ENSAI - 2A — Année 2024-2025

Chargés de TD/TP :

Koffi Amezouwui — koffi.amezouwui@ensai.fr

Julien Jamme — julien.jamme@insee.fr

Théo Leroy — theo.leroy@insee.fr

Clotilde Patarin — clotilde.patarin@square-management.com

Responsable du cours :

Frédéric Lavancier — frederic.lavancier@ensai.fr

Régression linéaire et généralisée - TD 6-7-8

Exercice 1. Le jeu de donnée "chdage.txt", disponible dans Moodle, contient les données de 100 patients âgés de 20 à 69 ans (variable *age*), présentant pour certains une maladie coronarienne (variable *chd* valant Yes ou No).

1. Importer ces données sous `R` sous la forme d'un `data.frame`. Observer le contenu de chaque variable et transformer leur `class` si nécessaire.

2. Proposer une ou des visualisation(s) graphique(s) permettant d'analyser le lien éventuel entre les variables *chd* et *age*.

3. Un lien est-t-il apparent ? Quel test statistique simple permettrait de le confirmer ?

4. On note $Y = \mathbf{1}_{chd=Yes}$ et $p(x) = \mathbb{P}(Y = 1|age = x)$. Donner la loi de Y sachant que $age = x$ en fonction de $p(x)$.

5. En s'appuyant sur cette loi, donner la vraisemblance des observations $(y_1, \dots, y_n)$ en fonction de $(p(x_1), \dots, p(x_n))$ où x_i désigne l'âge de l'individu i et $y_i = 1$ si ce dernier a une maladie coronarienne.

6. En guise de première estimation de $p(x)$, on met en oeuvre la démarche suivante :

   (a) Utiliser les 8 groupes d'âge proposés via la variable *agegrp* du jeu de données, ce qui forme 8 groupes d'individus associés à ces classes. Calculer $\bar{x}_1, \dots, \bar{x}_8$ le milieu de chaque classe d'âge.

   (b) Calculer les proportions de $chd = Yes$ dans chaque groupe, notées $\hat{p}_1, \dots, \hat{p}_8$ (on pourra utiliser les fonctions `table` et `prop.table`).

7. Afin d'analyser la qualité de l'estimation précédente, transformer la variable *chd* en variable numérique prenant les deux valeurs 0 et 1, et représenter sur un même graphique le nuage de points de *age* et *chd* (recodé) et les proportions estimées dans chaque classe $(\bar{x}_k, \hat{p}_k)$ pour $k = 1, \ldots, 8$.
   Quelles vertus et quelles limites cette procédure d'estimation a-t-elle ?

8. On décide de modéliser $p(x)$ à l'aide d'un modèle de régression logistique de paramètre $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$. Quelle hypothèse cela signifie-t-il sur l'expression de $p(x)$? Cela est-il compatible avec le graphique précédent ? Quelle(s) autre(s) alternative(s) de modélisation pourrait-on suggérer ?

9. Ecrire la log-vraisemblance du modèle logistique en fonction de β et en déduire le système que doit résoudre le maximum de vraisemblance $\widehat{\beta}$. Peut-on résoudre ce système analytiquement ?

10. Calculer l'estimateur du maximum de vraisemblance via la fonction `glm`.

11. Rappeler la définition théorique du rapport de cotes (odds ratio) d'avoir une maladie coronarienne entre un individu d'âge x_1 et un individu d'âge x_2. Que vaut-il pour le modèle précédent lorsque les 2 individus ont 10 ans d'écart (c'est à dire $x_1 = x_2+10$) ?

12. On s'intéresse à présent au rapport de probabilités (et non de cotes) d'avoir une maladie coronarienne entre un individu d'âge x_1 et un individu d'âge x_2. Que vaut ce rapport pour le modèle précédent lorsque les 2 individus ont 10 ans d'écart (c'est à dire $x_1 = x_2 + 10$) ? On pourra représenter ce rapport en fonction de x_2, pour x_2 prenant des valeurs de 20 à 70 ans.

On rappelle que sous "de bonnes conditions", l'estimateur du maximum de vraisemblance $\hat{\beta}$ dans un modèle de régression logistique portant sur p variables explicatives et n individus vérifie la convergence en loi suivante, lorsque $n \to \infty$,

$$J_n(\beta)^{1/2} \left(\widehat{\beta} - \beta \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0_p, I_p),$$

où $J_n(\beta)$ est la matrice d'information de Fisher. Cette dernière admet l'expression

$$J_n(\beta) = \mathbf{X}' W_\beta \mathbf{X}$$

où $\mathbf{X}$ est la matrice de design et W_β est la matrice diagonale

$$W_\beta = \begin{pmatrix} p_\beta(x_1)(1 - p_\beta(x_1)) & & 0 \\ & \ddots & \\ 0 & & p_\beta(x_n)(1 - p_\beta(x_n)) \end{pmatrix}.$$

13. Justifier qu'une approximation asymptotique de la loi de $(\widehat{\beta} - \beta)$ est $\mathcal{N}(0_p, J_n^{-1}(\beta))$.

14. Comment peut-on exploiter ce résultat pour estimer l'écart-type de chaque coordonnée de $\hat{\beta}$? On donnera la démarche concrète à appliquer, mais on ne demande pas de la mettre en pratique numériquement.

15. Cette procédure d'estimation est utilisée par la fonction `glm`. En utilisant sa sortie, donner une estimation de l'écart-type de $\hat{\beta}_0$ et $\hat{\beta}_1$.

16. Construire un intervalle de confiance asymptotique à 95% pour le paramètre β_1.

17. D'après la question précédente, le paramètre β_1 est-il différent de 0 au seuil d'erreur asymptotique 5% ? Quel est le nom de cette procédure de test ? Donner la p-value associée à ce test et vérifier qu'elle concorde bien avec la sortie de `glm`.

18. Calculer la statistique du test de déviance de significativité du modèle GLM (par rapport au modèle nul). En déduire la p-value et conclure aux seuils d'erreur 10%, 5% et 1%. Comparer avec les résultats du test effectué sous `R` à l'aide de la fonction `anova` appliqué au modèle, avec l'option `test="Chisq"`.

19. Reprendre le graphique effectué à la question 7 et superposer (sous forme d'une courbe) les valeurs prédites $\hat{p}(x)$ par le modèle logistique, calculées pour une grille de valeurs de x couvrant l'étendue prise par les observations. On pourra utiliser la fonction `predict` associée à l'option `type="response"`.

20. En partant de la convergence en loi de $\hat{\beta}$ rappelée ci-dessus, en déduire un intervalle de confiance au niveau asymptotique 95% pour $p(x)$. Ajouter cet intervalle de confiance pour chaque x considéré au graphique précédent. On pourra exploiter avec profit l'option `se=TRUE` de la fonction `predict` dans le cas `type="link"`.

Exercice 2 (Le modèle logistique est naturel)**.** On dispose d'un couple de variables aléatoires (X, Y) où Y est binaire et X est à valeurs dans $\mathbb{R}^d$. On note $p = \mathbb{P}(Y = 1)$, $f_0(.)$ la densité conditionnelle de X sachant que $Y = 0$ et $f_1(.)$ la densité conditionnelle de X sachant que $Y = 1$. On note de plus

$$h(x) = \log \frac{f_1(x)}{f_0(x)} + \log \frac{p}{1-p}, \quad x \in \mathbb{R}^d.$$

1. Montrer que pour tout x

$$\mathbb{P}(Y = 1|X = x) = \frac{1}{1 + e^{-h(x)}}.$$

2. On rappelle qu'une loi sur $\mathbb{R}^d$ fait partie de la famille exponentielle si sa densité s'écrit $a(x)b(\theta)e^{\theta' T(x)}$, pour un certain paramètre $\theta \in \mathbb{R}^q$, où a et b sont des fonctions positives et $T : \mathbb{R}^d \to \mathbb{R}^q$ est une fonction appelée statistique suffisante.
   Montrer que si les densités conditionnelles f_0 et f_1 font partie de la même famille exponentielle et ne diffèrent que par la valeur de leur paramètre associé, alors $\mathbb{P}(Y = 1|X = x)$ suit exactement un modèle de régression logistique, dont on précisera le paramètre et les variables.

Exercice 3 (Le modèle logistique est naturel (bis))**.** Préambule : L'entropie est une quantité qu'on trouve en thermodynamique pour mesurer l'état de *désordre* ou d'*aléa* d'un système. Dans le même esprit, on la trouve également en théorie de l'information et en probabilité pour quantifier le désordre ou la quantité d'aléa qu'intègre une loi de probabilité. Un système physique a tendance à évoluer naturellement vers un état d'entropie maximale. Suivant ce principe, il est naturel, pour décrire une expérience aléatoire donnée, de choisir des lois de probabilité qui maximisent l'entropie. C'est ce principe que nous allons appliquer pour chercher à choisir au mieux $\mathbb{P}(Y = 1)$ lorsque Y est binaire.

Mathématiquement, étant donné Y une variable binaire et $p = \mathbb{P}(Y = 1)$, l'entropie de la loi de Y vaut

$$-p \log(p) - (1 - p) \log(1 - p).$$

L'entropie d'un vecteur de variables binaires indépendantes $Y_1, \ldots, Y_n$ est simplement la somme des entropies individuelles.

1. Sans aucune source de contrainte, quelle est la loi d'entropie maximale d'une variable binaire ?

2. Supposons à présent qu'on dispose d'un échantillon de n couples (Y_i, X_i) où X_i est une variable aléatoire dans $\mathbb{R}^d$. On note $p_i(x_i) = \mathbb{P}(Y_i = 1|X_i = x_i)$, $i = 1, \ldots, n$. A priori, sans utiliser aucune information contenue dans l'échantillon, que valent les $p_i(x_i)$ qui maximisent l'entropie ?

3. On souhaite trouver les $p_i(x_i)$ qui maximisent l'entropie tout en étant cohérentes avec les observations. Cela revient à inclure des contraintes sur les $p_i(x_i)$ possibles.

On choisit les contraintes :

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} p_i(x_i) x_i.$$

(Puisque x_i est un vecteur de taille d, il s'agit bien d'un système de d contraintes). Ces contraintes sont assez naturelles : on souhaite que la moyenne des x_i des individus du groupe positif ($y_i = 1$) coincide avec la moyenne des x_i pondérée par la probabilité que y_i vale 1. En particulier (pour la variable constante 1), on souhaite que la proportion des $y_i = 1$ coincide avec la somme des probabilités.
Trouver les $p_i(x_i)$ qui maximisent l'entropie tout en satisfaisant les contraintes précédentes. On pourra donner la solution à une constante (vectorielle) inconnue près.

4. Quel rapport avec la régression logistique ?

Exercice 4 (Données `break`)**.** On considère les données `break`, disponibles sur la page Moodle du cours. Ce jeu de données, de taille $n = 33$, comporte trois variables ayant trait à l'état d'une automobile :

- Une variable `fault` qui vaut 1 si la voiture concernée a connu une panne, 0 sinon ;
- Une variable `age` qui donne l'âge de la voiture ;
- Une variable `brand` qui donne la marque de la voiture.

Le but de l'exercice est de modéliser la variable `fault`.

1. Importer les données sous `R` et recoder la classe de chaque variable si nécessaire.
2. Observer graphiquement le lien éventuel entre `fault` et `age` d'une part, et entre `fault` et `brand` d'autre part.
3. On souhaite mettre en oeuvre un modèle de régression logistique expliquant la probabilité d'avoir une panne en fonction de l'âge de la voiture et de sa marque. Lancer cette modélisation sous `R` et écrire la formule mathématique du modèle obtenu. On précisera en particulier le modèle estimé spécifique aux voitures de la marque 0, puis de la marque 1, puis de la marque 2.
4. Analyser la qualité globale du modèle.

5. Recommencer la modélisation en incluant seulement la variable `age` dans le modèle. Est-ce mieux ?

6. Le nuage de points entre la variable `age` et la variable `fault` semble suggérer que les pannes ont lieu en début de vie du véhicule ("défauts de jeunesse" ou de "rodage") et en fin de vie (pannes d'usure). Ce comportement de la probabilité de pannes en forme de parabole nous incite à essayer d'inclure un terme quadratique en la variable `age`. Effectuer cet ajout dans le modèle en incluant également la variable `brand` et analyser les résultats. Le modèle est-il significatif ?

7. Que vaut l'odds ratio associé à la marque "2" par rapport à la marque "0" de la variable `brand` ? Interpréter cette valeur et donner un intervalle de confiance à 95% autour de cette estimation. L'inclusion de la variable `brand` dans le modèle est-il pertinent ?

Exercice 5 (Données `mental`)**.** On considère les données `mental`, contenues dans le fichier `mental.txt` disponible sur la page Moodle du cours. Ce jeu de données, de taille $n = 40$, est extrait d'une étude sur la santé mentale d'adultes vivant dans le comté américain d'Alachua, en Floride. Il contient trois variables :

- Une variable `impair` décrivant l'état mental de la personne concernée, de 1 (sain) à 4 (en mauvaise santé),
- Une variable `ses` qui vaut 1 si la personne a un statut socio-économique élevé, 0 sinon,
- Une variable `life` mesurant le nombre et l'intensité des bouleversements qu'a connus la personne au cours des trois dernières années, de 0 (aucun changement) à 9 (changements très importants).

Le but de l'exercice est de modéliser la variable `impair`.

1. Importer les données sous `R`. La variable `ses` est-elle qualitative ou quantitative ? Même question pour la variable `life`. Changer leur classe sous R si besoin.

2. Effectuer une petite étude descriptive pour identifier un lien éventuel entre la variable `impair` et les autres variables du jeu de données.

3. Ecrire mathématiquement le modèle de régression logistique cumulatif proportionnel sans terme d'interaction liant `impair` à `ses` et `life`, et permettant d'estimer les probabilités que `impair` $= 1$, ≤ 2 et ≤ 3. Combien ce modèle a-t-il de coefficients ?

4. Estimer les coefficients de ce modèle à l'aide de la fonction `vglm` du package `VGAM`, associée à l'option `family = cumulative(parallel=TRUE)`. Contrôler que le nombre de coefficients est bien celui attendu.

5. Ecrire la définition mathématique de l'odds ratio associé à la variable `ses` et interpréter cette quantité.

6. Donner un intervalle de confiance asymptotique au niveau de confiance 95% pour le paramètre lié à la variable `ses`

7. En déduire un intervalle de confiance asymptotique au niveau de confiance 95% pour l'odds ratio associé.

8. Y a-t-il une influence du statut socio-économique sur l'état de santé mentale au seuil 5% ? Et au seuil 10% ?

9. On souhaite voir si un modèle plus complexe s'ajusterait mieux aux données. Ecrire mathématiquement puis estimer un modèle de régression logistique cumulatif proportionnel avec terme d'interaction. Interpréter le résultat obtenu. Ce modèle est-il significativement meilleur, au seuil d'erreur asymptotique 5% ?

10. Même question avec un modèle de régression logistique cumulatif sans structure proportionnelle, mais sans terme d'interaction.

11. Inversement, pourrait-on proposer un modèle plus simple ?

12. Pour finir, on décide de ne pas exploiter le fait que la variable `impair` est ordinale, et de la modéliser par un modèle logistique multinomial. Comparer cette approche à la modélisation précédente.

Exercice 6 (Données `Fourmis`)**.** Le but de l'étude est d'étudier la diversité des fourmis sur le site expérimental des Nourragues en Guyane Française. On a prélevé 1 m^2 de litière en plusieurs endroits de 4 forêts différentes (la forêt de plateau `GPWT`, la foret de liane `FLWT`, la foret de transition `FTWT`, et la forêt d'Inselberg `INWT`). Chaque échantillon a été

pesé (variable `Weight`) et le nombre d'espèces différentes présentes dans l'échantillon a été relevé (variable `Effectif`). Enfin les conditions de recueil (humides ou sèches, variable `Conditions`) ont été notées pour tester leur influence sur la présence des fourmis.

1. Importer les données sous `R` et transformer les variables catégorielles en facteur.

2. Effectuer une petite étude descriptive pour identifier un lien éventuel entre le nombre d'espèces observé et les autres variables à disposition.

3. Modéliser par un modèle log-linéaire de Poisson la variable `Effectifs` en fonction de toutes les variables à disposition, en incluant toutes leurs interactions possibles. Analyser la sortie du modèle.

4. En utilisant la fonction `step`, effectuer une sélection stepwise backward du meilleur sous-modèle du modèle précédent selon le critère AIC, puis selon le critère BIC. Effectuer de même une sélection stepwise forward. Comparer les choix obtenus.

5. L'incohérence dans les sélections backward et forward précédentes suggère qu'il y a peut-être un sous-modèle alternatif (non testé par ces algorithmes) encore meilleur. Cela nous incite à effectuer une sélection exhaustive du meilleur sous-modèle, comme le propose la fonction `regsubsets` en régression linéaire. Malheureusement cette dernière ne fonctionne pas avec le modèle de Poisson. Si on voulait implémenter cette sélection exhaustive nous-même, justifier qu'il y aurait 30 sous-modèles (avec constante) à tester, en comptant le modèle le plus général.

6. On admet qu'à l'issue d'une telle sélection exhaustive, le meilleur sous-modèle au sens de l'AIC et du BIC est celui dont les coefficients de Weight sont déclinés en autant de modalités croisées que contiennent les facteurs Site et Conditions (c'est à dire 8), mais qui a une ordonnée à l'origine identique pour toutes les modalités croisées de Site et Conditions. Estimer ce modèle, calculer son AIC et son BIC et comparer avec ceux des modèles sélectionnés précédemment.

7. En guise d'alternative, nous souhaitons essayer d'ajuster un modèle généralisé binomial négatif. Si on inclut toutes les interactions possibles, cette approche semble-t-elle préférable au modèle de Poisson ?

8. Après une sélection exhaustive, on admet que le meilleur sous-modèle binomial négatif au sens de l'AIC fait intervenir les mêmes variables que le meilleur modèle de Poisson. Par contre, pour le critère BIC, il s'agit du modèle ne faisant intervenir que

Weight et Site, dans lequel le coefficient de Weight varie selon Site, mais l'ordonnée à l'origine est constante. Estimer ces deux modèles et calculer leur AIC et BIC.

9. Etant donné l'avis des experts, il semble important que le modèle tienne compte des conditions d'humidité. Quel modèle final retenir ?

10. Ecrire l'équation du modèle retenu suivant les différents sites et les conditions de recueil.

11. D'après le modèle sélectionné, quelle est la probabilité d'observer plus de 15 espèces sur un sol de type INWT par temps sec, basé sur un échantillon de terre qui pèse 10kg ? Même question si le temps est humide.

Exercice 7 (Données `Horshoe Crabs`)**.** Le jeu de données `crabs` contient l'observation de 173 limules femelles (Horshoe crabs). Il s'agit d'animaux marins qui ressemblent à des crabes ayant une forme de fer à cheval. Pour chaque limule femelle, on relève sa couleur `color` (codée de 1 à 4, du plus clair au plus sombre), sa largeur `width`, son poids `weight` et `satell` : le nombre de limules mâles satellites (c'est à dire accrochés à la femelle). La couleur est un signe de l'âge de la limule, cette dernière ayant tendance à s'assombrir au cours du temps. On cherche à modéliser le nombre `satell` en fonction des variables à disposition.

1. Mettre en oeuvre un modèle log-linéaire de Poisson et un modèle binomial négatif. Evaluer leur qualité. On pourra en particulier discuter de la pertinence de considérer la variable `color` comme une variable quantitative ou un facteur.

2. Améliorer la modélisation en tenant compte de l'inflation de zéros.