Classification supervisée
Régression logistique

Charlotte Baey (charlotte.baey@univ-lille.fr)

Un compte rendu est à rendre pour ce TP.

Quelques consignes :

- le travail est à effectuer en binôme
- seuls les fichiers aux formats .pdf ou .html seront acceptés
- **il n'est pas demandé d'inclure le code** : c'est la présentation et l'interprétation des résultats qui sont importantes
- soyez concis (maximum 5 pages) : choisissez judicieusement les graphes à afficher, n'incluez pas de sorties brutes du logiciel
- à chaque figure et chaque tableau doit être associée une légende
- rédiger ce rapport comme s'il était adressé à quelqu'un n'ayant pas de connaissances précises en statistiques

Quelques remarques générales :

- **bien lire l'énoncé !** bien souvent, la réponse à votre question s'y trouve ...
- l'objectif d'une séance de TP est que chacun s'approprie les outils proposés : lorsque l'énoncé suggère l'utilisation d'une fonction, **commencez par chercher dans l'aide de R** pour savoir comment l'utiliser pour répondre à la question (choix des arguments, ...)
- en général, lorsqu'un message d'erreur s'affiche, il suffit de le lire pour comprendre comment résoudre le problème ... (ex. : "la variable XXX n'est pas définie" signifie que R ne connaît pas cette variable : vérifiez l'orthographe, la table d'où elle provient, ...)
- il est préférable de trouver soi-même la réponse à une question que de recopier la solution de l'enseignant, même si votre solution vous paraît moins élégante. Il sera toujours possible de simplifier votre code lorsque votre pratique s'améliorera.

1 Prédiction du diabète

On reprend la base de données sur le diabète du TP précédent. L'objectif est, ici aussi, de décrire et prédire la présence de diabète chez un patient en fonction de certaines charactéristiques cliniques. Nous verrons que l'apport de la régression logistique tient dans l'interprétation des résultats du modèle.

***Objectif** : identifier les facteurs de risques associés à la présence du diabète.*

1. Importer la base de données et en faire une analyse descriptive.
2. Ajuster un modèle de régression logistique, à l'aide de la fonction `glm()` et de l'option `family="binomial"`.
3. Afficher les résultats du modèle, à l'aide de la fonction `summary()`.
4. Calculer les odds-ratio et leurs intervalles de confiance. Plusieurs solutions pour cela :
   — à l'aide des fonctions `coef`, `confint` et `exp`
   — à l'aide de la librairie `broom` et de la fonction `tidy`
   — à l'aide de la librairie `gtsummary` et de la fonction `tbl_regression`
   — à l'aide de la librairie `forestmodel` et de la fonction `forest_model`

   ***N.B.** : on peut également visualiser graphiquement ces informations soit à l'aide de `ggplot` en utilisant les sorties de la fonction `tidy` du package `broom`, ou utiliser la librairie `GGally` et la fonction `ggcoef` appliquée directement à la sortie de la fonction `glm`.*
5. Proposer une interprétation du modèle. Quelles sont les variables qui agissent comme des facteurs de risque ? y a-t-il des variables non significatives ? que peut-on faire lorsqu'un modèle de régression multiple présentent une ou plusieurs variables non significatives ? Proposer une méthode pour construire un modèle réduit contenant moins de variables, en faisant un compromis entre ajustement et complexité.
6. Visualiser l'effet de chaque variable sur la variable à prédire. Pour cela on pourra utiliser la librairie `ggeffects` et la fonction `ggeffect`.

   ***N.B.** : cette fonction s'utilise aussi sur des objets obtenus à l'aide de la fonction `lm()` par exemple*
7. Tracer la courbe ROC associée au modèle.

2 Prédiction du mode de contraception

Dans l'exercice précédent, on s'est intéressé à la prédiction d'une variable binaire. Ici, on va s'intéresser à la prédiction d'une variable qualitative à trois modalités : on parle de régression logistique **multinomiale** ou **polytomique**. L'interprétation des résultats est un peu différente dans le cas d'une régression logistique multinomiale. En effet, au lieu d'obtenir un odds-ratio pour chaque variable, on obtient un odds-ratio pour chaque variable et chaque comparaison entre les différentes modalités de la variable à prédire.

La base de données `cmc.data` qui nous intéresse provient d'une étude sur la contraception effectuée chez 1473 femmes Indonésiennes. Les variables de la table sont les suivantes :

— `age` : l'âge en années
— `education` : le niveau d'éducation (codé de 1 : faible à 4 : élevé)
— `husband_education` : le niveau d'éducation du mari (codé de 1 : faible à 4 : élevé)
— `nbchildren` : le nombre d'enfants
— `religion` : la religion de la femme (1 : mulsumane, 0 : autre)
— `working` : est-ce que la femme travaille (1 : oui, 0 : non)
— `husband_occupation` : le niveau d'occupation du mari (codé de 1 : faible à 4 : élevée)
— `standard_of_living` : le niveau de vie du ménage (codé de 1 : faible à 4 : élevé)
— `media` : l'exposition aux médias (1 : oui, 0 : non)
— `contraceptive` : le type de contraception (1 : aucune, 2 : court-terme, 3 : long-terme)

***Objectif** : identifier les facteurs influençant la prise d'une contraception, et étudier les différences éventuelles entre le cas d'une contraception court-terme ou long-terme.*

1. Importer le jeu de données et faire une analyse descriptive de la base.
2. Recoder les variables qualitatives en facteurs, à l'aide de la fonction `as.factor`, et pour plus de lisibilité, renommer les niveaux de la variable cible `contraceptive` en toutes lettres. Refaire l'analyse descriptive.
3. Ajuster un modèle de régression logistique multinomial, à l'aide de la fonction `multinom` de la librairie `nnet`.
4. Afficher les odds-ratio comme dans l'exercice précédent. Proposer une interprétation des résultats du modèle. On pourra s'aider des outils graphiques.
5. Tracer les courbes ROC associées au modèle.