LSTAT2100 - Exercices - Série 2

Énoncés

Exercice 1

Dans cet exercice, nous allons reprendre un exercice de la séance précédente, et le réaliser **en utilisant les modèles log-linéaires**.

On s'intéresse à la couleur des yeux d'individus dans un pays. Pour $n = 300$ individus pris au hasard, on a obtenu :

bleu	marron	noir	vert
48	122	95	35

(a) Peut-on dire que toutes les couleurs sont équiprobables ?

(b) Peut-on dire que les yeux foncés (marron et noir) sont deux fois plus probables que les yeux clairs (bleu et vert) ?

(c) Peut-on dire que les yeux bleus et verts sont équiprobables ?

Exercice 2

Dans cet exercice, nous allons utiliser le jeu de données data.csv. Ce jeu de données comprend des observations liées à des clients qui ont contracté un crédit. Nous avons à notre disposition un certain nombre de variables, dont la variable Y qui nous dit si le client a pu rembourser dans les temps le crédit ($Y = 1$) ou pas ($Y = 2$) et la variable tel nous renseigne si la personne a un numéro de téléphone (A191 : Non, A192 : Oui).

Charger les données avec la commande suivante (cette commande suppose que votre répertoire de travail contient un répertoire `data` qui contient le fichier `data.csv`).

```
mdata <- read.csv(file = "data/data.csv", stringsAsFactors = TRUE)
# cette dernière option permet d'automatiquement transformer toute
# variable de type caractère en un facteur
```

(a) Ajustez un modèle log-linéaire saturé pour modéliser la table de contingence $tel \times Y$. Construisez des IC à 95% pour les différents paramètres. Pouvez-vous reconstruire la table de contingence à partir des coefficients estimés ? Que valent les résidus pour ce modèle ?

(b) Visualisez graphiquement $\hat{p}(Y|tel)$. Que suggère votre graphique quand à l'association entre Y et tel ?

(c) Utilisez le modèle log-linéaire ajusté précédemment pour tester l'indépendance entre Y et tel. Proposez deux approches différentes, dont une ne nécessitant aucun calcul ou ajustement supplémentaire. Que peut-on conclure ?

Exercice 3

Dans cet exercice, nous allons utiliser le même jeu de données (**data.csv**) que celui utilisé dans l'Exercice 2. Nous considérons ici les variables *check_ac* et *employm*.

- *check_ac* représente le montant sur le compte à vue du client selon 4 modalités :
  - A11 : < 0
  - A12 : $[0; 200[$
  - A13 : ≥ 200
  - A14 : pas de compte.
- *employm* donne des informations sur la situation professionnelle du client :
  - A71 : sans emploi
  - A72 : < 1 an
  - A73 : $[1; 4[$
  - A74 :$[4; 7[$
  - A75 : ≥ 7 ans

(a) Créez une nouvelle vriable *check_ac1* qui regroupe les modailités "A11" et "A14" de la vraible *check_ac* en une seule modalité nommée "A11" et les modailités "A12" et "A13" en une seule modalité nommée "A12". Réalisez une table de contingence pour les vraibles *employm* et *check_ac1*. Puis construisez un modèle log-linéaire saturé sur vos données en choisissant "A11" comme référence.

(b) Certains termes d'interaction sont-ils non significatifs ? Si oui, combien ? À partir de ces termes, que pouvez-vous conclure concernant l'association entre ces deux variables (il n'est pas demandé d'effectuer un calcul supplémentaire) ? Que signifie la p-valeur qui figure dans la ligne "check_ac1A12" du "summary" de votre modèle.

(c) Utilisez le modèle ajusté pour tester l'indépendance entre ces deux variables.

(d) Indépendamment du résultat du test précédent, construisez un modèle log-linéaire en considérant les deux variables indépendantes. Que signifie la p-valeur qui figure dans la ligne "check_ac1A12" du "summary" de votre modèle. Calculez le R^2 de ce modèle et comparez le au R^2 du modèle saturé.

Exercice 4

Dans cet exercice, nous allons utiliser le même jeu de données (**data.csv**) que celui utilisé dans l'Exercice 2. Nous considérons ici les variables Y, *tel* et *employm*.

(a) Réalisez un tableau de contingence et un mosaicplot impliquant ces trois variables. Est-ce qu'une association homogène entre ces trois variables est envisageable ? Il n'est pas demandé ici de réaliser un test.

(b) Confirmer ou infirmer votre réponse (a) à l'aide d'un test adéquat.

(c) Construisez un modèle log-linéaire saturé et simplifiez-le le plus possible à l'aide des tests entre modèles emboîtés. Refaites la même analyse en utilisant l'AIC ou la BIC.

Exercice 5

Le but de cet exercice est de montrer que l'indépendance deux à deux n'implique pas l'indépendance mutuelle.

Pour cela, considérant l'expérience qui consistent à lancer deux pièces (indépendamment l'une de l'autre). On définit les variables suivantes:

- X la variable aléatoire binaire prenant la valeur 1 si *la première* pièce montre face et 0 sinon,
- Y la variable aléatoire binaire prenant la valeur 1 si *la deuxième* pièce montre face et 0 sinon.
- Z la variable aléatoire prenant la valeur 1 si les deux pièces montrent le même coté et 0 sinon; i.e., $Z = I(X = Y)$.

Les deux variables X et Y sont supposées être indépendantes.

(a) Montrer que X et Z sont indépendantes. On en déduit que X, Y et Z sont indépendantes deux à deux.

(b) Montrer que X, Y et Z ne sont pas mutuellement indépendants.

Exercice 6

(a) Montrer que

$$\text{(a) } X \perp\!\!\!\perp Y|Z \text{ et } \text{ (b) } X \perp\!\!\!\perp Z|Y \iff \text{ (c) } X \perp\!\!\!\perp (Y, Z)$$

(b) Montrer que

$$\text{(a) } X \perp\!\!\!\perp (Y, Z) \text{ et } \text{ (b) } Y \perp\!\!\!\perp (X, Z) \iff \text{ (c) } X \perp\!\!\!\perp Y \perp\!\!\!\perp Z$$