Méthodes d’analyse biostatistique projet 1

Wen, Zehai

2023-10-13 14:47:46-04:00

0.1 Exercice 1a

Démontrez que la loi de Poisson appartient à la famille exponentielle sous la mesure de comptage.

Solution Soit ν la mesure de comptage supportée sur $\mathbb{N}$. La loi de Poisson avec paramètre $\lambda > 0$ est définie par $f(y)d\nu(y)$, où la densité $f : \mathbb{N} \to (0, \infty)$ est donnée par $f(y) = \frac{\lambda^y}{y!}e^{-\lambda}$ pour tout $y \in \mathbb{N}$. Écrire $\lambda = e^\theta$ pour un $\theta \in \mathbb{R}$. On a alors:

$$f(y) = \exp\{-\lambda + y \ln \lambda - \ln y!\} = \exp\{\theta y - e^\theta - \ln y!\}$$

Par conséquent, la loi de Poisson appartient à la famille exponentielle avec l'espace paramétrique naturelle $\mathbb{R}$. Sous la forme $f(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$, on a que $a(\phi) \equiv 1$, $b(\theta) = e^\theta$ et $c(y, \phi) = -\ln y!$.////

0.2 Exercice 1b

Afin de faire la régression logistique, soit Y, X_1, X_2 et X_3 quatre variables aléatoires à valeur $\mathbb{R}$ définies sur l'espace de probabilité commune $(\Omega, \mathcal{F}, \mathbb{P})$ telles que $Y \in \{0, 1\}$, $X_1 \in \{0, 1\}$ et X_2 sont variables binaires catégoriels et X_3 est continue. Supposons qu'il y a quatre constantes réelles β_0, β_1, β_2 et β_3 telles que :

$$P\{Y = 1 \mid X_1 = x_1, X_2 = x_2, X_3 = x_3\} = \pi_1(x_1, x_2, x_3) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)\}}$$

Assumer premièrement que X_2 et X_3 sont constantes presque partout, montrez que β_1 peut être interprété comme un log-rapport de cotes.

Dans le deuxième cas, trouvez la différence du log-rapport de cotes pour deux individus. Le premier individu a $X_1 = 1$ et $X_3 = 7$, le deuxième individu a $X_1 = 0$ et $X_3 = 5$, et les deux individus ont la même valeur de X_2.

Solution Si $f : \mathbb{R} \to (1, \infty)$ et $f(x) = \frac{1}{1+e^{-x}}$, alors f est une bijection parce que la fonction exponentielle est une bijection. La fonction inverse de f est $f^{-1} : (1, \infty) \to \mathbb{R}$ et $f^{-1}(y) = \ln \frac{y}{1-y}$.

Supposons que $X_2 = x_2$ et $X_3 = x_3$ presque partout. On a :

$$\beta_0 + \beta_1 X_1 + \beta_2 x_2 + \beta_3 x_3 = \ln \frac{\pi_1(X_1, x_2, x_3)}{1 - \pi_1(X_1, x_2, x_3)}$$

Si $X_1 = 1$, on a:

$$\beta_1 = \ln \frac{\pi_1(1, x_2, x_3)}{1 - \pi_1(1, x_2, x_3)} - \beta_2 x_2 - \beta_3 x_3 - \beta_0$$

Donc, β_1 peut être interprété comme un log-rapport de cotes. Dans le deuxième cas, supposons que $X_2 = x_2$ pour tous les deux individus, alors on a :

$$\ln \frac{\frac{\pi_1(1,x_2,7)}{1-\pi_1(1,x_2,7)}}{\frac{\pi_1(0,x_2,5)}{1-\pi_1(0,x_2,5)}} = \ln \frac{\pi_1(1, x_2, 7)}{1 - \pi_1(1, x_2, 7)} - \ln \frac{\pi_1(0, x_2, 5)}{1 - \pi_1(0, x_2, 5)} = \beta_1 + 2\beta_3$$

Le calcul est complet.////

0.3 Exercice 1c

Soit X, Y et Z trois variables aléatoires à valeur $\mathbb{R}$ définies sur l'espace de probabilité commune $(\Omega, \mathcal{F}, \mathbb{P})$ telles que :

1. La probabilité conditionnelle régulière $\mu_{Y|X,Z}(\omega, \cdot)$ est la loi de Poisson avec paramètre $\exp(\beta_0 + \beta_1 X(\omega) + \beta_2 Z(\omega$ pour tout $\omega \in \Omega$, où β_0, β_1 et β_2 sont des constantes réelles.
2. On a :

$$\mathbb{E}[Y \mid X] = \exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X]$$

Calculer Var$(Y \mid X)$.

Solution On a, pour chaque $\omega \in \Omega$:

$$\mathbb{E}[Y^2 \mid X, Z](\omega) = \int_{\mathbb{R}} y^2 d\mu_{Y|X=X(\omega),Z=Z(\omega)}(y) = \exp\{2(\beta_0 + \beta_1 X(\omega) + \beta_2 Z(\omega))\} + \exp\{\beta_0 + \beta_1 X(\omega) + \beta_2 Z(\omega)\}$$

Alors :

$$\mathbb{E}[Y^2 \mid X] = \mathbb{E}[\mathbb{E}[Y^2 \mid X, Z] \mid X] = \exp\{2(\beta_0 + \beta_1 X)\}\mathbb{E}[e^{2\beta_2 Z} \mid X] + \exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X]$$

Donc, en utilisant la formule de la variance conditionnelle :

$$\begin{aligned} &\text{Var}(Y \mid X) = \mathbb{E}[Y^2 \mid X] - \mathbb{E}[Y \mid X]^2 \\ &= \exp\{2(\beta_0 + \beta_1 X)\}\mathbb{E}[e^{2\beta_2 Z} \mid X] + \exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X] - \big(\exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X]\big)^2 \\ &= \exp\{2(\beta_0 + \beta_1 X)\}\text{Var}(e^{\beta_2 Z} \mid X) + \exp\{\beta_0 + \beta_1 X\}\mathbb{E}[e^{\beta_2 Z} \mid X] \end{aligned}$$

Le calcul est complet. ////

0.4 Exercice 1d

Fix $\alpha > 0$. For a $\theta \in (0, \alpha)$, let $\{X_i\}_{i=1}^n$ be a sequence of real independent identically distributed random variable defined on some probability space (Ω, F, P) with common probability density function with respect to the Lebesgue measure:

$$f_\theta(x) = \begin{cases} \frac{2x}{\alpha\theta} & x \in [0, \theta] \\ \frac{2(\alpha - x)}{\alpha(\alpha - \theta)} & x \in [\theta, \alpha] \\ 0 & \text{otherwise} \end{cases}$$

Prove that the maximum likelihood estimation of θ must be one of the given observation but not necessarily any particular observation. In case $\alpha = 5$ and $n = 3$, compute the maximum likelihood estimate of θ when the observations are $(1, 2, 4)$ or $(2, 3, 4)$.

Solution Let $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ be given. Write $g_\theta(x)$ for the joint probability density function of X_i's and $x_{(i)}$ for the ith smallest coordiante in x. If $x_i < 0$ or $x_i > \alpha$ for some i, then $g_\theta(x) = 0$ for any θ so that any estimate would be a maximum likelihood estimate. We exclude this pathology and prove that:

Théorème 1. *If* $0 \leqslant x_{(1)} \leqslant \cdots \leqslant x_{(n)} \leqslant \alpha$, *then* $\theta_0 = x_{(i)}$ *for some* i.

Proof. It never happens that $\theta_0 < x_{(1)}$ because $g_{\theta_1}(x) > g_{\theta_0}(x)$ whenever $\theta_1 \in (\theta_0, x_{(1)})$. Similarly, it never happens that $\theta_0 > x_{(n)}$ because $g_{\theta_1}(x) > g_{\theta_0}(x)$ whenever $\theta_1 \in (x_{(n)}, \theta_0)$. We assume from now on that $\theta_0 \in [x_{(i)}, x_{(i+1)}]$ for some $i \in \{1, \cdots, n-1\}$. Suppose, for the sake of contradiction, that $x_{(i)} < \theta_0 < x_{(i+1)}$. We have:

$$g_{\theta_{\theta_0}}(x) = \left(\frac{2}{\alpha}\right)^n \frac{x_{(1)}}{\theta_0} \cdots \frac{x_{(i)}}{\theta_0} \frac{\alpha - x_{(i+1)}}{\alpha - \theta_0} \cdots \frac{\alpha - x_{(n)}}{\theta_0}$$

The numerator does not depend on θ_0. This motivates us to define function $h : [x_{(i)}, x_{(i+1)}] \to \mathbb{R}$ by:

$$h(\theta) = \frac{1}{\theta^i} \frac{1}{(\alpha - \theta)^{n-i}}$$

Then the second derivative is:

$$h''(\theta) = i(i+1)\theta^{-i-2}(\alpha - \theta)^{i-n} + (n-i)(n-i+1)\theta^{-i}(\alpha - \theta)^{i-n-2} > 0$$

Therefore, h is strictly convex and the maximum can only be at the boundary points. □

We now demonstrate that the choice of i is not unique in the above theorem. The simplest case will be $x_i = x_j$ for any i and any j. For a nontrivial example, let $\alpha = 5$, $n = 3$. If $x = (2, 3, 4)$, then the maximum likelihood estimate is one of $\{2, 3, 4\}$. An estimate of 3 or 4 yields maximum likelihood $\frac{8}{375}$ while an estimate of 2 yields likelihood $\frac{16}{1125}$. Therefore, the maximum likelihood estimate can be 3 or 4 and is not unique.

Finally, the additional example $x = (1, 2, 4)$, estimate $\theta = 1, 2, 4$ gives likelihood $\frac{3}{250}, \frac{4}{375}, \frac{1}{125}$ repsectively. We conclude that $\theta = 1$ is the maximum likelihood estimate in this case.////

0.5 Exercice 1e

Fix $\alpha > 0$. For a $\theta \in (0, \alpha)$, let $\{X_i\}_{i=1}^n$ be a sequence of real independent identically distributed random variable defined on some probability space (Ω, F, P) with common probability density function with respect to the Lebesgue measure:

$$f_\theta(x) = \begin{cases} \frac{2x}{\alpha\theta} & x \in [0, \theta] \\ \frac{2(\alpha - x)}{\alpha(\alpha - \theta)} & x \in [\theta, \alpha] \\ 0 & \text{otherwise} \end{cases}$$

Prove that the maximum likelihood estimation of θ must be one of the given observation but not necessarily any particular observation. In case $\alpha = 5$ and $n = 3$, compute the maximum likelihood estimate of θ when the observations are $(1, 2, 4)$ or $(2, 3, 4)$.

Solution Let $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ be given. Write $g_\theta(x)$ for the joint probability density function of X_i's and $x_{(i)}$ for the ith smallest coordiante in x. If $x_i < 0$ or $x_i > \alpha$ for some i, then $g_\theta(x) = 0$ for any θ so that any estimate would be a maximum likelihood estimate. We exclude this pathology and prove that:

Théorème 2. *If* $0 \leqslant x_{(1)} \leqslant \cdots \leqslant x_{(n)} \leqslant \alpha$, *then* $\theta_0 = x_{(i)}$ *for some* i.

Proof. It never happens that $\theta_0 < x_{(1)}$ because $g_{\theta_1}(x) > g_{\theta_0}(x)$ whenever $\theta_1 \in (\theta_0, x_{(1)})$. Similarly, it never happens that $\theta_0 > x_{(n)}$ because $g_{\theta_1}(x) > g_{\theta_0}(x)$ whenever $\theta_1 \in (x_{(n)}, \theta_0)$. We assume from now on that $\theta_0 \in [x_{(i)}, x_{(i+1)}]$ for some $i \in \{1, \cdots, n-1\}$. Suppose, for the sake of contradiction, that $x_{(i)} < \theta_0 < x_{(i+1)}$. We have:

$$g_{\theta_{\theta_0}}(x) = \left(\frac{2}{\alpha}\right)^n \frac{x_{(1)}}{\theta_0} \cdots \frac{x_{(i)}}{\theta_0} \frac{\alpha - x_{(i+1)}}{\alpha - \theta_0} \cdots \frac{\alpha - x_{(n)}}{\theta_0}$$

The numerator does not depend on θ_0. This motivates us to define function $h : [x_{(i)}, x_{(i+1)}] \to \mathbb{R}$ by:

$$h(\theta) = \frac{1}{\theta^i} \frac{1}{(\alpha - \theta)^{n-i}}$$

Then the second derivative is:

$$h''(\theta) = i(i+1)\theta^{-i-2}(\alpha - \theta)^{i-n} + (n-i)(n-i+1)\theta^{-i}(\alpha - \theta)^{i-n-2} > 0$$

Therefore, h is strictly convex and the maximum can only be at the boundary points. □

We now demonstrate that the choice of i is not unique in the above theorem. The simplest case will be $x_i = x_j$ for any i and any j. For a nontrivial example, let $\alpha = 5$, $n = 3$. If $x = (2, 3, 4)$, then the maximum likelihood estimate is one of $\{2, 3, 4\}$. An estimate of 3 or 4 yields maximum likelihood $\frac{8}{375}$ while an estimate of 2 yields likelihood $\frac{16}{1125}$. Therefore, the maximum likelihood estimate can be 3 or 4 and is not unique.

Finally, the additional example $x = (1, 2, 4)$, estimate $\theta = 1, 2, 4$ gives likelihood $\frac{3}{250}, \frac{4}{375}, \frac{1}{125}$ repsectively. We conclude that $\theta = 1$ is the maximum likelihood estimate in this case.////

0.6 Exercice 2abcd

Fix $\alpha > 0$. For a $\theta \in (0, \alpha)$, let $\{X_i\}_{i=1}^n$ be a sequence of real independent identically distributed random variable defined on some probability space (Ω, F, P) with common probability density function with respect to the Lebesgue measure:

$$f_\theta(x) = \begin{cases} \frac{2x}{\alpha\theta} & x \in [0, \theta] \\ \frac{2(\alpha - x)}{\alpha(\alpha - \theta)} & x \in [\theta, \alpha] \\ 0 & \text{otherwise} \end{cases}$$

Prove that the maximum likelihood estimation of θ must be one of the given observation but not necessarily any particular observation. In case $\alpha = 5$ and $n = 3$, compute the maximum likelihood estimate of θ when the observations are $(1, 2, 4)$ or $(2, 3, 4)$.

Solution Let $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ be given. Write $g_\theta(x)$ for the joint probability density function of X_i's and $x_{(i)}$ for the ith smallest coordiante in x. If $x_i < 0$ or $x_i > \alpha$ for some i, then $g_\theta(x) = 0$ for any θ so that any estimate would be a maximum likelihood estimate. We exclude this pathology and prove that:

Théorème 3. *If* $0 \leqslant x_{(1)} \leqslant \cdots \leqslant x_{(n)} \leqslant \alpha$, *then* $\theta_0 = x_{(i)}$ *for some* i.

Proof. It never happens that $\theta_0 < x_{(1)}$ because $g_{\theta_1}(x) > g_{\theta_0}(x)$ whenever $\theta_1 \in (\theta_0, x_{(1)})$. Similarly, it never happens that $\theta_0 > x_{(n)}$ because $g_{\theta_1}(x) > g_{\theta_0}(x)$ whenever $\theta_1 \in (x_{(n)}, \theta_0)$. We assume from now on that $\theta_0 \in [x_{(i)}, x_{(i+1)}]$ for some $i \in \{1, \cdots, n-1\}$. Suppose, for the sake of contradiction, that $x_{(i)} < \theta_0 < x_{(i+1)}$. We have:

$$g_{\theta_{\theta_0}}(x) = \left(\frac{2}{\alpha}\right)^n \frac{x_{(1)}}{\theta_0} \cdots \frac{x_{(i)}}{\theta_0} \frac{\alpha - x_{(i+1)}}{\alpha - \theta_0} \cdots \frac{\alpha - x_{(n)}}{\theta_0}$$

The numerator does not depend on θ_0. This motivates us to define function $h : [x_{(i)}, x_{(i+1)}] \to \mathbb{R}$ by:

$$h(\theta) = \frac{1}{\theta^i} \frac{1}{(\alpha - \theta)^{n-i}}$$

Then the second derivative is:

$$h''(\theta) = i(i+1)\theta^{-i-2}(\alpha - \theta)^{i-n} + (n-i)(n-i+1)\theta^{-i}(\alpha - \theta)^{i-n-2} > 0$$

Therefore, h is strictly convex and the maximum can only be at the boundary points. $\square$

We now demonstrate that the choice of i is not unique in the above theorem. The simplest case will be $x_i = x_j$ for any i and any j. For a nontrivial example, let $\alpha = 5$, $n = 3$. If $x = (2, 3, 4)$, then the maximum likelihood estimate is one of $\{2, 3, 4\}$. An estimate of 3 or 4 yields maximum likelihood $\frac{8}{375}$ while an estimate of 2 yields likelihood $\frac{16}{1125}$. Therefore, the maximum likelihood estimate can be 3 or 4 and is not unique.

Finally, the additional example $x = (1, 2, 4)$, estimate $\theta = 1, 2, 4$ gives likelihood $\frac{3}{250}, \frac{4}{375}, \frac{1}{125}$ repsectively. We conclude that $\theta = 1$ is the maximum likelihood estimate in this case.////

0.7 Exercice 2e

Soit $\{X^{(i)} = (X_1^{(i)}, X_2^{(i)})\}_{i=1}^n$ et $\{Y_i\}_{i=1}^n$ deux suites de variables aléatoires indépendantes et identiquement distribuées à valeur $\mathbb{R}^2$ et $\mathbb{R}$ respectivement définies sur un espace de probabilité commun $(\Omega, \mathcal{F}, \mathbb{P})$ tels qu'il existe trois constantes β_0, β_1 et β_2 réels avec, pour chaque $i \in \{1, \cdots, n\}$:

$$\mathbb{E}\left[Y_i \mid X^{(i)}\right] = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)}$$

Avec les hypothèses appropriées, postulez une équation d'estimation pour β_0, β_1 et β_2. Calculez une expression pour β_0, β_1 et β_2 et ses variances étant donné les X. Réalisez tout utilisant le donnée `addhealth` avec $Y_i = \texttt{Weight}_i$, $X_1^{(i)} = \texttt{age}_i$ et $X_2^{(i)} = \texttt{SES}_i$.

Solution On fait les hypothèses suivantes:

1. La règle de décision est moindre carrés.
2. Seulement pour le calcul des variances, on suppose que Y_1 a distribution $N(0, \sigma^2)$.

Pour simplicité, écrire :

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}; \quad X = \begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ \vdots & \vdots & \vdots \\ 1 & X_1^{(n)} & X_2^{(n)} \end{bmatrix}; \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}; \quad S(X, Y, \beta) = (Y - X\beta)^T (Y - X\beta)$$

L'espace de paramètre est $\mathbb{R}^3$. Selon la définition d'éaquation d'estimation, il s'agit de trouver, pour chaque $i \in \{1, \cdots, n\}$, une fonction $\psi_i : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ telle que, pour tout $\beta \in \mathbb{R}^3$, on a :

$$\mathbb{E}\left[\sum_{i=1}^n \psi_i((Y_i, X_1^{(i)}, X_2^{(i)}), \beta)\right] = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

On écrit $\hat{\beta}$ pour la solution de l'équation d'estimation, qui est une variable aléatoire dépendant Y et X, telle que $\sum_{i=1}^n \psi_i((Y_i, X_1^{(i)}, X_2^{(i)}), \hat{\beta}) = 0$ presque partout. Selon la première hypothèse, la règle de décision est :

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^3}{\arg\min}\, S(X, Y, \beta)$$

Donc on utilise :

$$\frac{\partial}{\partial \beta} S(X,Y,\beta) = -2X^T(Y - X\beta) = \sum_{i=1}^{n} \begin{bmatrix} (-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)}) \\ (-2Y_i X_1^{(i)} + 2\beta_0 X_1^{(i)} + 2\beta_1 (X_1^{(i)})^2 + 2\beta_2 X_1^{(i)} X_2^{(i)}) \\ (-2Y_i X_2^{(i)} + 2\beta_0 X_2^{(i)} + 2\beta_1 X_1^{(i)} X_2^{(i)} + 2\beta_2 (X_2^{(i)})^2) \end{bmatrix}$$
$$= \sum_{i=1}^{n} \psi_i((Y_i, X_1^{(i)}, X_2^{(i)}), \beta) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Comme :

$$\mathbb{E}[\psi_i((Y_i, X_1^{(i)}, X_2^{(i)}), \beta)] = \begin{bmatrix} \mathbb{E}\left[(-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)})\right] \\ \mathbb{E}\left[(-2Y_i X_1^{(i)} + 2\beta_0 X_1^{(i)} + 2\beta_1 (X_1^{(i)})^2 + 2\beta_2 X_1^{(i)} X_2^{(i)})\right] \\ \mathbb{E}\left[(-2Y_i X_2^{(i)} + 2\beta_0 X_2^{(i)} + 2\beta_1 X_1^{(i)} X_2^{(i)} + 2\beta_2 (X_2^{(i)})^2)\right] \end{bmatrix}$$
$$= \begin{bmatrix} \mathbb{E}\left[\mathbb{E}\left[(-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)}) \mid X^{(i)}\right]\right] \\ \mathbb{E}\left[X_1^{(i)} \mathbb{E}\left[-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)} \mid X^{(i)}\right]\right] \\ \mathbb{E}\left[X_2^{(i)} \mathbb{E}\left[-2Y_i + 2\beta_0 + 2\beta_1 X_1^{(i)} + 2\beta_2 X_2^{(i)} \mid X^{(i)}\right]\right] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

On conclut que l'équation définie est une équation d'estimation. On a donc :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

L'inverse peut être un inverse généralisé si $X^T X$ n'est pas inversible. Selon la deuxième hypothèse, on sait immédiatement que $\text{Var}(\hat{\beta} \mid X) = \sigma^2 (X^T X)^{-1}$. Maintenant, il s'agit de réaliser tout utilisant le donnée `addhealth` avec $Y_i = \texttt{Weight}_i$, $X_1^{(i)} = \texttt{age}_i$ et $X_2^{(i)} = \texttt{SES}_i$. Ce travail ne consiste que l'utilisation de la fonction `lm` en R :

```
set.seed(1234)

donnee <-
   read.delim("Chapters\\biostat_projet_1\\resource_content_1_addhealth.txt",
   header = TRUE)
#donnee <- read.delim("resource_content_1_addhealth.txt", header = TRUE)

attach(donnee)

#On commence par le netoyage
donnee$feeling_depressed <- as.factor(donnee$feeling_depressed)
donnee$feeling_depressed[is.na(donnee$feeling_depressed)] <- as.factor(
   floor(runif(sum(is.na(donnee$feeling_depressed)), min = 1, max = 4.9999)))

donnee$smoking <- as.factor(donnee$smoking)

donnee$weight <- as.numeric(donnee$weight)
```

```
donnee$weight[is.na(donnee$weight)] <- mean(donnee$weight, na.rm = TRUE)

donnee$time <- as.factor(donnee$time) #identiquement 1

donnee$age <- as.numeric(donnee$age)
donnee$age[is.na(donnee$age)] <- as.factor(mean(donnee$age, na.rm = TRUE))

donnee$sex <- as.factor(donnee$sex)
donnee$sex[is.na(donnee$sex)] <- as.factor(floor(runif(sum(is.na(donnee$sex)),
   min = 1, max = 2.9999)))

donnee$SES <- as.numeric(donnee$SES)
donnee$SES[is.na(donnee$SES)] <- as.numeric(floor(mean(donnee$SES, na.rm =
   TRUE)))

attach(donnee)

modele_moindre_carre <- lm(weight ~ SES + age, data = donnee)
print(summary(modele_moindre_carre))
```

Le code donne que, selon les donné, on a $\beta_0 = 59.5309$, $\beta_1 = 5.5819$ et $\beta_2 = -0.3637$. Les variances sont respectivement $3.7017^2 = 13.70258289$, $0.2223^2 = 0.04941729$ et $0.2149^2 = 0.4618201$. Le calcul et la démonstration sont complets. ////

0.8 Exercice 3

Fix $\alpha > 0$. For a $\theta \in (0, \alpha)$, let $\{X_i\}_{i=1}^n$ be a sequence of real independent identically distributed random variable defined on some probability space (Ω, F, P) with common probability density function with respect to the Lebesgue measure:

$$f_\theta(x) = \begin{cases} \frac{2x}{\alpha\theta} & x \in [0, \theta] \\ \frac{2(\alpha - x)}{\alpha(\alpha - \theta)} & x \in [\theta, \alpha] \\ 0 & \text{otherwise} \end{cases}$$

Prove that the maximum likelihood estimation of θ must be one of the given observation but not necessarily any particular observation. In case $\alpha = 5$ and $n = 3$, compute the maximum likelihood estimate of θ when the observations are $(1, 2, 4)$ or $(2, 3, 4)$.

Solution Let $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ be given. Write $g_\theta(x)$ for the joint probability density function of X_i's and $x_{(i)}$ for the ith smallest coordiante in x. If $x_i < 0$ or $x_i > \alpha$ for some i, then $g_\theta(x) = 0$ for any θ so that any estimate would be a maximum likelihood estimate. We exclude this pathology and prove that:

Théorème 4. *If $0 \leqslant x_{(1)} \leqslant \cdots \leqslant x_{(n)} \leqslant \alpha$, then $\theta_0 = x_{(i)}$ for some i.*

Proof. It never happens that $\theta_0 < x_{(1)}$ because $g_{\theta_1}(x) > g_{\theta_0}(x)$ whenever $\theta_1 \in (\theta_0, x_{(1)})$. Similarly, it never happens that $\theta_0 > x_{(n)}$ because $g_{\theta_1}(x) > g_{\theta_0}(x)$ whenever $\theta_1 \in (x_{(n)}, \theta_0)$. We assume from now on that $\theta_0 \in [x_{(i)}, x_{(i+1)}]$ for some $i \in \{1, \cdots, n-1\}$. Suppose, for the sake of contradiction, that $x_{(i)} < \theta_0 < x_{(i+1)}$. We have:

$$g_{\theta_{\theta_0}}(x) = \left(\frac{2}{\alpha}\right)^n \frac{x_{(1)}}{\theta_0} \cdots \frac{x_{(i)}}{\theta_0} \frac{\alpha - x_{(i+1)}}{\alpha - \theta_0} \cdots \frac{\alpha - x_{(n)}}{\theta_0}$$

The numerator does not depend on θ_0. This motivates us to define function $h : [x_{(i)}, x_{(i+1)}] \to \mathbb{R}$ by:

$$h(\theta) = \frac{1}{\theta^i} \frac{1}{(\alpha - \theta)^{n-i}}$$

Then the second derivative is:

$$h''(\theta) = i(i+1)\theta^{-i-2}(\alpha - \theta)^{i-n} + (n-i)(n-i+1)\theta^{-i}(\alpha - \theta)^{i-n-2} > 0$$

Therefore, h is strictly convex and the maximum can only be at the boundary points. $\square$

We now demonstrate that the choice of i is not unique in the above theorem. The simplest case will be $x_i = x_j$ for any i and any j. For a nontrivial example, let $\alpha = 5$, $n = 3$. If $x = (2, 3, 4)$, then the maximum likelihood estimate is one of $\{2, 3, 4\}$. An estimate of 3 or 4 yields maximum likelihood $\frac{8}{375}$ while an estimate of 2 yields likelihood $\frac{16}{1125}$. Therefore, the maximum likelihood estimate can be 3 or 4 and is not unique.

Finally, the additional example $x = (1, 2, 4)$, estimate $\theta = 1, 2, 4$ gives likelihood $\frac{3}{250}, \frac{4}{375}, \frac{1}{125}$ repsectively. We conclude that $\theta = 1$ is the maximum likelihood estimate in this case.////