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

N'oubliez pas que vous avez accès aux corrigés de nos précédentes séances !

2 Données

Nous travaillons sur une version de la base TITANIC « **titanic.xlsx** » où l'on cherche à prédire la survie (SURVIVANT = YES ou NO) des voyageurs à partir de leur CLASSE_ (1st, 2nd,3rd ou Crew [membre d'équipage]), AGE_ (Adult ou Child) et SEXE_ (Male ou Female). Dans la première feuille (RAW), les données sont présentées de manière usuelle, la ligne représente un voyageur dont on connaît le sort. Dans la seconde (PATTERN), les données ont été groupées selon les modalités de (CLASSE, AGE et SEXE). Chaque ligne représente un « pattern » (un profil), une combinaison des explicatives. Leur nombre est réduit, et la variable cible est représentée de manière différente. Ce mode d'organisation des données ouvre la porte à un prisme d'analyse original où l'on peut mieux situer plus finement l'influence des explicatives via l'étude des profils.

3 Exercices – Analyse des données groupées

3.1 Importation et régression sur données individuelles

0. Chargez la première feuille (RAW) du fichier « **titanic.xlsx** » (**TUTO 1**, page 7).
1. Affichez les premières lignes pour appréhender la configuration des données (head). Affichez les caractéristiques de la base (str). Quel est le nombre de lignes et de variables ? (2201 obs., 4 variables).
2. Affichez la distribution de fréquence de chaque variable (summary). Combien y-t-il eu de survivants ? (711) De membres d'équipage ? (885) D'enfants ? (109) D'hommes ? (1731)
3. Lancez la régression (SURVIVANT ~ CLASSE_ + AGE_ + SEXE_). Affichez le détail des résultats (summary). Comment ont été traitées les explicatives dans la régression ? Quelle est la modalité de référence pour chaque variable explicative ? Interprétez rapidement les coefficients de la régression. Quelles sont les caractéristiques des personnes qui ont péri dans la catastrophe ?

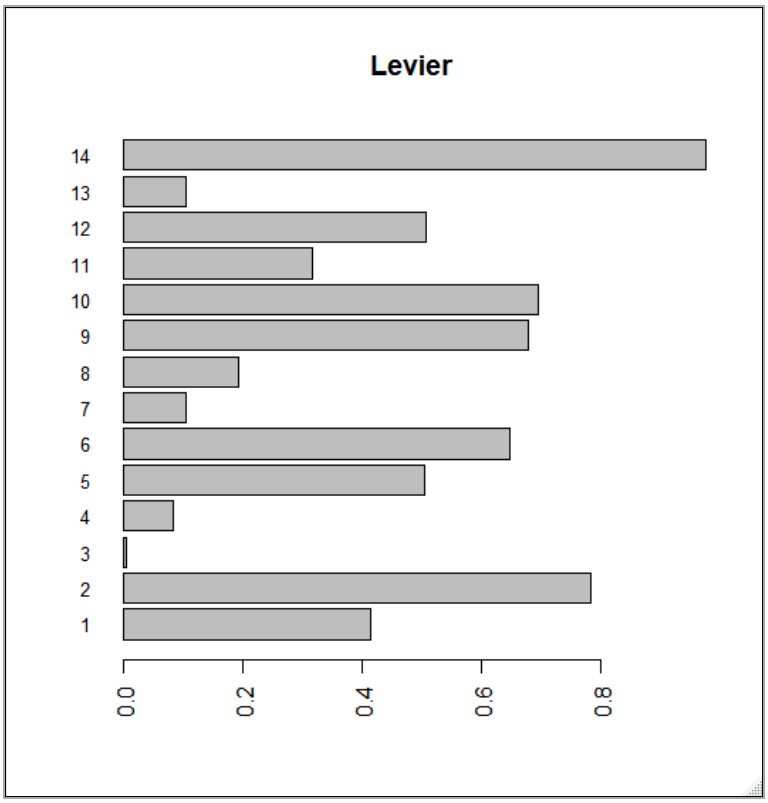
4. Notez attentivement les résultats : coefficients estimés, écarts-type estimés des coefficients, déviance, degrés de liberté. Nous en aurons besoin par la suite.
5. Quelle est le nombre théorique de combinaisons (pattern, profils) possibles des modalités des variables explicatives ? (16)
6. Calculez les effectifs pour chaque pattern (table). Quel est le nombre de (AGE = ADULT, SEXE = FEMALE, CLASSE = 1ST) ? (144) De (AGE = CHILD, SEXE = MALE, CLASSE = 2ND) (11)
7. Il existe des combinaisons pour lesquels nous n'avons aucune observation. Lesquelles ? (2 configurations). Quel est le nombre de « profils » que l'on observe réellement dans notre base finalement ? (14)

3.2 Importation et régression sur données groupées

8. Chargez la seconde feuille (PATTERN) du fichier « **titanic.xlsx** ». Affichez le data frame (print). Comment se présente maintenant la variable dépendante ?
9. Affichez les caractéristiques de la base (str). Combien d'observations disposons-nous maintenant, combien de variables (14 obs., 5 variables).
10. Créez une matrice correspondant aux deux colonnes (dans cet ordre) SURV_YES et SURV_NO (14 lignes, 2 colonnes).
11. Lancez la régression avec cette matrice en variable cible vs. les explicatives (AGE_, CLASSE_, SEXE_) (lire attentivement la documentation de glm). Affichez le détail des résultats et comparez-les avec ceux de la régression sur données individuelles ci-dessus. Que constatez-vous ?

3.3 Levier

12. Le levier (« hat value ») donne une indication sur l'influence globale des « pattern » dans la régression. Calculez le levier de chaque observation (hatvalues).
13. La somme des leviers est égale au nombre de paramètres estimés dans le modèle. Est-ce le cas ? (oui)
14. Affichez graphiquement la valeur des leviers. Quels sont les 5 « pattern » qui semblent peser le plus dans la régression ? (je suis parti sur un barplot en ce qui me concerne) A quels profils correspondent ces « pattern » ?



15. En pratique, on se rend compte que la valeur du levier, en régression sur données groupées, repose en grande partie sur les effectifs associés aux profils. Pouvez-vous vérifier cela ? (j'ai calculé la somme des colonnes SURV_YES et SURV_NO et j'ai intégré le tout dans un data frame)

	N	DPat.CLASSE	DPat.AGE	DPat.SEXE
1	144	1ST	ADULT	FEMALE
2	175	1ST	ADULT	MALE
3	1	1ST	CHILD	FEMALE
4	5	1ST	CHILD	MALE
5	93	2ND	ADULT	FEMALE
6	168	2ND	ADULT	MALE
7	13	2ND	CHILD	FEMALE
8	11	2ND	CHILD	MALE
9	165	3RD	ADULT	FEMALE
10	462	3RD	ADULT	MALE
11	31	3RD	CHILD	FEMALE
12	48	3RD	CHILD	MALE
13	23	CREW	ADULT	FEMALE
14	862	CREW	ADULT	MALE

3.4 Résidus de Pearson

16. Calculez la probabilité d'affectation à (SURV_YES) (predict) (**TUTO 1**, page 10)

17. Calculez le résidu de Pearson qui indique la qualité de restitution par le modèle de chaque pattern (**TUTO 2**, section 9.3.1 ; dans la formule 9.2, n est le nombre d'observations pour un pattern « m », Y représente le nombre d'observations positives [SURV_YES], Y^ les

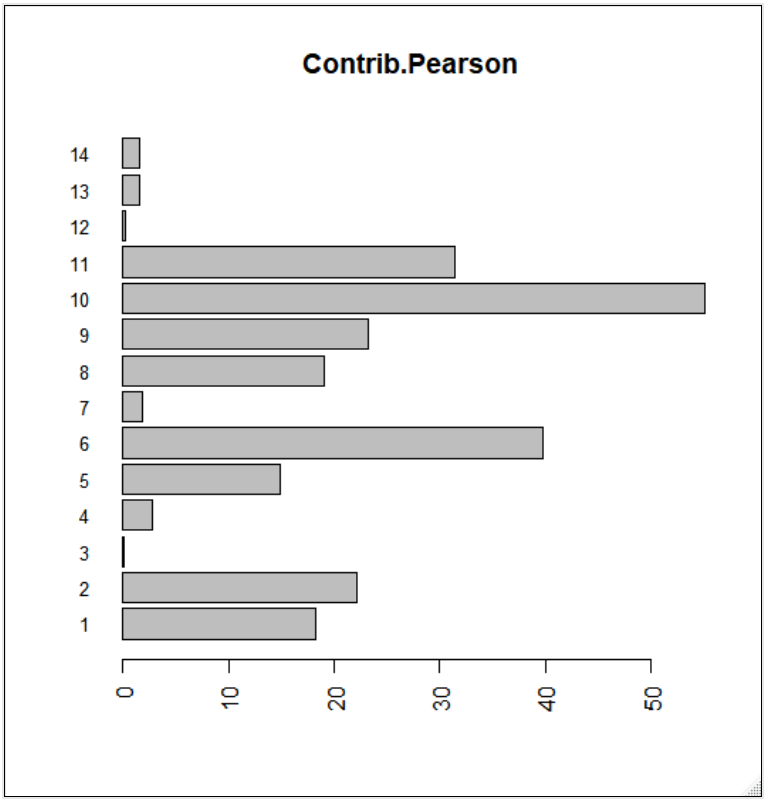
prédictions du même nombre ; PI est la probabilité d'occurrence des observations positives) (nous avons : 3.272, -2.189, …, -0.194)

18. Faites appel à la fonction residuals() en veillant à lui passer les bons paramètres. Obtenez-vous des valeurs de résidus de Pearson identiques à votre calcul précédent ? (oui, gros souci sinon).

19. Calculez la statistique de Pearson (**TUTO 2**, équation 9.3) (103.8296) Est-ce que la régression est compatible avec les données ? (ddl = 8, p-value = 7.026538E-19 ; non hélas)

20. Pour identifier les « pattern » mal modélisés, fauteurs de trouble, nous calculons les résidus standardisés de Pearson (TUTO 2, équation 9.4) (cf. rstandard, soyez attentifs aux paramètres de la fonction).

21. Passez-les au carré pour obtenir les contributions aux KHI-2 de Pearson (**TUTO 2**, équation 9.5). Réalisez un graphique permettant d'identifier les « pattern » qui perturbent la régression. Lesquels sont-ce ?



22. Un « pattern » présente une contribution à la statistique de Pearson importante parce que l'écart entre le nombre de survivant estimé est éloigné de la réalité, et que l'effectif concerné est important. Est-ce le cas ? (oui)

23. Retranchez les contributions des profils n°10 et n°6 de la statistique de Pearson. Est-ce que l'approximation de (SURV = YES) avec le modèle serait satisfaisante dans ce cas ? (oui ; KHI-2 = 9.05, ddl = 6, p-value = 0.1707)

3.5 Résidu déviance

24. Réitérer les opérations de la section précédente avec le résidu déviance (**TUTO 2**, section 9.3.2) (utilisez directement residuals et rstandard, avec les paramètres appropriés) : calculez la statistique déviance, puis la contribution à la déviance des profils.

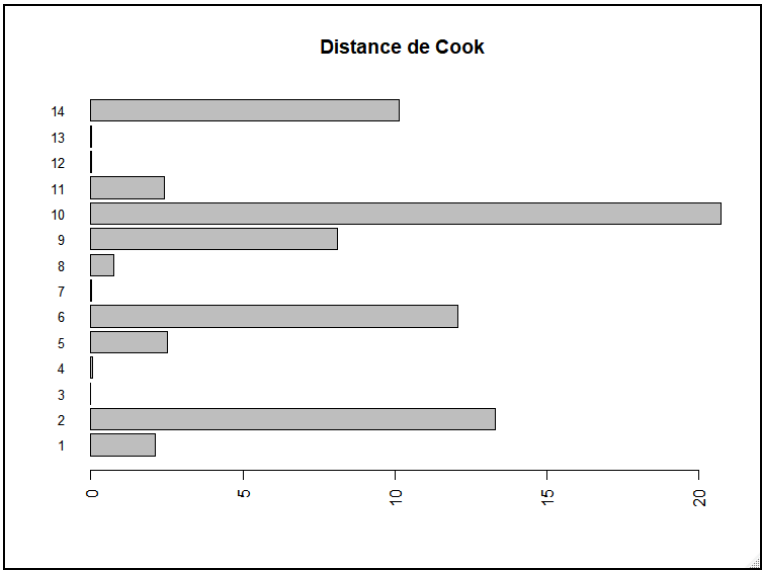
25. Rapprochez la statistique déviance avec le champ $deviance de la régression sur les données groupées, que constatez-vous ? (on obtient bien les mêmes valeurs)

26. Quels sont les profils problématiques au sens de la déviance ? (toujours n°10 et n°6)

3.6 Distance de Cook

27. Pour identifier les profils qui ont le plus d'impact sur les coefficients estimés, nous souhaitons calculez la distance de Cook (**TUTO 2**, section 9.4.1) (cooks.distance)

28. Quels sont les 5 profils qui engendrerait le plus de modifications des coefficients estimés si nous les retirions de la régression ?



29. Au regard des profils concernés, on se dit qu'il y a quelque chose autour de l'interaction entre CLASSE et SEXE. On souhaite l'introduire dans la régression. Comment faire ?

30. Est-ce que l'introduction de cette interaction est pertinente ? (un test du rapport de vraisemblance de significativité des coefficients associés peut-être ?) (oui)

31. Est-ce que pour autant le modèle explique convenablement la probabilité de survie au sens du test de la déviance ? (non)