TD2 régression logistique (correction)

Marie Fourcot

03/2022

```
library("ggplot2")
library(ROCR)
library("plotROC")
```

Chargement des données et reprise modèles du TD précédent:

Pour rappel : Dans le cadre d'une étude sur les facteurs prénataux liés à un accouchement prématuré chez les femmes déjà en travail prématuré, on dispose de 13 variables explicatives sur 388 femmes incluses dans l'étude.

La variable à expliquer (PREMATURE) est l'accouchement prématuré.
L'objectif est de définir les facteurs prédictifs d'un accouchement prématuré (Y). Pour chaque modèle considéré, on notera π la probabilité d'un accouchement prématuré sachant les variables $X_1, \cdots, X_p$ incluses.
Les données contiennent les variables suivantes :

Var	Description	Commentaire
GEST	l'âge gestationnel à l'entrée dans l'étude	en semaine
DILATE	la dilatation du col utérin	en cm
EFFACE	l'effacement du col	en %
CONSIS	la consistance du col	1 : mou
		2 : moyen

Var	Description	Commentaire
		3 : ferme
CONTR	la présence de contractions	1 : oui
		2 : non
MEMBRAN	état des membranes	1 : rupturées
		2 : non rupturées
		3 : incertain
AGE	l'âge de la mère	en années
STRAT	période de la grossesse	1-4
GRAVID	la gestité	nombre de grossesses antérieures, y compris celle en cours
PARIT	la parité	nombre de grossesses à terme antérieures
DIAB	diabète	1 : présence
		2 : absence
TRANSF	le transfert vers un hôpital en soins spécialisés	1 : oui
		2 : non
GEMEL	type de grossesse	1 : simple
		2 : multiple

Pour remplir cet objectif, nous avons construit pour le moment deux modèles :

- un premier modèle avec comme variable explicative une variable binaire, la variable GEMEL
- un deuxième, avec comme variable explicative une variable quantitative, la variable EFFACE.

Nous allons maintenant construire (et évaluer) des modèles plus complexes avec une approche plus proche de ce que nous ferions réellement pour construire un modèle de régression logistique.

1 Etude de plusieurs variables explicatives

13. Ajuster le modèle expliquant l'accouchement prématuré par le type de grossesse et l'effacement du col.
14. Comparer les deux modèles model2 et model3 en utilisant le test du rapport de vraisemblance.
    Utiliser la fonction `anova(, test= 'LRT')`, se référer à `anova.glm` pour l'aide.
    Quel modèle gardez-vous ?
15. Estimer le modèle complet (fullmodel) :
16. Combien de patientes sont inclues dans ce modèle?
17. Évaluer la significativité de chaque coefficient de fullmodel.
    Utilisez la fonction step pour la sélection automatique de variables dans le modèle et interpréter. On appelle reduced le modèle réduit aux variables sélectionnées. Comparer les deux modèles (complet et réduit).
18. Comparez les différents modèles (model2, model3, fullmodel, reduced) à l'aide de leur AIC.
19. Interpréter les coefficients de reduced. Quels sont les facteurs de risque pour l'accouchement prématuré ? Quels sont les facteurs protecteurs ?

2 Evaluation de la règle de décision

20. Calculer les valeurs prédites des probabilités d'intérêt en utilisant la fonction `predict` ou le champ fitted.values de reduced. On nommera ce nouveau score S.
    Visualiser et commenter la qualité de prédiction (tracer par exemple des boites à moustache).
21. Calculer la matrice de confusion pour un seuil de décision à 0,5.

22. On décide arbitrairement d'affecter toutes les valeurs qui ont un score S supérieur au score de la dernière ligne au groupe 1 et les autres au groupe 0. Calculer alors sensibilité et spécificité pour ce seuil.

23. Tracer la courbe ROC associée au score S en utilisant les fonctions `prediction` et `performance` du package ROCR

24. Explorer les objets qui permettent de calculer la courbe ROC :

```
perf@x.values[[1]]
perf@y.values[[1]]
perf@alpha.values[[1]]
```

25. Calculer l'aire sous la courbe ROC en utilisant les commandes suivantes :

26. Calculer le seuil le plus proche du point idéal pour la courbe ROC liée au score S. Calculer la nouvelle matrice de confusion associée à ce seuil de décision.