Université d'Aix-Marseille **Année 2024-2025**
Master 1 Mathématiques Appliquées, Statistique
Cours de classification, fiche 5

Objectif : savoir implémenter et interpréter sous `R` une régression logistique (modèle logit) univariée et bivariée.

Ce TP est une initiation à la régression logistique (modèle logit) avec le logiciel `R`. Vous trouverez de la documentation sur la fonction `glm()` et d'autres fonctions utiles dans le fichier `logitR.pdf` sur AMeTICE.

Nous allons étudier les données *infarctus.txt* issues d'une étude présentées dans le livre de D.G. Kleinbaum et M. Klein[1]. Le jeu de données *infarctus.txt*, disponible sur AMeTICE, est constitué de 609 individus et 3 variables. La table 1 en présente un descriptif. L'objectif est de modéliser au moyen d'un modèle logistique la probabilité d'avoir un infarctus en fonction de l'age et du fait d'avoir de l'hypertension. Un objectif supplémentaire sera de construire une règle permettant de prédire le risque d'infarctus des patients.

colonne	nom	type	description
1	`y`	qualitative	infarctus du myocarde (0=non / 1=oui)
2	`age`	quantitative	age
3	`hpt`	qualitative	hypertension (0=non / 1=oui)

Table 1: Descriptif des variables du jeu de données considéré dans le TP.

Lecture du jeu de données

1. Créer dans le répertoire *tpdc* un nouveau répertoire au nom de *tplogit* et un nouveau projet R du même nom.

2. Après avoir téléchargé et observé le fichier *infactus.txt*, importer les données sous R dans un data.frame *infarctus*. Vous associerez un label à chaque variable en utilisant la fonction `var_label()` du package `labelled` et vous définirez un format pour chacune des modalités de la variable qualitative *hpt* en utilisant la fonction interactive `irec()` du package `questionr`. Recoder la variable en facteur.

Modèle 1 : modèle logistique univarié, variable explicative htp

3. Que représente la table suivante ?

`y \hpt`	oui	non
oui	$n_{11} = 43$	$n_{10} = 28$
non	$n_{01} = 212$	$n_{00} = 326$

[1] D.G. Kleinbaum and M. Klein. 2002. Logistic regression, a self learning text. Springer

4. Calculez les probabilités empiriques d'avoir eu un infarctus du myocarde conditionnellement à chacune des modalités de la variable *hpt*.

5. Calculez la cote empirique d'avoir eu un infarctus du myocarde pour chaque modalité de la variable *hpt*. En déduire le rapport des cotes empirique.

6. On note Y la variable aléatoire valant 1 si le sujet a eu un infarctus du myocarde et 0 sinon. On note π les probabilités conditionnelles $\pi_j = P(Y = 1/hpt = j),\ j = 0, 1$, où $hpt = j$ désigne la population des individus prenant la modalité j pour la variable hypertension. Donner l'expression du modèle logistique expliquant π_j, $j \in \{0, 1\}$ par l'état de l'hypertension.

7. Exprimez, puis calculez les estimations $\hat{\beta}_j$. Comparez les valeurs empiriques et estimées des probabilités d'avoir eu un infarctus.

8. Expliquez les effets des deux paramètres du modèle sur la variation des probabilités d'avoir eu un infarctus.

9. Calculez la déviance du modèle et du modèle vide. En déduire un test de la significativité de l'effet de l'hypertention sur la probabilité de développer un infarctus.

10. En vous servant de la fonction `glm()` estimer le modèle précédent, puis décrire chaque terme de la sortie `R`.

Modèle 2 : modèle logistique univarié, variable explicative age

12. Quelle est la loi de Y sachant $age = x$? En déduire la vraisemblance et la log-vraisemblance de l'échantillon $(Y_1, \cdots, Y_n)$ sachant $(age_1 = x_1, \cdots, age_n = x_n)$.

13. On note $\pi_x = P(Y = 1|age = x)$. Définir le modèle logistique expliquant π_x par l'age.

14. En vous servant de la fonction `glm()` (on nommera l'objet créé modele2), donner la valeur $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ de estimation du maximum de vraisemblance du modèle sur nos données, puis décrire chaque terme de la sortie détaillée (la sortie détaillée s'obtient avec la fonction summary()).

15. Donner une estimation de Σ_B, la matrice de variance-covariance de l'estimateur $B = (B_0, B_1)'$. La matrice estimée sur nos données est accessible avec l'instruction R : summary(modele2)$cov.scaled. Quelle loi de probabilité suit approximativement B lorsque n est grand ?

16. On considère ici la question du test de nullité du coefficient β_1

    (a) **Le test de Wald.** Expliquer le calcul de la statistique de Wald, l'utilisation de la loi du khi-deux et le calcul de la p-valeur. Que concluez-vous d'après les sorties de la fonction glm()?

(b) **Test du rapport de vraisemblance.** Après avoir rappelé brièvement en quoi consiste ce test dans le cas étudié, utiliser la fonction `drop1()` avec l'option test = "LRT" pour réaliser ce test et conclure.

(c) **Test du score.** Même question pour le test du score en utilisant la fonction `drop1()` avec l'option test="Rao".

17. Editer sous R les odd-ratios à partir des coefficients du modèle. Quelle est l'augmentation estimée de l'odd lorsqu'on passe d'une personne à une personne âgée d'une année de plus ? Comment appelle-t-on cette quantité ?

18. Utilisation du modèle pour la prévision

(a) Calculer à la main la valeur estimée $\widehat{\pi(51)}$ de la probabilité d'être atteint d'un intarctus à 51 ans.

(b) Calculer l'intervalle de confiance pour $\pi(51)$ au niveau 95% à partir de la formule du cours.

(c) Au vu de ce résultat à quel groupe, affecteriez-vous une personne âgée de 51 ans: malade ou non malade? Préciser la règle de décision choisie.

(d) Pour éditer les valeurs prédites sous R de la prévision $\hat{\pi}(x)$, on peut utiliser la fonction `predict.glm()` appliquée au modele1 avec l'option type="response". On peut également éditer les écart-types estimés de ces valeurs en rajoutant dans la formule l'option se.fit=TRUE. En utilisant la formule l'intervalle de confiance autour de $logit\hat{\pi}(x)$ p16 du cours, on peut calculer les intervalle de confiances de chaque valeur de $logit\hat{\pi}(x)$ à partir des sorties de `predict.glm()` et en déduire les intervalles de confiance autour des valeurs prédites.

(e) Calculer $\hat{\pi}(51)$ et un intervalle de confiance à 95% autour de $\pi(51)$.

(f) **Représentation graphique**: exécuter le programme suivant:

```
library(glm.predict)
k=length(unique(infarctus$age))
pi=matrix(NA,k,3)
for(i in 1 :k)
pi[i,]=basepredict(modele1, c(1,unique(infarctus$age)[i]),type="simulation")
tab.pi=cbind(unique(infarctus$age),pi) colnames(tab.pi)=c("age","pi","pi.inf",
"pi.sup")
v=order(tab.pi[,1]) ;v
tab.pi[v,]
plot(tab.pi[,1],tab.pi[,2],col=2, cex=0.2)
points(tab.pi[,1],tab.pi[,3],col=3, cex=0.2)
points(tab.pi[,1],tab.pi[,4],col=3, cex=0.2)
```

Commenter le graphique obtenu.

Modèle 3 : modèle logistique univarié, variable explicative agec

19. On considère maintenant le codage de la variable `age` en une variable catégorielle `agec` en 4 classes: "< 45", "$[45, 50[$", "$[50, 60[$", "≥ 60". Construire cette variable en utilisant le code

```
infarctus$agec=cut(infarctus$age, breaks=c(min(infarctus$age),45,50,60,max(infarctus$age)),
right = FALSE, labels=c("<45","45-49","50-59",">=60")
```

20. Etudier la liaison des deux variables `y` et `agec` à l'aide d'un test du chi2.

21. Comme en régression linéaire lorsqu'une variable explicative est catégorielle, on choisit généralement pour éviter le problème de multicolinéarité exacte, une modalité de référence et on introduit dans le modèle uniquement les indicatrices des autres modalités.

    Si nous choisissons pour modalité de référence "< 45", on cherche donc à estimer le paramètre $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$ tel que

    $$\pi(agec) = P(y = 1 \,|\, agec) = \frac{e^{X\beta}}{1 + e^{X\beta}} \text{ avec } X = (1, 1_{45-49}, 1_{50-59}, 1_{\geq 60})$$

    Donner l'expression du modèle précédent sans utiliser l'écriture matricielle. Quelles sont les expressions de $\pi(< 45)$?, de $\pi(45 - 49)$?

22. En utilisant la fonction `glm()` écrire et exécuter le programme permettant d'estimer le modèle précédent. Ensuite :

    (a) Interpréter les résultats

    (b) Comparer ses résultats à ceux obtenus dans la section précédente avec la variable *age* quantitative.

Modèle 4 : modèle logistique multivarié, variables explicatives age et htp

24. On souhaite maintenant prendre en compte conjointement les effets de l'age et de l'hypertension sur la probabilité de développer un infarctus. Définir le modèle logistique expliquant de manière additive π par l'age (`age`) et l'hypertension (`hpt`). On fixera comme modalité de référence pour htp la modalité "non".

25. En vous servant de la fonction `glm()` estimer le modèle, puis interpréter les résultats obtenus.

26. Si c'était la modalité "oui" qui constituait la référence, quelles seraient les modifications dans la sortie `R` ? On pourra mettre en oeuvre le modèle, en utilisant la fonction `relevel()` permettant de fixer la modalité de référence de htp à "oui".

27. Comparer la valeur de l'OR de l'age à la valeur obtenue dans le modèle 2. Donner l'explication de la différence entre les deux valeurs obtenues. Comment appelle-t-on ce phénomène ?

28. A partir des statistiques de déviances des sorties de `glm()`, réaliser un test de déviance (du rapport de vraisemblance) de significativité du modèle global. En utilisant la fonction `logitgof()` du package `generalhoslem`, effectuer également un test de Hosmer-Lemeshow. Interpréter

les résultats obtenus.

29. On peut utiliser le modèle logistique pour affecter un individu à une des classes définies par la variable réponse Y. Proposer une règle d'affectation. Définir les notions de vrai positif, faux positif, vrai négatif, faux négatif, sensibilité et spécificité. En utilisant la fonction `roc()` du package `pROC`, construire la courbe ROC associée au modèle et donner l'AUC. Que représente cette courbe et en quoi est-elle indicatrice de la qualité de la règle d'affectation du modèle ? conclure sur la qualité de cette règle