# LSTAT2100 - Exercices - Série 4

## Énoncés

Les exercices de ce TP sont en très grande partie tirés d'examens antérieurs.

### Excercice 1

Considérons $n$ variables aléatoires indépendantes $Y_1, \dots, Y_n$ telles que $Y_i$ est distribuée selon une loi binomiale négative ($NBin$) avec les paramètres $(k, \pi_i)$, où $k \geq 2$ est un entier connu et $\pi_i \in (0, 1)$ est inconnu. La loi binomiale négative $(k, \pi_i)$ modélise, dans le contexte d'une suite d'épreuves indépendantes de Bernoulli, le nombre d'essais nécessaires pour obtenir $k$ succès, $\pi_i$ représente la probabilité de succès. Nous avons

$$f(y_i; \pi_i) = C_{y_i-1}^{k-1} \pi_i^k (1 - \pi_i)^{(y_i - k)}, \;\; y_i = k, k+1, \dots .$$

#### (a)

Montrer que cette loi appartient à la famille exponentielle. En suivant les notations du cours quant aux familles exponentielles, on explicitera les paramètres $\theta$ et $\phi$, et la fonction $b$.

#### (b)

Calculez $E(Y_i)$ et $Var(Y_i)$.

En plus du fait que $Y_i \sim NBin(k, \pi_i)$, $i = 1, \dots, n$, on suppose dans ce qui suit que

$$\log(1 - \pi_i) = \beta_0 + \beta_1 x_i,$$

où $x_i$ est une variable à valeurs dans $\mathbb{R}$.

#### (c)

Montrez qu'il s'agit d'un modèle linéaire généralisé. La fonction de lien canonique a-t-elle été utilisée ? Justifiez et détaillez votre raisonnement.

### (d)

Pour le modèle défini ci-dessus, donnez les équations de vraisemblance qui déterminent les estimateurs du maximum de vraisemblance de $\beta_0$ et $\beta_1$. Développez et simplifiez vos calculs. Est-il possible d'obtenir des expressions explicites pour ces estimateurs ? Dans l'affirmative, procédez aux calculs qui s'imposent.

### (e)

Donnez les expressions de la déviance et de la déviance nulle correspondant au modèle défini ci-dessus. Simplifiez au maximum vos formules.

## Excercice 2

Le tableau suivant donne le nombre de candidats (admis ou rejetés) à une certaine prestigieuse université américaine pour six grands départements (A, B, ..., F). La question de la discrimination à l'égard des femmes dans l'accès à l'enseignement supérieur étant au cœur de cette étude, le genre est rapporté.

| Gender | Male | | Female | |
|---|---|---|---|---|
| Admit | Rejected | Admitted | Rejected | Admitted |
| Dept | | | | |
| A | 313 | 512 | 19 | 89 |
| B | 207 | 353 | 8 | 17 |
| C | 205 | 120 | 391 | 202 |
| D | 279 | 138 | 244 | 131 |
| E | 138 | 53 | 299 | 94 |
| F | 351 | 22 | 317 | 24 |

### (a)

Calculez la proportion de candidats admis à cette université (tous départements confondus) par genre. Représentez ces chiffres à l'aide d'un graphique approprié. Que constatez-vous ?

### (b)

Au vu de ces données (tous départements confondus), peut-on affirmer que les femmes font l'objet d'une discrimination significative ? Pour répondre à cette question, vous devrez réaliser *un test LR sur un rapport de cotes*. Formulez vos hypothèses $H_0$ et $H_1$, et décrivez clairement votre démarche. Proposez une autre approche, toujours basée sur un test LR de rapport de cotes, en utilisant cette fois une modélisation différente. Parvenez-vous à la même conclusion ? Commentez.

### (c)

Répétez la question (a), mais cette fois-ci par département. Peut-on parler d'une *discrimination significative*, fondée sur le genre, au sein de l'un ou l'autre des départements considérés ? Justifiez. Comparez et mettez en contraste les deux analyse (celle du point (a) et la présente).

### (d)

Ces données relèvent du paradoxe de Simpson. Expliquez brièvement sa manifestation ici. Identifiez la cause de ce paradoxe, càd un (ou des) élément(s) dans les données qui l'explique(nt) clairement. Pour répondre à cette question, vous devez fournir un ou plusieurs graphiques, accompagnés d'une argumentation.

### (e)

En prenant "Male" et "A" comme catégories de référence (pour *Gender* et *Dept*, respectivement), ajustez un modèle de régression logistique pour modéliser la probabilité d'admission en fonction du genre et du département. Peut-on conclure à une association homogène entre les trois variables considérées ? Si nous nous concentrons sur la ligne `GenderFemale:DeptB` dans `summary()`, comment pouvez-vous interpréter précisément les valeurs figurant dans les colonnes `Estimate` et `Pr(>|z|)` ?

## Excercice 3

Le fichier lung_cancer_data contient des données sur le nombre de cas de cancer du poumon (variable `cases`) dans quatre villes du Danemark (Fredericia, Horsens, Kolding, Vejle) et pour différentes catégories d'âge ($40-54$, $55-59$, $60-64$, $65-69$, $70-74$, $\geq 75$). La taille de la population de chaque groupe d'âge de chaque ville est rapportée dans la variable `city`. Et pour chaque tranche d'âge, `age_midpt` donne le point central, sauf pour la dernière tranche où 75 est utilisé.

Pour charger des données et les enregistrer dans l'objet `lc`, vous pouvez saisir la commande suivante.

```
lc <- read.table("Data/lung_cancer_data.txt")
```

### (a)

Faites un (seul) graphique montrant l'évolution du taux de cancer (*cas/population*) en fonction de l'âge. Utilisez une couleur différente pour chaque ville.

### (b)

Le modèle suivant est proposé pour expliquer le nombre de cas de cancer en fonction des variables disponibles. Ce modèle présente une déficience qui doit être corrigée. Expliquez cette déficience et proposez un remède qui prenne mieux compte les données disponibles *tout en utilisant la régression de Poisson.*

```
glm(cases ~ city * age_midpt, data = lc, family = poisson)
```

### (c)

Considérons le modèle de Poisson (corrigé), cette fois avec les variables explicatives *city*, *age_midpt*, l'interaction entre les deux et $age\_midpt^2$. Simplifiez ce modèle en ne conservant que les termes *significatifs à* 1%. En précisant vos notations, écrivez l'équation mathématique de votre modèle final. Diagnostiquez sa qualité d'ajustement et concluez.

### (d)

Refaites l'analyse précédente mais cette fois-ci en utilisant une régression logistique. Comparez les deux modèles (Poisson et logistique).