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

TUTO 2 – http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf

TUTO 3 – http://tutoriels-data-mining.blogspot.com/2014/10/la-discretisation-des-variables.html

2 Données

On cherche à expliquer l'occurrence du diabète chez des femmes enceintes. Dans la base « **pima-non-linearite.xlsx** », « diabète » est la variable cible, « diabète = positive » est la modalité cible, les autres variables constituent les explicatives (plasma, diastolic, triceps, serum).

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 – Identification et traitement des non-linéarités

3.1 Importation, modélisation, évaluation

0. Importez la première feuille du classeur « **pima-non-linearite.xlsx** » (read.xlsx si package 'xlsx') (**TUTO 1**, page 7). Affichez les caractéristiques de la base (str) (292 obs., 5 variables).
1. Importez la seconde feuille du classeur (1160 obs., 5 variables).
2. Construisez le modèle sur les données d'apprentissage (**TUTO 1**, page 8) (glm). Aux fins de vérification, quelle est la valeur de l'AIC ? (AIC : 292.84). Quelles sont les variables pertinentes à 5% ? (plasma, triceps).
3. Appliquez le modèle sur le second échantillon pour obtenir les prédictions en test (predict) (**TUTO 1**, page 10).
4. A l'aide du package « caret » qu'il faut au préalable installer et charger, calculez la matrice de confusion et les indicateurs de performances en test (confusionMatrix). N'oublions pas que « diabète = positive » est la modalité cible. Quelles sont les valeurs respectives du taux de reconnaissance (taux de succès), rappel et précision ? (0.756, 0.4397, 0.7447). Que remarquez-vous dans le mode de présentation de la matrice de confusion de « caret » ?

3.2 Test de Box-Tidwell – Identification des non-linéarités

Certaines variables paraissent non-pertinentes, mais il est possible qu'elles agissent non-linéairement sur le LOGIT (**TUTO 2**, section 8.2). Nous souhaitons implémenter le test « omnibus » de Box-Tidwell destiné à détecter l'existence d'éventuelles non-linéarités.

5. Programmez la boucle de calcul décrite dans le support (**TUTO 2**, section 8.2.3). Combien de régression devons-nous effectuer ? (4, parce que 4 variables explicatives candidates pour lesquelles nous devons former Z = X * log(X) à intégrer dans la régression)

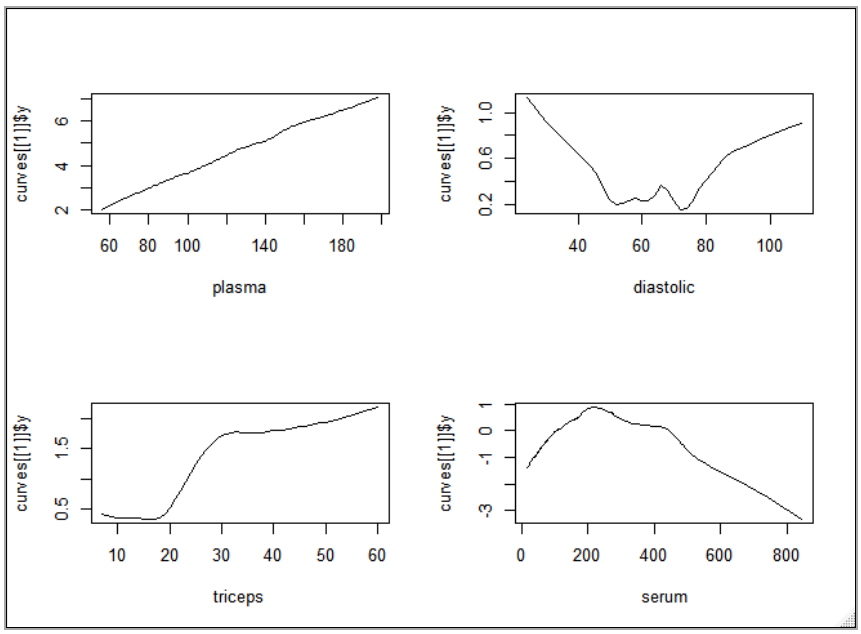
6. Quelles sont les variables suspectées d'influence non-linéaire dans la régression à 5% ? (serum).

3.3 Forme de non-linéarité – Graphiques des résidus partiels

Le test précédent indique l'éventuelle présence de non-linéarité mais ne donne aucune indication sur sa forme. Elle ne donne pas non plus par conséquent de suggestions sur les transformations de variables à opérer pour remédier au problème. Pour pallier ces insuffisances, nous nous tournons vers le graphique des résidus partiels (**TUTO 2**, section 8.2.4).

7. Installez et chargez le package « rms ». Lancez la même régression (lrm) que précédemment pour obtenir un objet que nous pourrons mettre à contribution pour le calcul des résidus partiels par la suite. Le modèle est-il globalement significatif à 5% au sens du test de rapport de vraisemblance ? (cf. les sorties de l'objet 'lrm' avec print) (oui, p-value < 0.0001).

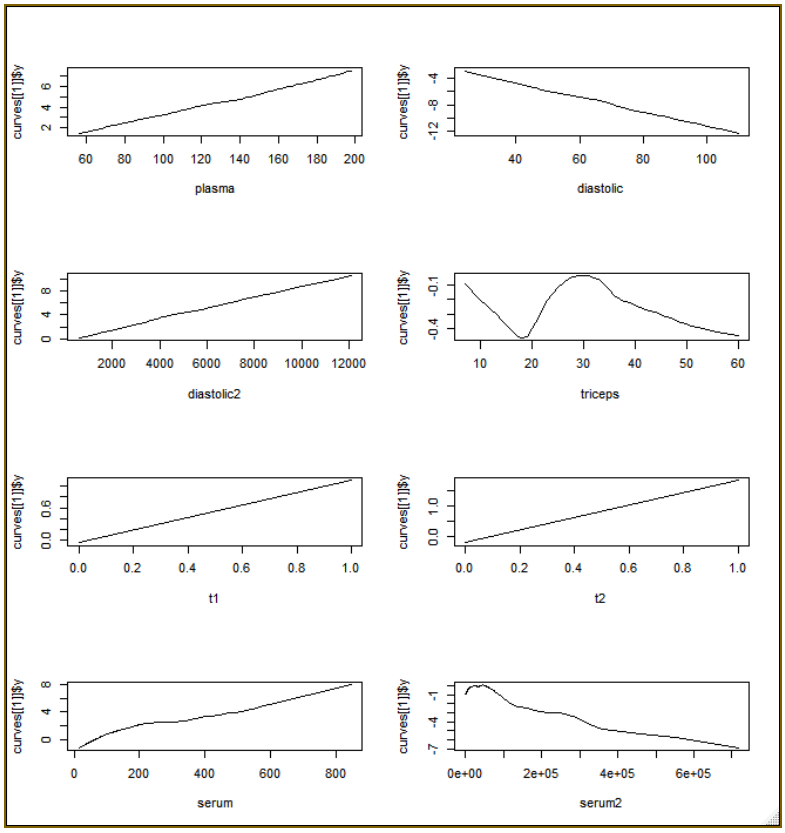
8. Affichez les graphiques des résidus partiels (plot.lrm.partial, attention il fallait ajouter des options dans l'appel de lrm() pour que l'outil puisse produire les graphiques). Arrangez-vous pour les graphiques tiennent dans un même seul panneau (avec par(mfrow…) peut-être).



9. Pour quelles variables la non-linéarité de l'influence sur le LOGIT semble avérée ? (diastolic, triceps, serum). Quelles transformations de variables peut-on suggérer ?

3.4 Régression sur variables transformées

10. Accordons-nous sur les transformations suivantes :
    a. Nous ne modifions pas la variable « plasma ».
    b. « diastolic » est passée au carré (diastolic2).
    c. « triceps » est discrétisée en 3 intervalles (]-oo, 20[; [20, 30[; [30, +oo[). Il faut donc créer 2 indicatrices (T1, T2) en prenant comme modalité de référence le premier intervalle. Nous adoptons un codage non-emboîté (simple codage 0/1 pour les deux derniers intervalles) (cut ou ifelse sont les outils possibles, au choix).
    d. « serum » est aussi passée au carré (serum2).
11. Réalisez la régression (lrm) en incluant ces nouvelles variables. Quelles sont les variables pertinentes à 5% ? (plasma, T1, T2, serum, serum2).
12. Affichez de nouveau les graphiques des résidus partiels. Que constatez-vous ?



13. Réaliser la même régression mais au format glm (glm).

14. Puis effectuez la prédiction sur l'échantillon test (predict), après avoir recodé de manière appropriée les variables concernées bien évidemment.
15. Quel est le taux de reconnaissance cette fois-ci ? (accuracy = 74.05 %). Les transformations opérées ont-elles été probantes ? (non, pas vraiment).

3.5 Traitement des non-linéarités par la discrétisation

La discrétisation des variables explicatives est souvent mise en avant pour appréhender les non-linéarités (**TUTO 2**, section 8.2.2). Nous l'avons réalisé sur une des variables (triceps) dans la section précédente. Mais déterminer manuellement le nombre d'intervalles et les bornes de découpage se révèle très vite fastidieux lorsque nous avons à traiter une grande base de données composée d'un nombre élevé de variables. Nous nous tournons vers les méthodes automatisées dans cette section (TUTO 3).

16. Installez et chargez la librairie « discretization ».
17. Nous optons pour une discrétisation supervisée (mdlp) (**TUTO 3**, page 22). Pourquoi ?
18. Lancez la discrétisation pour l'ensemble des variables explicatives candidates ($cutp). Pour chaque variable : Quel est le nombre d'intervalle ? Quelles sont les bornes de discrétisation ?
19. Affichez la nouvelle version des données d'apprentissage ($Disc.data).
20. (Normalement) A l'issue de l'opération, « plasma », « triceps », « serum » sont binaires, « diastolic » est constituée d'une constante. Créez une nouvelle version du data frame d'apprentissage : en transformant les binaires en 0/1, en retirant la constante, et en lui adjoignant la variable cible « diabète ».
21. Relancez de nouveau la régression avec exclusivement les variables discrétisées (glm). Quelle est la valeur de l'AIC ? (275.04) Quelles sont les variables pertinentes à 5% ? (toutes)
22. Recodez les variables de l'échantillon test en utilisant les paramètres calculés sur l'échantillon d'apprentissage c.-à-d. nombre d'intervalles et bornes de découpage (cut ou ifelse)
23. Appliquez le modèle sur cette nouvelle version de l'échantillon test (predict) et calculez les performances en test (confusionMatrix) (accuracy = 73.62%). La discrétisation des variables explicatives a-t-elle été probante ? (non plus, comme quoi le mieux peut-être – parfois – l'ennemi du bien, il faut être prudent quand on commence à triturer les données…).