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/regression_logistique.pdf

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

TUTO 3 - http://tutoriels-data-mining.blogspot.com/2012/03/introduction-r-regression-logistique.html

2. Données

On dispose de **données sur l'infidélité conjugale** **(infidelites.xlsx)**. Voici la description de la base :

Table F22.2 from William Greene Fair ' s (1977) - Extramarital Affairs Data

Y: Number of affairs in past year (0-3 = real number, 7 = 4 to 10, 12 = more)
Sex : Gender (0 = female, 1 = male)
Age : Age in years
YearsMarried : Number of years married
Chlidren : 0 = no, y = yes
Religious : Religiousness (1 = anti to 5 = very)
Education : Education in Years
Occupation : "Hollingshead Scale" 1-7
RatingMarriage : Self rating of marriage (1 = very unhappy to 5 = very happy)

3. Exercices - Analyse de l'infidélité

0. Chargez les données dans un data frame. Combien y a-t-il d'observations et de variables dans la base ? (**TUTO 3**, page 7) (601 obs., 9 variables).

1. Recodez la variable Y en une variable binaire Z où 0 correspond aux personnes fidèles, 1 aux personnes ayant fauté au moins une fois. Combien y a-t-il de personnes fidèles dans la base (idem pour les infidèles) (table) (451 vs. 150)

2. Le genre (homme/femme) a-t-il un impact sur l'infidélité ? (glm + summary) (au sens du test de Wald).

3. Réalisez la régression logistique expliquant Z à l'aide des variables (Sex, Age, ..., Rating Marriage) (glm)

    a. Quelles sont les variables pertinentes (donnant lieu à des coefficients significativement non-nuls) au sens du test de Wald ?

   b. Parmi ces variables, comment interprétez-vous leur impact sur l'infidélité : quels sont les facteurs qui poussent les personnes à être fidèle ou l'inverse ?

4. Est-il possible de retirer le bloc de variables non significatives ? Utilisez deux procédés différents, vérifiez la cohérence des résultats (par ex. selon le test du rapport de vraisemblance d'une part, via la comparaison de la valeur de l'AIC d'autre part).

5. Nous partons maintenant de la régression intégrant (Z ~ Age + YearsMarried + Religious + RatingMarriage). N'y a -t-il pas quelque chose d'intriguant dans l'impact de Age et YearsMarried ? Calculez la corrélation entre ces deux variables. Que constatez-vous ?

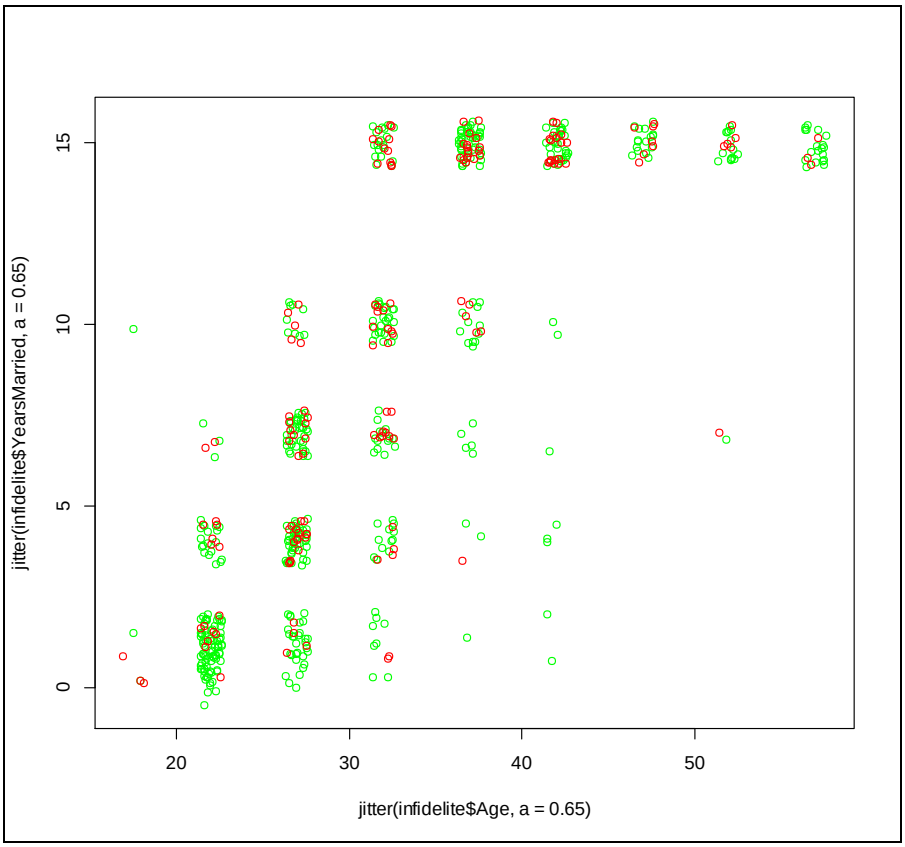
6. Concentrons-nous sur les variables Age et YearsMarried

   a. Réalisez la régression avec la seule variable Age, que constatez-vous ?

   b. Réalisez la régression avec la seule variable YearsMarried, que constatez-vous ?

   c. Réalisez la régression avec les deux variables Age et YearsMarried, que constatez-vous maintenant ? C'est étrange non ?

   d. Construisez un graphique avec en abscisse l'âge, en ordonnée les années de mariage. Coloriez les points avec la classe d'appartenance (Z). Attention, de nombreux point seront superposés, utilisez le principe du jittering (cf. jitter) pour rendre le graphique plus lisible. Que constatez-vous ?

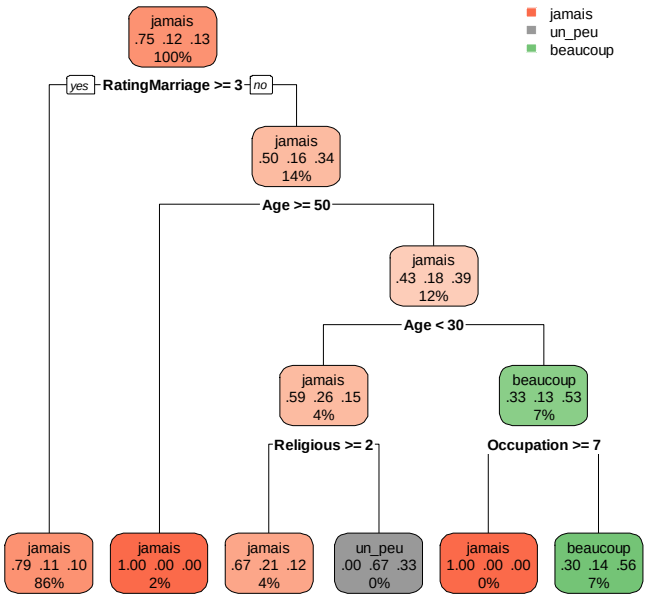


7. On souhaite analyser maintenant l'infidélité en constituant 3 blocs [« jamais » : Y = 0 ; « un peu » : Y = 1 à 3 ; « beaucoup » : Y = 7 et 12].

a. Recoder Y en conséquence en créant la variable G. Combien y a-t-il d'observations dans chaque groupe ? (resp. 451, 70, 80)

b. Quelle analyse mèneriez-vous pour expliquer l'infidélité sous cet angle ? Quels sont les facteurs déterminants pour différencier les individus dans ce cas ? Quel parallèle feriez-vous avec les résultats de l'analyse précédente ?

Ex.1. Avec un arbre de décision, sans tenir compte du caractère ordinal de la variable cible.



Ex. 2. Avec l'analyse discriminante peut-être.