INSA – GM 5 Année 2018-2019

ANTOUN Cyril
ROUFFIAC Jean-Eudes

Régression Non Linéaire avec Applications sous R

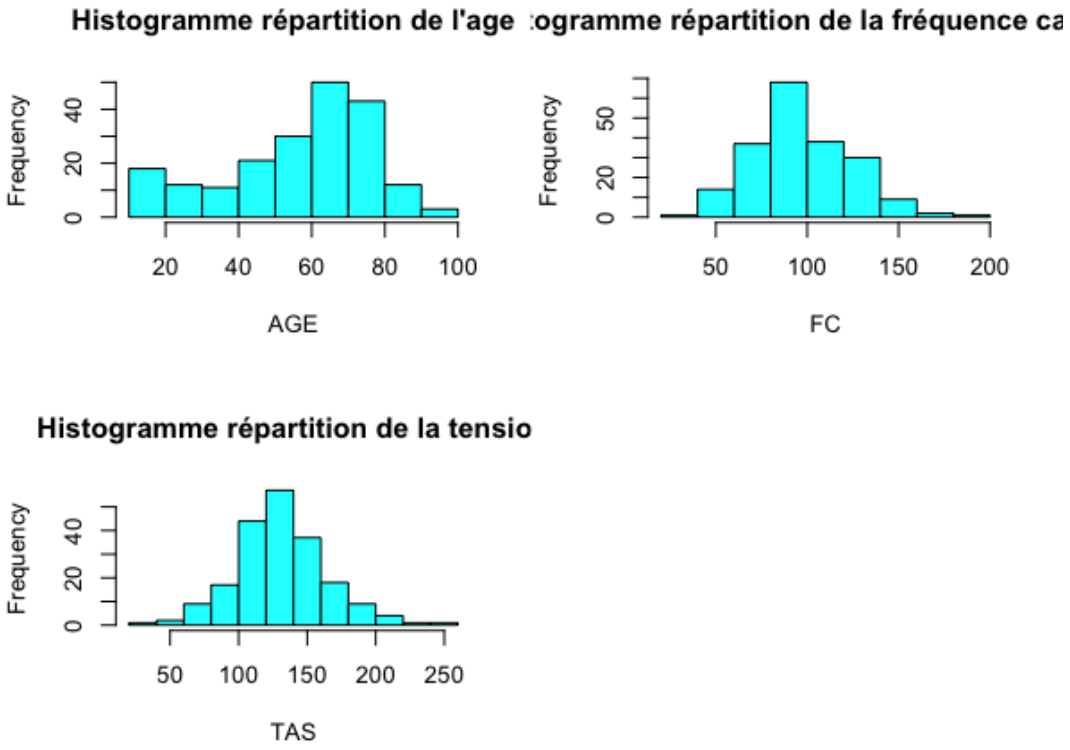
Rapport de TP n°7

Titre : *Régression logistique sur variables quantitatives*

L'objectif de ce TP est d'expérimenter, avec la fonction glm, quelques techniques statistiques dans le cadre des modèles de régression logistiques binaires. On cherche à expliquer le statut vital, mort ou vivant, de patients. Pour ce faire on dispose des 200 données multidimentionnelles relevées sur les variables suivantes :

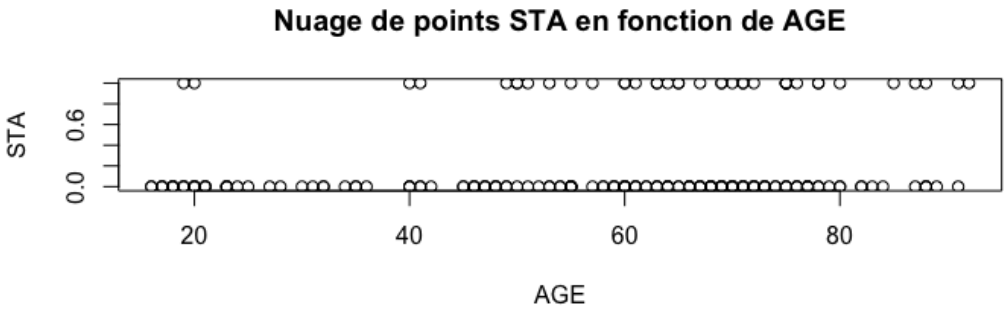
- STA : Statut vital (0=vivant, 1=mort)
- SER : Service lors de l'admission (0=medical, 1=chirurgical)
- SEX : (1 = homme, 2 = femme)
- RACE : (1 = blanc, 2 = noir, 3 = autre)
- IRC : Antécédents d'insuffisance rénale chronique (0=non, 1=oui)
- INF : Infection probable à l'admission (0=non, 1=oui)
- MCE : Massage cardiaque avant l'admission (0=non, 1=oui)
- CAN : Un cancer fait-il parti du problème ? (0 = Non, 1 = Oui)
- TYP : Type d'admission (0=normal, 1=urgence)
- ATC : Antécédents d'admission en USI dans les derniers 6 mois (0 = Non, 1 = Oui)
- FRA : Fracture (0 = Non, 1 = Oui)
- PO2 : PO2 de la gazométrie artérielle initiale (0 si > 60, 1 ≤ 60)
- PH : PH de la gazométrie artérielle initiale (0 si $\geq$ 7,25, 1 < 7.25)
- PCO : PCO2 de la gazométrie artérielle initiale (0 si ≤ 45, 1 > 45)
- BIC : Bicarbonate de la gazométrie artérielle initiale (0 si ≥ 18, 1 si < 18)
- CRE : Créatinine du prélèvement initial (0 si ≤ 2, 1 si > 2)
- CS : Niveau de conscience à l'admission (0=pas de coma ni stupeur, 1=stupeur, 2=coma)
- AGE : Age en années
- TAS : Tension artérielle systolique à l'admission en mmHg
- FC : Fréquence cardiaque à l'admission en USI (battements/min)

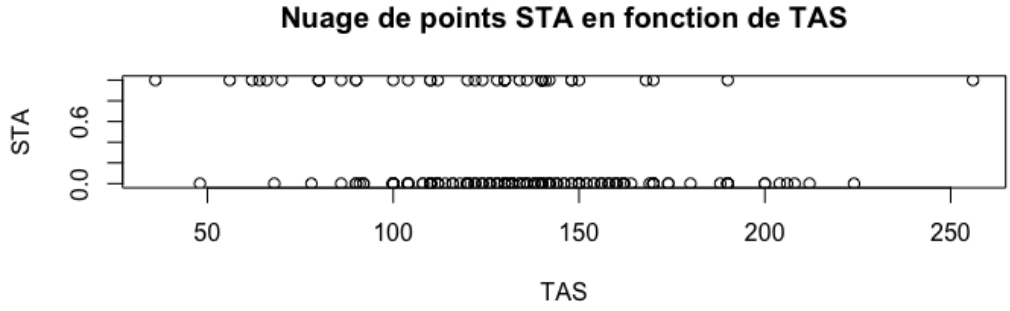
On commence par une étude univariée sur chacune des variables. On trace l'histogramme de répartition pour les trois variables quantitatives : âge, tension artérielle systolique et fréquence cardiaque. On obtient :

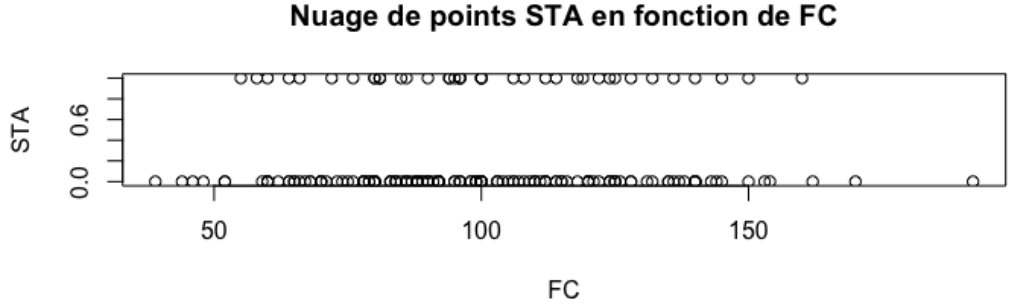


On remarque qu'un grand nombre de patients admis ont plus de 60 ans. Les tensions artérielles systoliques des patients sont majoritairement situées entre 100 et 150 mmHg et les fréquences cardiaques sont plutôt réparties autour de 100 bpm.

On trace maintenant les nuages de points pour ces trois mêmes variables quantitatives :







Les nuages de points ont une forme atypique. On observe que pour un âge élevé, supérieur à 50 ans, la proportion de décès est élevée. Alors qu'entre 0 et 40 ans on ne relève pratiquement aucun cas de décès. Concernant la tension artérielle systolique, lorsqu'elle est faible on observe beaucoup de décès et quasiment aucun lorqu'elle est supérieure à 150 mmHg. On conclut donc que ces deux variables quantitatives ont une influence sur le statut vital. Contrairement à la fréquence cardiaque. En effet pour cette variable on n'observe pas de comportement spécifique du statut vital pour une faible ou une forte fréquence cardiaque. Ainsi graphiquement on peut déduire que la variable FC n'influe pas sur la variable à expliquer STA.

On souhaite maintenant éliminer les variables non significatives du modèle. Par conséquent, on met en place un test du chi-deux. Les résultats sont indiqués dans le tableau suivant :

Variables	ID	SEX	RAC	SER	CAN	IRC	INF	MCE
Test Chi-deux	0,47	0,91	0,40	0,01	1	0,02	0,02	0,005

Variables	ATC	TYP	FRA	PO2	PH	PCO	BIC	CRE	CS
Test Chi-deux	0,8	0,001	1	0,4	0,52	1	0,31	0,043	$1{,}09*10^{-10}$

On décide de conserver les variables dont la p-value est inférieure à 5%. Ainsi nous supprimons au risque 5% les variables qualitatives suivantes : ID, SEX, RAC, CAN, ATC,

FRA, PO2, PH, PCO, BIC. On supprime également la variable quantitative FC qui n'influe pas sur STA.

On peut alors introduire un premier modèle de régression logistique qu'on appelle modèle complet du statisticien et que l'on note ($M_{S,comp}$).
Nous calculons les coefficients estimés du modèle et nous obtenons les tableaux résultats suivants :

Coefficient	Estimate	Std Error	z value	Pr(> |z|)
(Intercept	-5.667411	1.637457	-3.461	0.000538 ***
AGE	0.029164	0.012292	2.373	0.017661 *
TAS	-0.009061	0.006976	-1.299	0.194029
SER	-0.127594	0.482626	-0.264	0.791492
IRC	0.456918	0.664667	0.687	0.491806
INF	0.211498	0.460439	0.459	0.645991
MCE	0.515000	0.835857	0.616	0.537806
TYP	1.784387	0.826068	2.160	0.030765 *
CRE	0.518736	0.837440	0.619	0.535632
CS	1.746042	0.578524	3.018	0.002544 **

Null deviance	200.16	199 degree of freedom
Residual deviance	146.69	190 degree of freedom
AIC	166.69	

La deviance du modèle est de 146.69 contre 200.16 pour la déviance nulle. La déviance résiduelle est considérablement plus faible ce qui implique que les variables expliquent bien le statut vital. Par ailleurs, le nombre d'itérations est de 6. De plus, on observe que certaines variables possèdent des p-values importantes. Par conséquent on cherche à estimer l'influence de chaque variable sur le statut vital des clients. L'analyse de la deviance nous donne les informations suivantes :

	Df	Deviance Resid.	Df Resid.	Dev
NULL			199	200.16
AGE	1	7.8546	198	192.31
TAS	1	9.0513	197	183.25
SER	1	5.6043	196	177.65
IRC	1	3.5872	195	174.06
INF	1	0.8700	194	173.19
MCE	1	3.1053	193	170.09
TYP	1	8.3356	192	161.75
CRE	1	0.2873	191	161.47
CS	1	14.7761	190	146.69

Grâce à ce tableau, on peut observer l'importance de chaque variable sur la déviance finale. Ainsi les variables CS, AGE et TAS sont celles qui réduisent le plus la déviance résiduelle alors que CRE ou INF sont celles qui la réduisent le moins. On se rend compte alors que certaines variables ont peu d'influence sur la variable à expliquer et sont alors dispensables. On crée donc un nouveau modèle à l'aide d'une méthode descendante. On élimine une à une les variables jugées non significatives au risque 5%. On enlève du modèle les variables suivantes :

- étape 1 : SER. p-value = 0.79
- étape 2 : INF. p-value = 0.62
- étape 3 : CRE. p-value = 0.52
- étape 4 : MCE. p-value = 0.49
- étape 5 : TRC. p-value = 0.32
- étape 6 : TAS. p-value = 0.15

On obtient alors le modèle du statisticien réduit que l'on note ($M_{S,red}$) et qui utilise les variables AGE, TYP et CS pour expliquer le statut vital des patients. On calcule à nouveau les coefficients estimés du modèle. On obtient les tableaux résultats suivants :

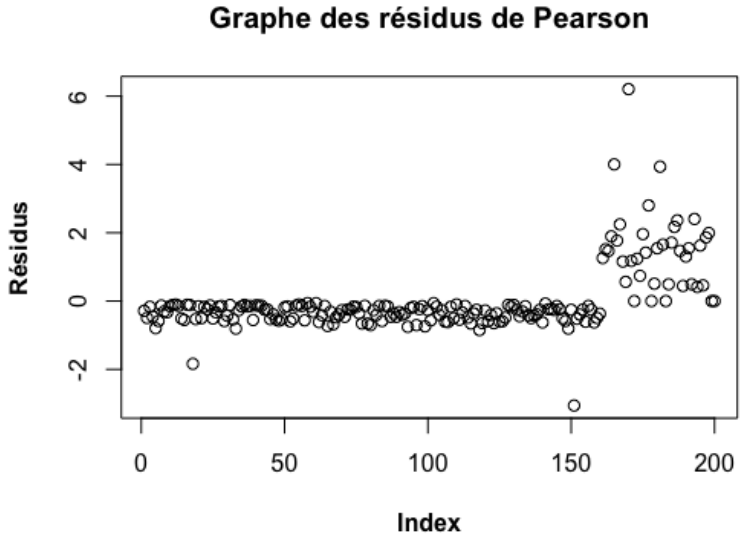
Coefficient	Estimate	Std Error	z value	Pr($>\|z\|$)
(Intercept	-7.35509	1.23929	-5.935	2.94e-09 ***
AGE	0.03291	0.01179	2.791	0.005247 **
TYP	2.18842	0.76276	2.869	0.004117 **
CS	1.83445	0.51609	3.555	0.000379 ***

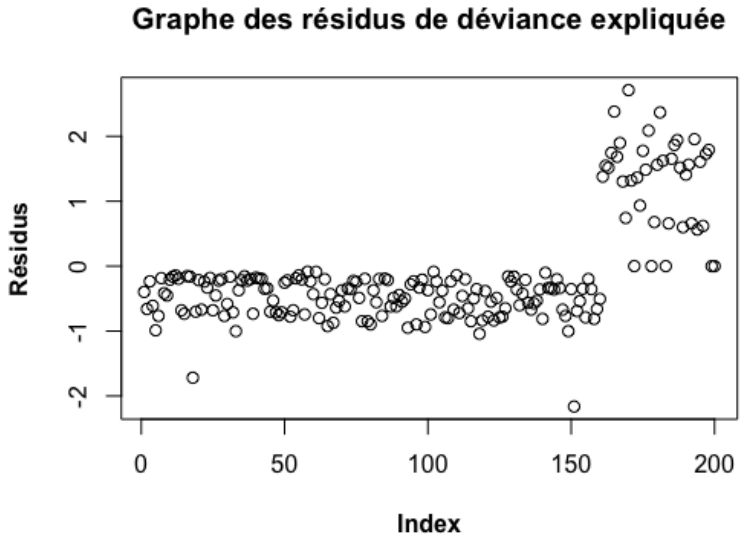
Null deviance	200.16	199 degree of freedom
Residual deviance	151.02	196 degree of freedom
AIC	159.02	

Ce nouveau modèle nous donne une deviance résiduelle de 151.02. Elle est proche de celle trouvée pour le précedent modèle et est supérieure de seulement 4.33. Cela malgré le passage de 9 à 3 variables explicatives. De plus, on remarque également que l'AIC du ($M_{S,red}$) est plus faible que celui du ($M_{S,comp}$). Cela est dû au fait que l'AIC prend en compte le nombre de variables du modèle. Enfin le nombre d'itérations est également de 6.

En conclusion on peut dire que le modèle complet n'est que légèrement plus performant au prix d'un nombre de variables beaucoup plus élevé. Il apparait plus intéressant de choisir le modèle réduit.

On étudie maintenant les graphiques des résidus de Pearson et de déviance associés à ce modèle réduit.





Les résidus de la déviance sont plus élevés que ceux de Pearson. Dans les deux cas, on remarque que les résidus sont très proches de zéro lorsque le statut vital vaut 0. En revanche, la valeur de ces résidus augmente lorsque le statut vital du patient devient 1.

Dans les deux modèles précédents, la variable CS est considérée comme une variable quantitative. On reprend ce modèle mais en considérant CS comme une variable qualitative. On note ce modèle $M_{S,qual}$. On calcule à nouveau les coefficients estimés du modèle et on obtient :

Coefficient	Estimate	Std Error	z value	Pr(> |z|)
(Intercept	-6.20071	1.30990	-4.734	2.2e-06 ***
AGE	0.03399	0.01204	2.824	0.00474 **
TYP	2.78080	1.03875	2.677	0.00743 **
CS2	18.93903	990.09093	0.019	0.98474
CS3	2.63052	0.83237	3.160	0.00158 **

Null deviance	200.16	199 degree of freedom
Residual deviance	141.87	195 degree of freedom
AIC	151.87	

On remarque tout de suite que la deviance du nouveau modèle est inférieure à celle des deux premiers modèles. On peut faire la même observation pour l'AIC. A priori, l'étude des données en considérant CS comme variable qualitative permet d'obtenir un modèle de bien meilleure qualité. On note en revanche que le nombre d'itérations est passé à 15.

On apprend que les médecins utilisent un modèle différent, mettant en jeu les variables AGE, CAN, IRC, INF, TAS et CS. On étudie donc ce modèle que l'on note ($M_{M,comp}$). Par le biais d'une analyse descendante on supprime les variables les moins significatives au risque 5%. On obtient le nouveau modèle ($M_{M,red}$) utilisant les variables AGE, TAS et CS. On compare les deux modèles dans le tableau suivant :

Modèle	($M_{M,comp}$)	($M_{M,red}$)
Déviance résiduelle	155.67	160.27
AIC	169.67	168.27
Nb itérations	5	5

Au vu des déviances résiduelles calculées et du nombre d'itérations trouvé, ces modèles du médecin semblent de moins bonne qualité que les modèles du statisticien étudiés précédemment.

Lors de ce TP, nous avons travaillé sur 5 modèles. Ce n'était pas demandé dans le TP, mais par curiosité nous avons ajouté un sixième modèle à notre étude. Ce modèle qu'on note $M_{T,red}$ est trouvé à la suite d'une méthode descendante faite sur l'ensemble des variables. Ce modèle met en jeu les 7 variables suivantes : AGE, TYP, CS, ID, CAN, PH et PCO. Nous présenterons les performances de ce modèle dans un tableau de confusion et un tableau de résultat qui vont suivre. En effet, pour évaluer les performances des modèles de manière empirique, on peut construire la matrice de confusion qui comptabilise les bonnes et mauvaises prédictions. On fixe le seuil de décision à 0.5 et on obtient les matrices de confusion suivantes :

$M_{S,comp}$		Prévus	
		0	1
Observés	0	157	3
	1	28	12

$M_{S,reduit}$		Prévus	
		0	1
Observés	0	158	2
	1	28	12

$M_{S,qual}$		Prévus	
		0	1
Observés	0	158	2
	1	27	13

$M_{M,comp}$		Prévus	
		0	1
Observés	0	157	3
	1	30	10

$M_{M,reduit}$		Prévus	
		0	1
Observés	0	158	2
	1	29	11

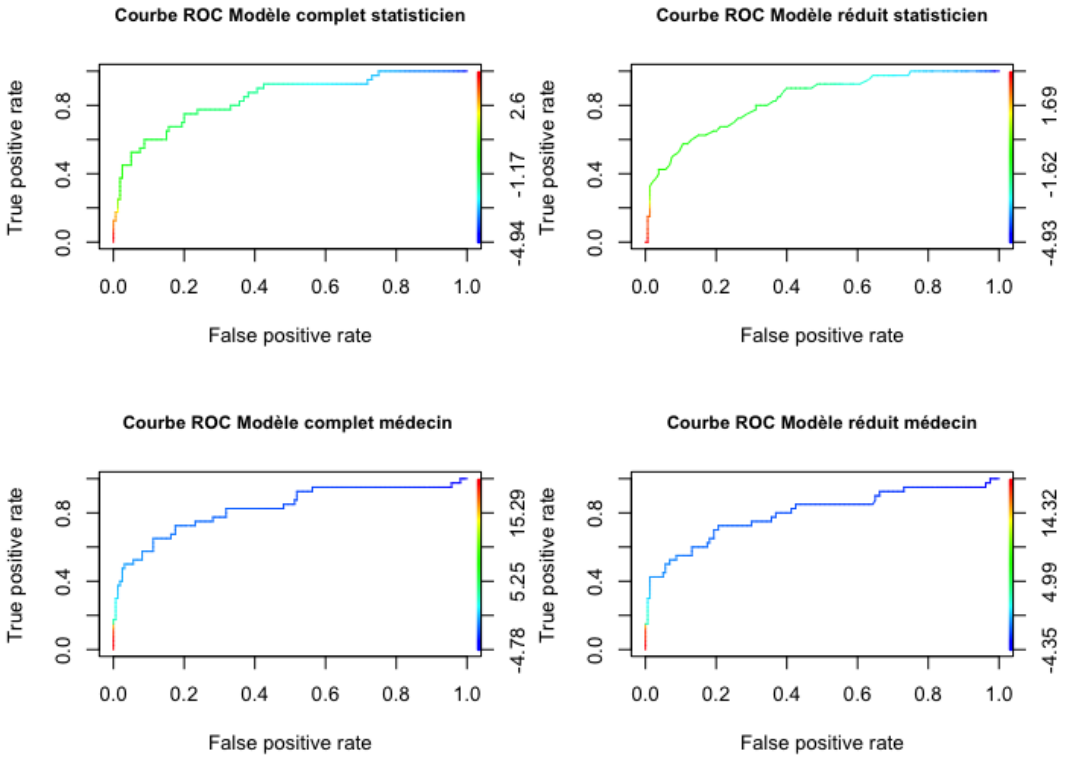
$M_{T,reduit}$		Prévus	
		0	1
Observés	0	154	6
	1	23	17

On peut alors calculer le taux d'erreurs, la spécificité et la sensibilité. Nous résumons alors les résultats trouvés pour chaque modèle dans un tableau en ajoutant les indicateurs cités précedemment. On obtient alors le tableau de résultat suivant :

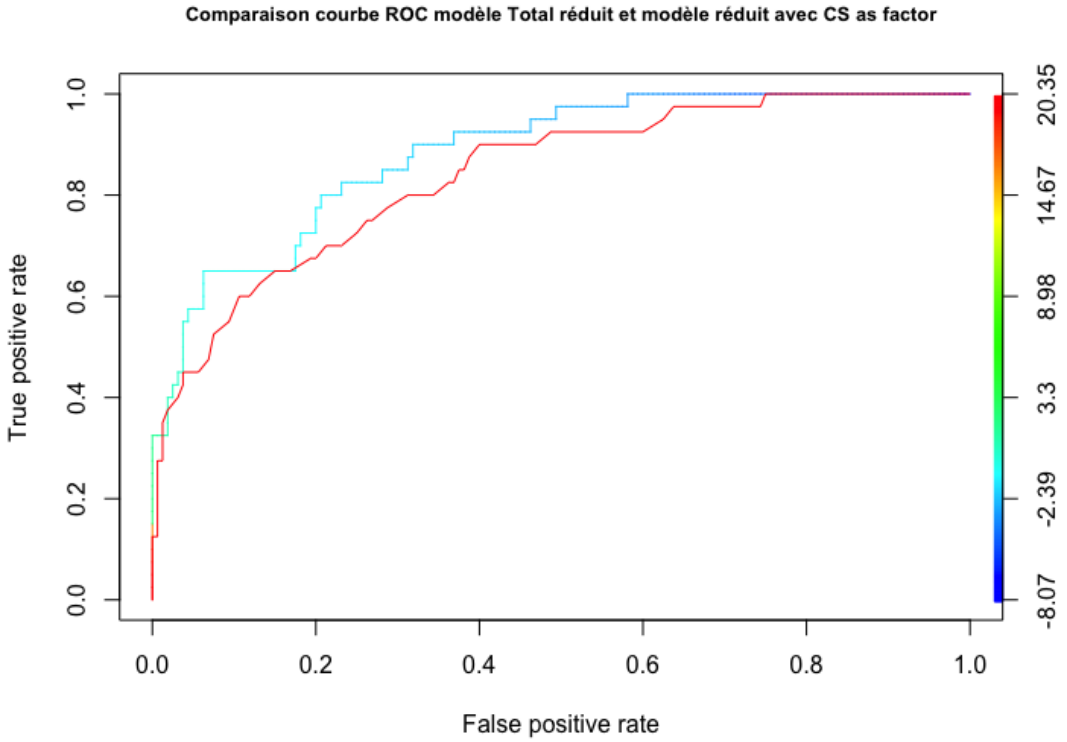
Modèle	$M_{S,comp}$	$M_{S,red}$	$M_{S,qual}$	$M_{M,comp}$	$M_{M,red}$	$M_{T,red}$
Deviance	146.69	151.02	141.87	155.67	160.27	132.33
AIC	166.69	159.02	151.87	169.27	168.27	148.33
Nombre de variables	9	3	3	6	3	7
Nombre d'itérations	6	6	15	5	5	6
Taux d'erreurs	15.5%	15%	14.5%	16.5%	15.5%	14.5%
Spécificité	30%	30%	32.5%	25%	27.5%	42.5%
Sensibilité	98.1%	98.75%	98.75%	98.1%	98.75%	96.25%

Tout d'abord, remarquons que les faibles spécificités obtenues (moins de 42.5%) ne sont pas surprenantes car nous avons pu voir sur les graphes des résidus que la valeur des résidus augmente lorsque le statut vital du patient devient 1. C'est le modèle $M_{T,red}$ qui a la meilleure spécificité avec une spécificité de 42.5%.
Les taux d'erreurs varient tous entre 14.5% et 16.5%. C'est le modèle $M_{T,red}$ qui a le plus faible taux d'erreurs.
Quant à la sensibilité, tous les modèles ont une spécificité supérieure à 96%. Après analyse de ce tableau, on remarque donc que le modèle $M_{T,red}$ semble le plus performant. D'ailleurs il donne une deviance résiduelle bien plus faible que les autres et il minimise l'AIC. Le modèle $M_{S,qual}$ donne des résultats un peu moins bon mais est tout de même performant, surtout qu'il met en jeu seulement 3 variables.

Pour confirmer nos conclusions quant aux performances des modèles, on trace leurs courbes ROC :



Le modèle dont la courbe ROC s'éloigne le plus de la bissectrice est considéré comme le meilleur. En supperposant les courbes, on s'aperçoit que 3 modèles sont très proches : $M_{S,comp}$, $M_{S,qual}$ et $M_{M,comp}$. Les courbes des modèles $M_{S,red}$ et $M_{M,red}$ sont situées légérement en dessous et indiquent donc que ces modèles sont de moins bonne qualité. On trace également en bleu la courbe ROC du modèle $M_{T,red}$ que l'on compare à celle du modèle $M_{S,qual}$ en rouge.



Cette fois-ci, on observe clairement que la courbe ROC du modèle $M_{T,red}$ est de bien meilleure qualité. Finalement ce modèle est bien le plus performant pour expliquer le statut vital.