Durant cette séance, pour chaque méthode à évaluer, recensez le nombre et la liste des variables sélectionnées, ainsi que les performances en test.

1 Supports

Nous utiliserons les supports suivants durant cette séance :

TUTO 1 – http://tutoriels-data-mining.blogspot.com/2012/03/introduction-r-regression-logistique.html

TUTO 2 - http://tutoriels-data-mining.blogspot.com/2008/10/rgression-logistique-comparaison-de.html

TUTO 3 - http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf

2 Données

On cherche à prédire et expliquer l'occurrence du diabète chez des femmes enceintes. Dans la base « **diabete_reg_logistique.xlsx** », « diabète » est la variable cible, « diabète = positive » est la modalité cible, les autres variables constituent les explicatives potentielles.

Le fichier a été scindé en 2 échantillons : « apprentissage » pour la construction du modèle (1ère feuille du classeur Excel) ; « test » pour son évaluation (2nde feuille).

3 Exercices

3.1 Importation et inspection des données

1. Installez et chargez la librairie « xlsx » (ou, à défaut, utilisez d'autres librairies telles que « openxlsx » ou « readxl »).
2. Importez les données "apprentissage" et "test " dans deux data.frame distincts (read.xlsx ; TUTO 1, page 7).
3. Affichez les caractéristiques des deux jeux de données (str). De combien d'observations et de variables disposons-nous en apprentissage et en test ? (apprentissage, 568 obs. x 11 var. ; test, 200 x 11).
4. Affichez résumé des données (summary). Quelle est la distribution des classes en apprentissage (372 négatifs, 196 positifs) et en test (128, 72).

3.2 Modélisation et évaluation du modèle (1)

5. A l'aide de glm(), sur les données d'apprentissage, construisez le modèle permettant d'expliquer « diabète » à partir des autres variables disponibles dans le data frame (TUTO 2, section 3.2.2, page 9). Affichez le modèle (print).
6. Quelle est la valeur de la déviance du modèle ? (540.9). Quelle est la déviance du modèle trivial réduit à la constante ? (TUTO 3, section 1.6.1) (732).

7. Affichez les propriétés de l'objet issu de la modélisation (attributes) (TUTO 1, page 9).
8. A partir des informations fournies par l'objet, calculez les différents pseudo-R^2 (McFadden, Cox and Snell et Nagelkerke) (resp. 0.261, 0.2856, 0.394) (TUTO 3, section 1.6, en particulier le tableau 1.1 pour les formules).
9. Testez la significativité globale du modèle à l'aide du rapport de vraisemblance (TUTO 1, page 9). Le modèle est-il globalement significatif à 1% ? (oui, largement, KHI-2 = 191.05, ddl = 10, p-value = 1.17E-35)

3.3 Evaluation du modèle (2)

10. Nous souhaitons évaluer les performances en généralisation du modèle. Appliquez le modèle sur l'échantillon test (predict) (TUTO 1, page 10 ; ou TUTO 2, section 3.2.3).
11. Convertissez les probabilités d'affectation en prédiction en les comparant à la valeur seuil 0.5 (ifelse). Sur l'échantillon test, combien d'observations sont prédites « positive » ? (table) (61).
12. Calculez la matrice de confusion en confrontant les classes observées et prédites sur l'échantillon test. Déduisez de cette la matrice le taux d'erreur (0.185).

3.4 Evaluation des variables

13. Nous cherchons à identifier les variables pertinentes via le test de Wald pour la significativité individuelle des coefficients. Affichez les caractéristiques détaillées du modèle (summary) (TUTO 1, page 8).
14. Quelles sont les variables pertinentes à 5 % ? (pregnant, plasma, diastolic, bodymass, pedigree, age).

3.5 Sélection de variables « backward » - Critère AIC (Akaike)

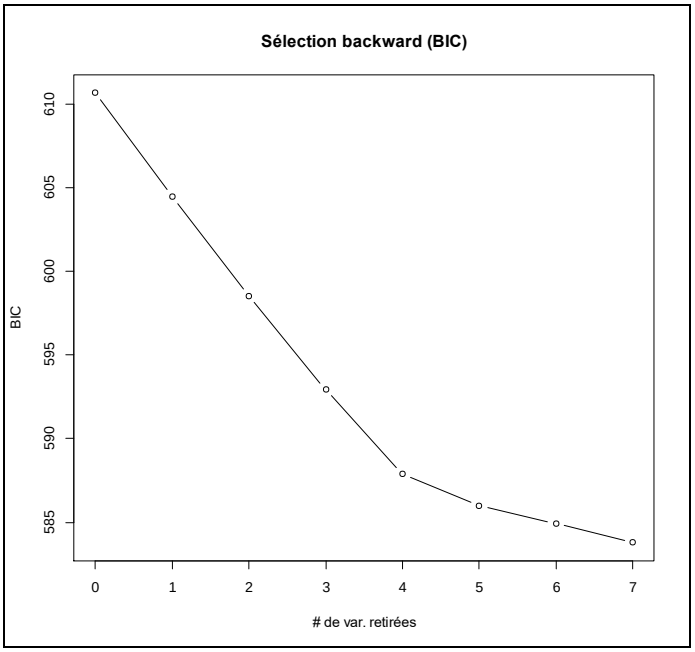
15. Nous souhaitons réaliser une sélection de variables « backward » via le critère Akaike (AIC). Chargez la librairie (MASS). Lancez le processus de sélection (stepAIC ; TUTO 1, page 10). Quelles ont été les variables retenues finalement ? (pregnant, plasma, diastolic, bodymass, pedigree, age).
16. Quelles sont ses performances sur l'échantillon test ? (taux d'erreur = 0.205).

3.6 Sélection de variables « backward » - Critère BIC (Schwartz)

17. Nous souhaitons réitérer la même opération, mais avec le critère BIC. Quel critère devons-nous modifier lors de l'appel de stepAIC (TUTO 2, page 10, dernier paragraphe ; spécifiez la valeur du paramètre k = log(nombre d'observations de l'échantillon d'apprentissage) -- https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/stepAIC.html) ? Combien

et quelles variables sont retenues cette fois-ci ? (3 variables ; pregnant, plasma, bodymass).

18. Quelles sont les performances en test de ce nouveau modèle ? (taux d'erreur = 0.205).
19. Affichez les propriétés de l'objet « modèle » issu de la sélection de variables (attributes).
20. Nous nous intéressons en particulier au champ « $anova ». Affichez son contenu (print). Combien de variables ont été retirées du modèle initial ? (7) Dans quel ordre ont-elles été retirées (triceps, alea1, serum, …, diastolic).
21. Quelle est la valeur initiale du critère BIC ? (même si le nom de la colonne est AIC, il s'agit bien du critère BIC dans le cas présent ; 610.6804) Sa valeur finale ? (583.7877)
22. Tracez un graphique avec en abscisse le nombre de variables retirées et en ordonnée l'évolution du critère BIC.



3.7 Sélection « forward » (BIC)

23. Réalisez une sélection « forward » cette fois-ci, toujours avec le critère BIC (TUTO 1, page 12). Quelles sont les variables retenues dans le modèle (les mêmes qu'en backward BIC).
24. Dans quel ordre ont été introduites les variables ? (plasma, bodymass, pregnant).