Nous travaillons sous R. Tous les tests sont à 5%. Tous les intervalles de confiance sont au niveau de confiance 95%.

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/cours/Dependance_Variables_Qualitatives.pdf

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

2. Données

Le fichier « **heart.xlsx** » retranscrit les caractéristiques (âge, sexe, etc.) d'un ensemble de personnes, atteintes ou pas d'une maladie cardiaque (coeur ϵ {absence, présence}).

Le fichier original est disponible sur le serveur UCI http://archive.ics.uci.edu/ml/datasets/statlog+(heart). Il a été légèrement modifié pour les besoins du TD.

3. Exercices - Analyse des risques cardiaques

a. Risque relatif et odds-ratio

0. Chargez les données dans un data frame (read.xlsx, si package 'xlsx'). Affichez la description de la base (str). Combien y a-t-il d'observations et de variables ? (270 obs., 13 variables).
1. Affichez les distributions de fréquences des variables « angine » et « cœur » (summary ou table). Combien de personnes présentent une angine de poitrine (angine = oui) ? (89) Combien présentent une maladie cardiaque (cœur = présence) ? (120)
2. Calculer le KHI-2 du test d'indépendance entre « cœur » et « angine » (**TUTO 1**, section 2.1). Vu les effectifs, il n'est pas nécessaire d'introduire le test de continuité (chisq.test) (KHI-2 = 47.47, ddl = 1). Que conclure ?
3. Construire le tableau croisé entre « cœur » (ligne) et « angine » (colonne) (table). Calculez les profils colonnes (prop.table). Quelle est la proportion des personnes malades (cœur = présence) parmi les « angine = non » (0.2983) ? Parmi les « angine = oui » (0.7415) ?
4. Calculez le risque relatif d'être malade (cœur = présence) lorsque la personne est atteinte d'une angine (angine = oui) (**TUTO 1**, section 5.3) (2.48)

5. Calculez les bornes de l'intervalle de confiance du risque relatif (**TUTO 1**, section 5.3.2) ([1.926 ; 3.207]). Peut-on en conclure que le risque est significativement différent de 1 ?
6. Calculez le même risque relatif par l'entremise du package 'epitools' (riskratio, attention, on s'appuie sur l'approximation normale de la distribution du logarithme du risque, cf. l'option method). Obtient-on les mêmes valeurs que précédemment (estimation ponctuelle + intervalle de confiance) ? (oui)
7. Calculez l'odds-ratio pour la même configuration (**TUTO 1**, section 5.5) (6.748)
8. Calculez les bornes de l'intervalle de confiance (**TUTO 1**, section 5.5.4) ([3.81 ; 11.95]).
9. Refaites les mêmes calculs avec 'epitools' (oddsratio). Les résultats sont cohérents ? (oui)

b. Régression logistique et odds-ratio (var. indép. binaire)

10. Réalisez une régression logistique (cœur ~ angine) (glm). Récupérez l'objet fourni par summary. Comment a été traitée la variable « angine » dans la régression ?
11. Affichez-en les propriétés (attributes). Nous nous intéressons plus particulièrement au champ '$coefficients'. Quel est son type ? (class) (matrix)
12. Affichez son contenu et ses dimensions (dim) (2 lignes x 4 colonnes)
13. Affichez les en-têtes de lignes (rownames) et de colonnes (colnames). Comment accéder alors au coefficient (Estimate) de l'indicatrice (angineoui) ?
14. Calculez l'exponentielle du coefficient estimé de l'indicatrice « angineoui » dans la régression (6.748). Peut-on rapprocher ce résultat avec ceux calculés plus haut ? (oui, l'odds-ratio).
15. Calculez les bornes de l'intervalle de confiance du coefficient de « angineoui » ([1.337 ; 2.48])
16. Passez ces bornes à l'exponentielle. Que constatez-vous ?

c. Régression avec deux explicatives binaires

17. On passe à la régression logistique multivariée. On souhaite expliquer « cœur ~ angine + sexe ». Comment sont traitées les 2 variables explicatives dans la régression ? (transformées en indicatrices 0/1 : angine = oui et sexe = masculin).
18. Calculez l'exponentielle du coefficient de « angineoui » (6.16). Pourquoi l'estimation de l'odds-ratio obtenue ici est différente de celle issue de la régression simple ?
19. Les variables sexe et angine sont-elles liées ? (chisq.test) (oui)
20. Calculez l'odds-ratio (cœur vs. angine) chez les femmes (7.375) et chez les hommes (5.769). Que constatez-vous ? (utilisez oddsratio de 'epitools' pour aller à l'essentiel)

21. Réalisez la régression en introduisant le terme d'interaction entre sexe et angine. Est-il significatif à 5% ? (non) Comment faut-il lire ce résultat ?

d. Régression avec une explicative nominale (> 2 modalités)

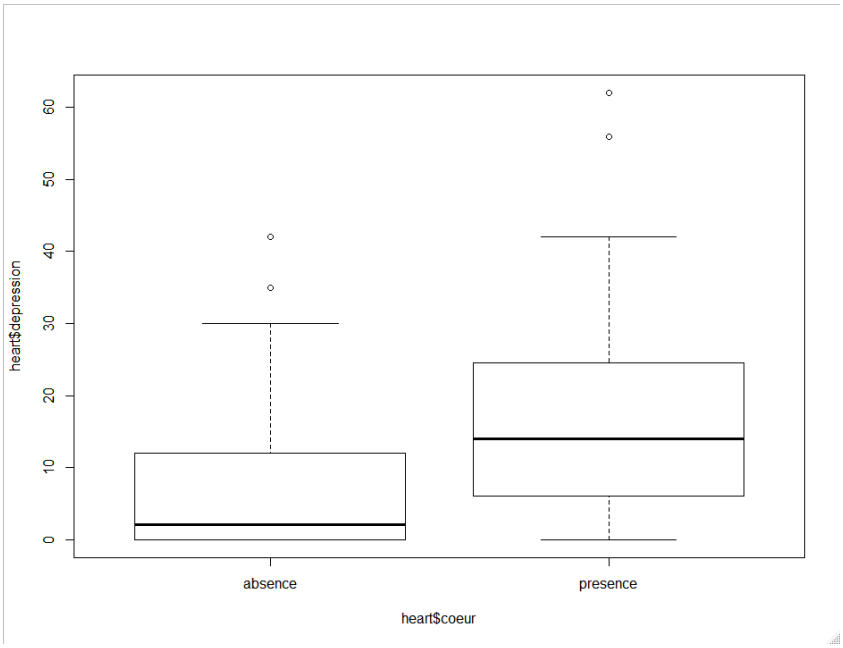
22. Nous nous intéressons à la variable « vaisseau ». Enumérez ses modalités et calculez sa distribution de fréquences (A, B, C, D ; 160, 58, 33, 19).
23. Réalisez la régression « cœur ~ vaisseau ». Quelle est la modalité de référence. Comment lire les résultats ? Les coefficients des indicatrices en particulier ?
24. Passez à l'exponentiel les coefficients de la régression. Comment lire maintenant les valeurs liées aux indicatrices ? (B : 5.7 ; C : 11.14 ; D : 16.0)
25. Avec la fonction oddsratio de 'epitools', calculez les odds-ratio des modalités de la variable « vaisseau ». Un rapprochement avec les résultats de la régression logistique est possible ? (oui, exponentiel des coefficients des indicatrices).

e. Régression avec une explicative ordinale (> 2 modalités)

26. On vient nous dire que la variable « vaisseau » est en fait ordinale. L'ordre des modalités est (A < B < C < D). Construisez le tableau croisé entre cœur (ligne) et vaisseau (colonne).
27. A partir de ce tableau, calculez les odds-ratio des modalités relativement à la précédente (**TUTO 2**, section 5.2.4) (B / A : 5.7 ; C / B : 1.95 ; D / C : 1.43). Quels rapprochements pouvons-nous faire avec la même analyse mais où nous considérions « vaisseau » comme qualitative nominale ?
28. Recodez la variable vaisseau en ordinale (ordered). Affichez son type (class) et les valeurs. Que constatez-vous ? Notamment dans l'affichage des modalités ?
29. Recodez cette nouvelle variable en 3 indicatrices emboîtés (**TUTO 2**, section 5.2.4) (vais_b, vais_c, vais_d). Lancez la régression « cœur ~ vais_b + vais_c + vais_d », affichez les résultats.
30. Passez les coefficients des indicatrices à l'exponentielle. Rapprochez les résultats avec les odds-ratio calculés précédemment. Sont-ils cohérents ? (oui).
31. En revenant sur les résultats de la régression. Comment interpréter la significativité des coefficients ?
32. Avec la fonction oddsratio de 'epitools', calculez les odds-ratio des modalités de la variable « vaisseau » rendue ordinale. Que constatez-vous ? L'outil a-t-il pris en compte le caractère ordinal de la variable ?

f. Régression avec une ou des explicatives quantitatives

33. Nous nous intéressons à la variable « dépression ». Construisez les « boxplot » de « dépression » selon « cœur ». Que constatez-vous ? La régression « cœur ~ dépression » a-t-elle des chances d'être probante ?



34. Lancez la régression « cœur ~ dépression ». La variable dépression est-elle significative ? (oui).

35. Calculez l'exponentielle du coefficient associé (1.094). Que représente cette nouvelle valeur ? (**TUTO 2**, section 5.2.2).

36. Nous souhaitons associer la variable « âge » à dépression. Lancez la régression « cœur ~ âge + dépression ». Commentez les résultats.

37. Calculez l'odds-ratio associé à dépression (1.089). La valeur est différente de précédemment dans la régression simple ? Pourquoi ? (**TUTO 2**, section 5.3.1).

38. Quelle est la variable qui a le plus d'impact dans la régression, « âge » ou « dépression » ? Pourquoi ?

39. Pour s'en assurer, nous souhaitons calculer les coefficients standardisés (**TUTO 2**, section 5.3.2 ; Solution 1 pour la régression logistique, équation 5.4) (âge : 0.32 ; dépression : 0.977).

40. Comment s'interprètent ces nouvelles valeurs des coefficients.

41. Si on vous dit que la dépression a 3.05 fois plus d'impact que l'âge sur le LOGIT, vous êtes d'accord ?