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

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) :

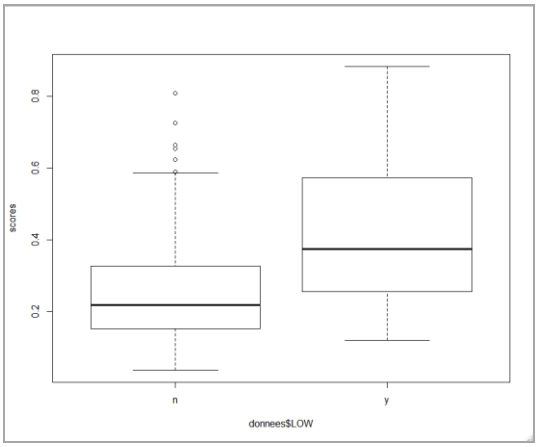
- `LOW : Low Birth Weight (y : < 2500 g ; n : >= 2500 g)`
- `AGE : Age of Mother (years)`
- `LWT : Weight of Mother at Last Menstrual Period (pounds)`
- `SMOKE : Smoking Status During Pregnancy (1 : yes, 0 : no)`
- `HT : Hystory of Hypertension (1 : yes, 0 : no)`
- `UI : Presence of Uterine Irritability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

3.1 Importation, modélisation, probabilités d'affectation aux classes

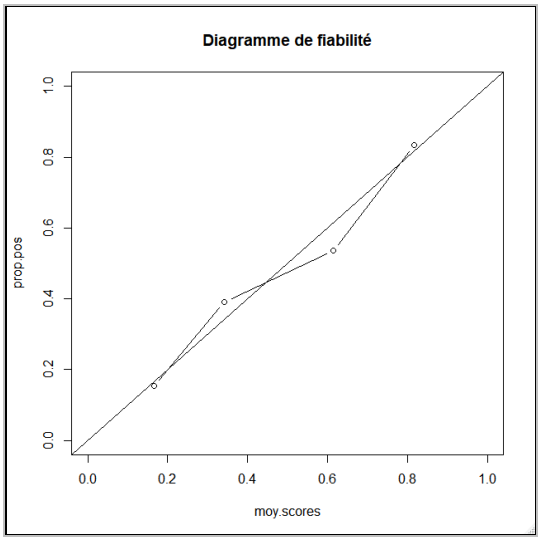
0. Importez le fichier « **Low_Birth_Weight_Data.xlsx** » (read.xlsx si package 'xlsx') (**TUTO 1**, page 7).
1. Affichez les premières lignes des données (head), puis leur description (str).
2. Réalisez la régression de LOW vs. l'ensemble des autres variables (`LOW ~ .`). Le modèle est-il globalement significatif à 5 % ? (**TUTO 1**, pages 8 et 9) (oui).
3. Affichez les caractéristiques détaillées du modèle (summary) (**TUTO 1**, page 8). 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 ; **TUTO 1**, page 10 ; ou encore voir la propriété $fitted.values de l'objet modèle).
5. Affichez la description de ce nouveau vecteur des scores. Quelles sont les valeurs minimales ? Maximales ? Médianes (0.03686, 0.88381, 0.26).
6. Affichez le boxplot de ces scores (boxplot).
7. Affichez le boxplot de ces scores conditionnellement aux classes d'appartenance (LOW). Que constatez-vous ? (boxplot)



3.2 Diagramme de fiabilité

8. Dans cette partie, nous essayons de reproduire les calculs décrits dans notre ouvrage de référence (**TUTO 2**, Section 2.2). Construisez un vecteur représentant le découpage des probabilités d'affectation en 4 groupes de largeur égales. Les bornes des groupes sont (0, 0.25, 0.5, 0.75, 1.0) (cut, voir l'option breaks pour délimiter les intervalles).
9. Calculez les effectifs par groupe (91, 64, 28, 6) (table sur la variable après découpage).
10. Calculez les moyennes des scores par groupe (0.165, 0.341, 0.614, 0.816) (tapply ou aggregate).
11. Calculez la proportion des positifs (individus appartenant à la classe LOW = y) dans chaque groupe (0.153, 0.390, 0.535, 0.833).
12. Construisez le diagramme de fiabilité c.-à-d. un graphique (plot) avec en abscisse la moyenne des scores et en ordonnée la proportion des observations positives. Rajoutez la diagonale principale (abline) pour repérer les éventuelles divergences.
13. Conclusion ? Le modèle est-il bien calibré ? (oui, plutôt).



3.3 Test de Hosmer et Lemeshow

14. Dans cette partie, nous essayons de mettre en œuvre le test de Hosmer et Lemeshow, basé sur la même idée que le diagramme de fiabilité, mais qui s'appuie sur un test statistique pour statuer sur la qualité de l'approximation des données par le modèle (**TUTO 2**, section 2.3).
15. Nous constituons de nouveaux des groupes basés sur les scores, mais en nous appuyant sur les déciles (10 intervalles de fréquences égales) (cut, voir ici comme paramétrer breaks de manière à ce qu'il utilise les quantile des scores).
16. Combien d'observations obtenons-nous dans les 10 intervalles ? (table) (19, 19, 19, 19, 19, 18, 19, 19, 19, 19).
17. Calculez la somme des scores dans chaque groupe (tapply) (1.617, 2.592, …, 13.67)
18. Calculez les différences entre effectifs observés et les sommes des scores dans chaque groupe (19 – 1.617 = 17.382, 19 – 2.592 = 16.407, …, 5.327)
19. Calculez les effectifs des négatifs (LOW = n) et positifs (LOW = y) dans chaque groupe (table) [(19, 0), (15, 4), …, (6, 13)]
20. A partir de ces informations, calculez la statistique de Hosmer et Lemeshow (**TUTO 2**, équation 2.3, 2.4 ou 2.5, au choix) (C = 7.5)
21. En faisant l'hypothèse que cette statistique suit une loi du KHI-2 à (10 – 2 = 8) degrés de liberté. Est-ce que les scores produits par le modèle sont compatibles avec les données observées (p-value = 0.4837).
22. Avec la fonction logitgof du package 'generalhoslem', est-ce que vous retrouvez ces résultats ? (oui)
23. Et avec la fonction hoslem.test du package 'ResourceSelection' ? (oui itou).