les modèles contenant les trois variables 'AGE', 'TYP' et 'SYS' servaient à expliquer plus de variabilité des données que ceux ne contenant qu'une d'elles. C'est aussi le cas si l'on teste ces mêmes conclusions sur le jeu de données complet.

	AGE	SYS	TYP
Dév. modèle 1 variables	192.31	191.34	185.05
Dév. modèle 3 variables	167.82	167.82	167.82

TABLE 3 – Déviation des résidus selon les modèles utilisés pour chacune des variables dans le cas du jeu de donnée entier.

C'est vrai que l'approche stepwise pour le dataset complet propose des effets croisés entre l'âge et la pression systolique comme explication d'une part de la variabilité des données. Une fois ajouté le terme

des effets croisés, les valeurs de AUC et de sensibilité s'améliorent relativement avec un AUC de 3.1 et une sensibilité de 3.1.

Conclusion

L'analyse permet donc de faire ressortir des variables intéressantes sur la survie des patients aux services de soins intensifs. Cependant cette analyse n'est pas exhaustive et la faible différence de modèles avant et après échantillonnage questionne sur la qualité du jeu de donnée, et sa taille réduite. Il faut également remarquer que le modèle proposé est réduit à une régression logistique et que d'autres interractions peuvent avoir leurs importances dans la survie du patient.

Bibliographie

R.Team. 2017. R : A language and environment for statistical computing (Version 3.4. 2)[Computer software]. *Vienna, Austria : R Foundation for Statistical Computing.*

Ressources

Ce document est disponible en ligne sous format ".Rnw", contenant tout le code néccessaire à la reproduction de l'analyse, réalisée avec un script en langage R (R.Team, 2017), ainsi que le jeu de données de départ. L'ensemble est situé sur Github : `https://github.com/gowachin/GIS`