

LSTAT2100 - Exercices - Série 3

Énoncés

Pour ce TP nous allons utiliser les données de 200 patients issus d'un hôpital. Le jeu de données se trouve dans le fichier [dt.csv](#). Il s'agit de données récoltées au cours d'une étude clinique sur une certaine maladie, dont la description précise n'a aucun intérêt ici.

Partie 1

Pour l'instant, nous n'utiliserons que les variables suivantes:

- **STA**: Variable binaire indiquant si le patient est décédé (1) ou pas (0).
- **AGE**: L'âge du patient au début de l'étude.

Commencez par charger le jeu de données dans R puis *examinez sa structure*.

(a) À partir de la variable **AGE**, créez la variable **AGECAT** qui correspond aux catégories d'âge suivantes:

(15 – 24] (24 – 34] (34 – 44] (44 – 54] (54 – 64] (64 – 74] (74 – 84] (84 – 94]

Quelle est la proportion de décès par catégorie d'âge? Visualisez graphiquement ces chiffres. Que pouvez-vous conclure quant à l'effet de l'âge sur la survie des patients ?

(b) Réalisez un scatterplot de **STA** versus **AGE**. Ajoutez la droite des moindres carrés à votre graphique et calculez le coefficient de détermination (R^2 classique).

(c) Écrivez l'équation du modèle logistique reliant $Y=STA$ (réponse) et $X=AGE$ (prédicteur). *L'événement de "succès" ($Y = 1$) doit être le décès*. Écrivez le log-vraisemblance de votre modèle et maximisez-le **numériquement** (sans passer par la fonction `glm()`) en utilisant la fonction `optim()`. Celle-ci permet de trouver le *minimum* d'une fonction donnée; voir le Help pour plus de détails.

(d) Utilisez la fonction `glm()` pour estimer le modèle tel que défini en (c). Comparez avec les estimations obtenues avec `optim()`. Visualisez le modèle à l'aide d'un graphique adéquat.

(e) Interprétez les paramètres de votre modèle. Selon ce dernier, quelle est la probabilité de mourir pour une personne âgée de 60 ans ? Construisez un intervalle de confiance, à 95%, pour cette probabilité.

(f) Utilisez la fonction `fitted()` pour calculer les résidus de la **déviance** à l'aide de la formule qui définit ces derniers dans le cours. Vérifiez vos calculs à l'aide de la fonction `residuals()`.

Représentez ces résidus en fonction des valeurs ajustées. Que pouvez-vous en conclure ? Au besoin, proposez une autre approche pour une analyse des résidus adaptée aux données.

(g) Que pensez-vous de la qualité d'ajustement du modèle ? Répondez à cette question de deux façons différentes: (i) en utilisant les calculs réalisés au point (a); (ii) en utilisant la courbe ROC.

Partie 2

(a) On aimerait ajuster un modèle logistique avec **STA** comme variable expliquée et **CPR** comme variable explicative. En précisant vos notations et en prenant $CPR = 2$ et $STA = 0$ comme niveaux de référence,

- (i) réalisez un tableau croisé entre les variables **STA** et **CPR**
- (ii) donnez l'équation théorique du modèle logistique en question
- (iii) estimez les paramètres "**à la main**", c-à-d sans utiliser la fonction `glm()`
- (iv) toujours à la main, calculez un intervalle de confiance pour la pente (β_1)
- (v) comparez vos calculs avec un ajustement fait avec la fonction `glm()`

(b) Qu'advient-il des paramètres estimés du modèle si l'on modifie le niveau de référence de la **CPR** ? Répondez **sans utiliser R**, puis vérifiez avec R.

(c) Réalisez un tableau croisé entre les variables **STA** et **RAC** (c'est la race des individus: 1 = white, 2 = black, 3 = other). Prenez $RAC = 1$ comme niveau de référence. Calculez les deux *log-OR*'s de la table de contingence; vous devez utiliser les groupes de références tel qu'indiqué ci-dessus. Ajustez ensuite un modèle logit avec **STA** comme variable expliquée et **RAC** comme variable explicative. Commentez.

(d) Utilisez le modèle ajusté au point précédent pour tester l'indépendance entre les variables **STA** et **RAC**.

(e) Ajustez un modèle logit avec **STA** comme variable expliquée et **CRN** et **AGE** comme variables explicatives. Prenez $CRN = 2$ comme référence et incluez l'interaction entre **CRN** et **AGE** dans votre modèle. Cette interaction est-elle significative ? Le cas échéant, mettez à jour le modèle en supprimant cette interaction. Utilisez le modèle, éventuellement réduit, pour calculer la probabilité de décès d'une personne âgée de 30 ans dont le $CRN = 2$.

Partie 3

En plus de la variable **AGE**, nous souhaitons ici expliquer la variable **STA** à l'aide des variables **CPR**, **CAN**, **INF**, ainsi que la variable **RAC** qu'on vous demande de *recoder* de façon à ce qu'elle soit dichotomique (1 = white, 0 = black or other); prenez "0" comme référence. Pour **CPR**, **CAN** et **INF**, prenez "2" comme niveau de référence.

(a) Écrivez à la main l'équation complète d'un modèle logistique, sans interactions, incluant les variables citées ci-dessus et estimez les paramètres de ce modèle.

(b) Utilisez les fonction `logLik()` et `pchisq()` pour réaliser un test LR pour tester le modèle actuel (H_1) versus un modèle avec seulement l'intercept (H_0). Vous ne devez utiliser ni la fonction `anova()` ni la fonction `drop1()`. Écrivez explicitement vos hypothèses H_0 et H_1 . Que concluez-vous ?

(c) Simplifiez le modèle actuel en supprimant toutes les variables/termes non-significatives à 5%. Effectuez cette simplification étape par étape en supprimant un élément à la fois. Écrivez l'équation de votre modèle ainsi construit.

(d) Maintenant que vous n'avez que des variables explicatives significatives, complétez le modèle en (c) en y ajoutant toutes les interactions. Peut-on simplifier ce dernier? Utilisez le BIC pour répondre à cette dernière question.

(e) Utilisez le modèle que vous avez choisi pour prédire la probabilité **de survie** pour deux patients avec $AGE = 25$ et $AGE = 80$ et un $CPR = "1"$. Même question pour un $CPR = "2"$. Accompagnez vos calculs d'intervalles de confiance à 95%.

(f) Selon le modèle choisi, quel est l'effet de l'âge sur la mortalité ? Répondez par un **graphique** approprié. Même question, mais cette fois concernant l'effet de la variable **CPR**.