Nous travaillons sous R. Tous les tests sont à 5%.

1 Supports

Le site de notre cours est http://eric.univ-lyon2.fr/~ricco/cours/cours_regression_logistique.html

Plus spécifiquement pour cette séance, nous nous référerons à :

TUTO 1 – http://eric.univ-lyon2.fr/~ricco/cours/slides/roc_curve.pdf

TUTO 2 – http://tutoriels-data-mining.blogspot.com/2015/02/validation-croisee-bootstrap-diapos.html

2 Données

On cherche à déterminer **le faible poids des bébés** à partir des caractéristiques et du comportement de la mère. Les données proviennent du célèbre ouvrage de Hosmer & Lemeshow (2ème édition, 2000) dont la description est disponible dans la Section 1.6.2, Tableau 1.6, page 26.

Voici la liste des variables (LOW est la variable cible ; LOW = y, la modalité cible) :

- `LOW : Low Birth Weight (y : < 2500 g ; n : >= 2500 g)`
- `AGE : Age of Mother (years)`
- `LWT : Weight of Mother at Last Menstrual Period (pounds)`
- `SMOKE : Smoking Status During Pregnancy (1 : yes, 0 : no)`
- `HT : Hystory of Hypertension (1 : yes, 0 : no)`
- `UI : Presence of Uterine Irritability (1 : yes, 0 : no)`
- `FTV : Number of Physicians Vistis During the First Trimester`
- `PTL : History of Premature Labor (1 : yes, 0 : no)`

3 Exercices – Bébés à faible poids – Courbe ROC et AUC

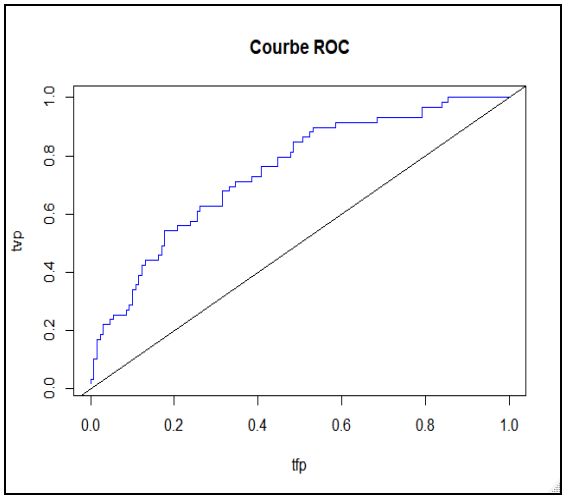
3.1 Importation, modélisation, probabilités d'affectation aux classes

0. Importez le fichier « **Low_Birth_Weight_Data.xlsx** » (read.xlsx si package 'xlsx').
1. Affichez les premières lignes des données (head), puis leur description (str) (189 obsv., 8 variables).
2. Réalisez la régression de LOW versus l'ensemble des autres variables (`LOW ~ .`) (glm).
3. Affichez les caractéristiques détaillées du modèle (summary). Quelles sont les variables qui semblent pertinentes, au sens du test de significativité de Wald, dans la régression ? (3 variables). Nous décidons de conserver toutes les explicatives pour la suite.

4. Calculez les probabilités d'affectation à la modalité cible (LOW = y), nous les appellerons « **scores** » dans la suite de notre exercice (predict).
5. Affichez la description de ce nouveau vecteur des scores. Quelles sont les valeurs minimales ? Maximales ? Médianes (0.03686, 0.88381, 0.26) (summary).

3.2 Courbe ROC

6. Nous souhaitons construire la courbe ROC en resubstitution c.-à-d. en utilisant les données d'apprentissage qui ont servi à la construction du modèle. Recodez la variable cible (LOW) en variable binaire $\mathbf{Y} \in \{0, 1\}$ (1 : LOW = y, 0 : LOW : n) (différentes pistes, ifelse peut-être). Vérifiez la bonne tenue de l'opération en croisant LOW avec cette variable recodée (table).
7. Comptabilisez le nombre d'observations positives (Y = 1) (59) et négatives (Y = 0) (130).
8. Nous essayons de reproduire les calculs décrits dans notre support (**TUTO 1**, page 10) :
    a. Créez un data frame avec la variable Y et les scores. Affichez les premières lignes (head).
    b. Triez ce data frame selon les scores décroissants (order peut-être). Affichez les premières (head) et les dernières lignes (tail) du nouveau data frame. Que remarquez-vous pour la colonne scores ? (individu 4, Y = 1, score le plus élevé, 0.88381 ; individu 82, Y = 0, score le plus faible, 0.03686).
    c. Calculez la colonne des taux de faux positifs (TFP) en vous basant sur la description du support de cours (cumsum peut servir peut-être).
    d. Calculez la colonne des taux de vrais positifs (TVP).
    e. Affichez la courbe ROC sur la base des (TFP, TVP) (plot). Rajoutez la diagonale principale (abline).



9. Nous essayons de réitérer l'opération en utilisant le package spécialisé « ROCR ».

a. Installez et chargez la librairie « ROCR » (library).

b. Faites appel à la fonction prediction() auquel vous passez les scores et les étiquettes (LOW) (cf. par ex. https://rocr.bioinf.mpi-sb.mpg.de/). Affichez le contenu de l'objet généré, qu'observez-vous ? Faites le parallèle avec notre support (**TUTO 1**, page 9).

c. Passez cet objet à la fonction performance() en demandant à obtenir le taux de vrais positifs (« tpr » en anglais) et le taux de faux positifs (« fpr »). Que contient l'objet résultat ? (faire le parallèle avec **TUTO 1**, page 10).

d. Affichez la courbe ROC (plot) en lui passant ce dernier objet. Ajoutez toujours la diagonale principale (abline). Est-ce que cette dernière courbe correspond à celle que nous avions élaboré plus haut ? (oui, gros problème sinon).

3.3 Calcul de l'AUC

10. A partir du data frame des valeurs triées de Y et des scores, nous souhaitons calculer l'aire sous la courbe (AUC) en nous appuyant sur les étapes décrites dans notre support (**TUTO 1**, page 13).

    a. Construisez la colonne des rangs des valeurs (créer une série décroissante suffit puisque les données sont déjà triées, avec seq par ex.).

    b. Calculez la somme des rangs des observations positives c.à-d. les individus étiquetés (Y = 1) (7492).

    c. Calculez la statistique de Mann-Whitney (5722).

    d. Déduisez alors la valeur de l'AUC (0.7460235).

11. Faisons de nouveau appel à la libraire « ROCR » pour conforter nos résultats. Voyez comment vous pouvez exploiter la fonction performance() en lui passant les bons paramètres. La valeur de l'AUC obtenue est-elle cohérente avec la nôtre ? (oui, ouf !).

3.4 Courbe ROC et valeur de l'AUC en Leave-One-Out (LVO)

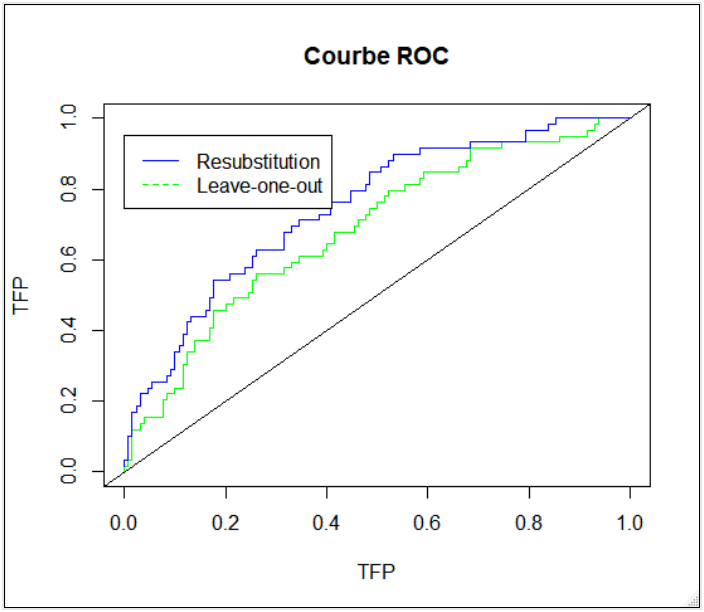
Nous savons que l'évaluation en resubstitution – c.-à-d. utiliser les mêmes données pour élaborer le modèle prédictif et en mesurer les performances – conduit à des indicateurs biaisés, souvent trop optimistes, laissant à penser à tort que la modélisation a bien fonctionné (**TUTO 2**, page 5).

L'AUC obtenue dans la section précédente par exemple est certainement surévaluée. Nous nous tournons vers les techniques de rééchantillonnage pour disposer d'une valeur plus réaliste, nous étudions en particulier l'approche « leave-one-out » (**TUTO 2**, page 12).

12. Préparez un vecteur de scores de la même taille que le vecteur Y, remplie de valeur 0.

13. En vous appuyant sur le principe de la LVO, calculez les scores des individus en les considérant comme individu supplémentaire c.-à-d. n'ayant pas participé à la construction du modèle (schématiquement, vous construisez [glm] le modèle sur les (n-1) observations, et vous effectuez la prédiction [predict] avec l'individu qui a été mis de côté ; vous itérez sur l'ensemble des observations ; il y a certainement une boucle dans l'affaire).

14. Avec ce nouveau vecteur des scores obtenus en LVO et les étiquettes observées (Y), construisez la courbe ROC, rajoutez la diagonale et l'ancienne courbe en resubstitution (allez à l'essentiel, utilisez ROCR directement cette fois-ci) (lines pour ajouter une courbe supplémentaire dans un graphique, legend pour ajouter une légende). Que constatez-vous ?



15. Calculez l'AUC en LVO (0.6827901). Conclusion ?