Université de Montpellier — Année 2020-2021
HMMA304 - Modèles linéaires généralisés

Examen final

Durée 2h - Sans document - 2 novembre 2020

Exercice 1 (5 pts)

Nous considérons un n-échantillons $(Y_1, x_1), \ldots, (Y_n, x_n)$ du couple (Y, x) où Y est une variable aléatoire à valeurs dans $\{0, 1\}$ et x est une variable réelle fixée.
Nous supposons qu'il existe une variable latente Y_i^* telle que $Y_i^*|x_i \sim \mathcal{L}(a + bx_i, 1)$ où $\mathcal{L}(\mu, s)$ est une loi de probabilité à valeurs réelles de densité

$$f_{\mathcal{L}(\mu,s)}(y) = \frac{\exp(-(y-\mu)/s)}{s(1+\exp(-(y-\mu)/s))^2}\,.$$

Nous supposons enfin que

$$Y_i = \begin{cases} 1 & \text{si} \quad Y_i^* \geq 0 \\ 0 & \text{si} \quad Y_i^* < 0 \end{cases}\,.$$

1 (2 pts) Calculer $\mathbb{P}(Y_i = 1|X = x_i)$.
Indication : la primitive de $\dfrac{\exp(-(y-\mu))}{(1+\exp(-(y-\mu)))^2}$ est $\dfrac{1}{1+\exp(-(y-\mu))}$.

2 (2 pts) Montrer qu'il s'agit d'un modèle linéaire généralisée. En suivant les notations du cours quant aux familles exponentielles à un paramètre de nuisance, on explicitera les paramètres ϕ, θ, les fonctions r et b ainsi que les régresseurs à considérer.

3 (1 pt) Est-ce le lien canonique qui a été choisi ?

Exercice 2 (4 pts)

Nous considérons la loi $\Gamma(6, 1)$ tronquée en $a > 0$, de densité

$$f(x) \propto x^5 \exp(-x)\mathbb{I}_{[a,\infty[}(x)\,.$$

1 (2 pts) Proposer une méthode pour générer des réalisations suivant cette loi et donner le code R associé.

2 (2 pts) Proposer une méthode pour approcher

$$\int_a^\infty x^5 \exp(-x)\mathrm{d}x\,.$$

Exercice 3 (6 pts)

Nous souhaitons estimer
$$\theta = \int_0^{2\pi} \exp(-x/10)\sin(x)\mathrm{d}x\,.$$

1 (2 pts) Proposer deux méthodes de Monte Carlo pour estimer θ.

2 (2 pts) Comparer ces deux stratégies.

3 (2 pts) Déterminer le nombre de simulation nécessaire pour que l'erreur absolue relative de l'une des deux méthodes soit inférieure à 1% avec une probabilité supérieure ou égale à 99%.

Exercice 4 (5 pts)

Décrire et commenter les résultats obtenus à l'aide du code R ci-dessous.

```
> library(gRim)
> data(reinis)
> summary(reinis)
Number of cases in table: 1841
Number of factors: 6
Test for independence of all factors:
Chisq = 809.5, df = 57, p-value = 1.564e-133
Chi-squared approximation may be incorrect
> model1 <- dmod(~ .^., reinis)
> model2 <- backward(model1,k=log(1841))
  change.AIC -108.0635 Edge deleted: mental,systol
  change.AIC  -52.9956 Edge deleted: phys,systol
  change.AIC  -48.7808 Edge deleted: mental,protein
  change.AIC  -23.7047 Edge deleted: systol,family
  change.AIC  -25.4955 Edge deleted: family,protein
  change.AIC  -23.0541 Edge deleted: phys,family
  change.AIC  -12.3780 Edge deleted: smoke,family
  change.AIC   -9.0485 Edge deleted: smoke,mental
  change.AIC   -2.7865 Edge deleted: mental,family
```

Description du jeu de données
reinis gRbase R Documentation Risk factors for coronary heart disease
Data collected at the beginning of a 15 year follow-up study of probable risk factors for coronary thrombosis. Data are from all men employed in a car factory.
A table with 6 discrete variables

- A smoking,
- B strenous mental work,
- D strenuous physical work,
- E systolic blood pressure,
- F ratio of lipoproteins,
- G Family anamnesis of coronary heart disease.

Correction examen final HMMA304
2 novembre 2020

Exercice 1

1] $\mathbb{P}(Y_i = 1 \mid X = x_i)$

$$= \int_0^{+\infty} \frac{e^{-(y-a-bx_i)}}{\left(1 + e^{-(y-a-bx_i)}\right)^2}\, dy$$

$$= \left[\frac{1}{1 + \exp\left[-(y-a-bx_i)\right]} \right]_0^{+\infty}$$

$$= 1 - \frac{1}{1+e^{(a+bx_i)}} = \boxed{\frac{e^{a+bx_i}}{1+e^{a+bx_i}}}$$

On reconnait le modèle ②
de régression logistique
Questions 2] & 3] triviales.

Exercice 2

1] On utilise l'algorithme acceptation-rejet ; on simule une réalisation suivant une loi gamma (6,1) ; on l'accepte si elle est plus grande que a ; si on rejette on recommence jusqu'à acceptation.

Code R pour générer une réalisation:

```
bool <- TRUE
while (bool) {
y <- rgamma(1,6,1)
if(y >= a){ bool <- FALSE; x <- y}}
```

2] $\int_a^{\infty} x^5 e^{-x} dx$ ③

$= \mathbb{E}_{\mathcal{E}xp(1)}\left(X^5 \mathbb{1}_{\{X \geq a\}}\right)$

On peut mettre en œuvre la méthode de Monte Carlo associée.

On simule $X_1, \ldots, X_N \overset{iid}{\sim} \mathcal{E}xp(1)$

On approche $\int_a^{\infty} x^5 e^{-x} dx$

par

$$\frac{1}{N} \sum_{i=1}^{N} X_i^5 \mathbb{1}_{\{X_i \geq a\}}$$

Exercice 3

1] $\widehat{\theta_1} = \left(\frac{1}{N}\right)(2\pi) \sum_{i=1}^{N} e^{-\frac{X_i}{10}} \sin(X_i)$

où $X_1, \ldots, X_N \overset{iid}{\sim} \mathcal{U}_{[0,2\pi]}$

$\widehat{\theta_2} = \left(\frac{1}{N}\right) 10 \sum_{i=1}^{N} \sin(X_i) \mathbb{1}_{[0,2\pi]}(X_i)$

où $X_1, \ldots, X_N \overset{iid}{\sim} \mathcal{E}xp\left(\frac{1}{10}\right)$

2] $\mathbb{E}(\widehat{\theta_1}) = \theta$

$\mathbb{V}(\widehat{\theta_1}) = \frac{1}{N}\left[(2\pi)^2 \mathbb{E}_{\mathcal{U}_{[0,2\pi]}}\left(e^{-\frac{2X}{10}} \sin^2(X)\right) - \theta^2\right]$

Par ailleurs, $\mathbb{E}_{\mathcal{U}_{[0,2\pi]}}\left(e^{-\frac{2X}{10}} \sin^2(X)\right)$

$= \int_0^{2\pi} \frac{e^{-\frac{2x}{10}} \sin^2(x)}{2\pi} dx$

(5)

$$\mathbb{E}(\widehat{\theta_2}) = \theta$$

$$\mathbb{V}(\widehat{\theta_2}) = \frac{1}{N}\left[10^2\,\mathbb{E}_{\mathcal{E}xp\left(\frac{1}{10}\right)}\left(\sin^2(X)\,\mathbb{1}_{[0,2\pi]}(X)\right) - \theta^2\right]$$

Par ailleurs,

$$\mathbb{E}_{\mathcal{E}xp\left(\frac{1}{10}\right)}\left(\sin^2(X)\,\mathbb{1}_{[0,2\pi]}(X)\right)$$

$$= \int_0^{2\pi} \frac{1}{10} e^{-\frac{x}{10}} \sin^2(x)\,dx$$

Il s'agit donc de comparer

$$10\int_0^{2\pi} e^{-\frac{x}{10}} \sin^2(x)\,dx \quad \text{et}$$

$$(2\pi)\int_0^{2\pi} e^{-\frac{x}{5}} \sin^2(x)\,dx$$

$$\frac{x}{10} < \frac{x}{5}$$

$$\Rightarrow \frac{-x}{10} > -\frac{x}{5}$$

$$\Rightarrow e^{-\frac{x}{10}} \sin^2(x) > e^{-\frac{x}{5}} \sin^2(x)$$

$$\Rightarrow \int_0^{2\pi} e^{-\frac{x}{10}} \sin^2(x)\,dx > \int_0^{2\pi} e^{-\frac{x}{5}} \sin^2(x)\,dx$$

$$\Rightarrow 10 \int_0^{2\pi} e^{-\frac{x}{10}} \sin^2(x)\,dx > (2\pi) \int_0^{2\pi} e^{-\frac{x}{5}} \sin^2(x)\,dx$$

$$\Rightarrow \mathbb{V}(\hat{\theta}_1) \leq \mathbb{V}(\hat{\theta}_2)$$

3) Nous allons utiliser le meilleur estimateur à savoir $\hat{\theta}_1$.

Erreur absolue relative de $\hat{\theta}_1$ est

$$|\hat{\theta}_1 - \theta| / |\theta|$$

Nous $^{(*)}$ la plus petite valeur de N telle que $\boxed{(*)\text{ cherchons}}$

$$\mathbb{P}\left(\frac{|\hat{\theta}_1-\theta|}{|\theta|}\leq 10^{-2}\right)\geq 1-10^{-2}$$

$\theta_1 > 0$ trivialement.

Par ailleurs, d'après l'inégalité de Bienaymé - Tchebyshev nous avons

$$\mathbb{P}\left(\frac{|\hat{\theta}_1-\theta|}{|\theta|}\leq 10^{-2}\right)\geq 1-\frac{\mathbb{V}(\hat{\theta}_1)}{\theta^2 10^{-4}}$$

$$1-\frac{\mathbb{V}(\hat{\theta}_1)}{\theta^2 10^{-4}}\geq \left[1-10^{-2}\right]$$

$$\Leftrightarrow \frac{\mathbb{V}(\hat{\theta}_1)}{\theta^2}\leq 10^{-6}$$

$$\Leftrightarrow N\geq 10^{6}\left(\frac{2\pi}{\theta^2}\int_0^{2\pi} e^{-\frac{x}{5}}\sin^2(x)\,dx-1\right)$$

8

$N \geq \alpha\, 10^6$

Si $\alpha \leq a$, alors

$10^6 a \geq 10^6 \alpha$

Et si $N \geq 10^6 a$ alors

$N \geq 10^6 \alpha$

$\alpha = \left[\dfrac{2\pi}{\theta^2}\displaystyle\int_0^{2\pi} e^{-\frac{x}{5}} \sin^2(x)\, dx - 1\right]$ voir (*)

$\alpha \leq \left[8\pi e^{\frac{\pi}{5}} \displaystyle\int_0^{2\pi} e^{-\frac{x}{5}}\, dx - 1\right]$

$\alpha \leq 8\pi e^{\frac{\pi}{5}} \left[-5 e^{-\frac{x}{5}}\right]_0^{2\pi} - 1$

$\alpha \leq 8\pi e^{\frac{\pi}{5}} \left(5 - 5 e^{-\frac{2\pi}{5}}\right) - 1$

$\# \; 167{,}5$

$N \geq 167{,}5 \times 10^6$

(*) $\theta = \displaystyle\int_0^{2\pi} e^{-\frac{x}{10}} \sin(x)\, dx$ [voir ou dos →]

8'

$$\Theta = \int_0^{2\pi} e^{-\frac{x}{10}} \sin(x)\, dx$$

$$= \int_0^{\pi} e^{-\frac{x}{10}} \sin(x)\, dx + \int_{\pi}^{2\pi} e^{-\frac{x}{10}} \sin(x)\, dx$$

$$= \int_0^{\pi} e^{-\frac{x}{10}} \sin(x)\, dx + \int_0^{\pi} e^{-\frac{x+\pi}{10}} \sin(x+\pi)\, dx$$

$$= \int_0^{\pi} \sin(x) \left[e^{-\frac{x}{10}} - e^{-\pi/10} e^{-x/10} \right] dx$$

$$= \int_0^{\pi} \sin(x) e^{-\frac{x}{10}} \left(1 - e^{-\pi/10}\right) dx$$

$$\geq \int_0^{\pi} \sin(x) e^{-x/10} \left(\frac{1}{4}\right) dx$$

$$\geq \frac{e^{-\pi/10}}{4} \int_0^{\pi} \sin(x)\, dx$$

$$\geq \frac{e^{-\pi/10}}{2} \Rightarrow \frac{1}{\Theta^2} \leq 4 e^{\pi/5}$$

Exercice 4

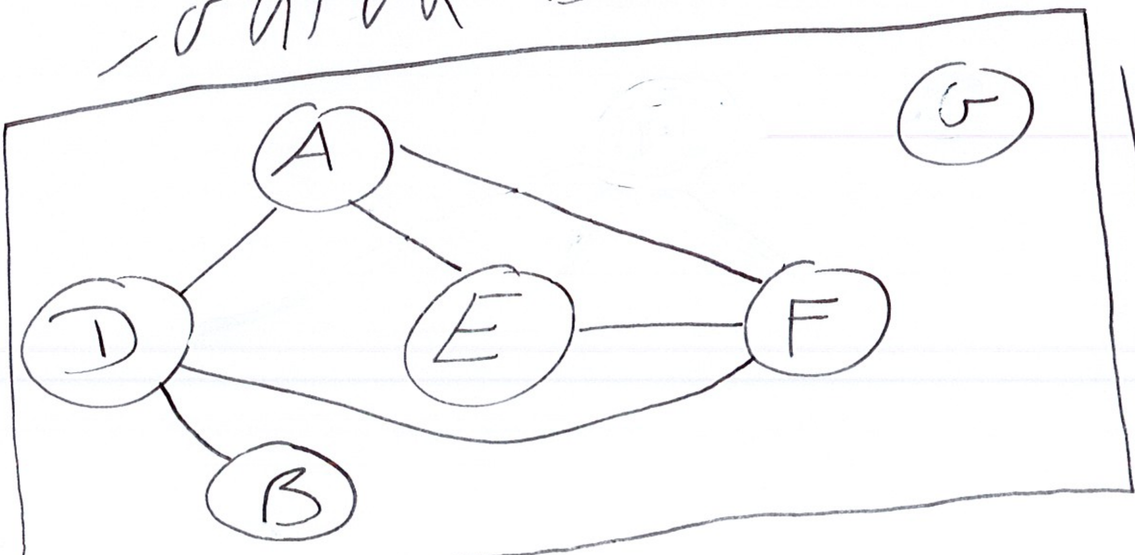
Mise en oeuvre d'un modèle log-linéaire graphique

6 variables qualitatives

1 841 individus

model 1 est le modèle saturé

model 2 est le meilleur modèle par une méthode descendante basée sur le critère BIC



$G \perp\!\!\!\perp A, B, D, E, F$

$B \perp\!\!\!\perp A, E, F \mid D$