TD3 régression logistique

Code

Marie Fourcot

15/03/2022

Hide

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

Chargement des données :

Hide

```
load("prema.RData")
```

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 %

Var	Description	Commentaire
CONSIS	la consistance du col	1 : mou
		2 : moyen
		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 d'abord construit 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.

Puis nous avons construit un modèle complet que nous avons affiné et évalué.

Nous allons maintenant construire des modèles affinés et permettre une meilleure évaluation de ceux-ci.

27. Imputer les données manquantes de la variable `DIAB` par la réponse majoritaire.
28. Séparer le jeu de données en jeu d'apprentissage et jeu de test. Avec 70% des données pour l'apprentissage et 30% pour le test. Pour cela, utiliser les fonctions `slice_sample` et `anti_join` du package `dplyr`.
29. Estimer le modèle complet des données d'apprentissage.
    Évaluer la significativité de chaque coefficient de ce modèle.
30. Réaliser la sélection automatique de variables dans le modèle.
31. Calculer et interpréter les odds ratio significatifs.
32. Changer la valeur de référence pour les variables DIAB et MEMBRAN à l'aide de la fonction `relevel`.
33. Prédire l'accouchement prématuré sur le jeu de test.
34. Tracer la courbe ROC.
35. Calculer l'aire sous la courbe ROC.