Méthodes d’analyse biostatistique projet 1

Wen, Zehai

2023-10-10 12:40:20-04:00

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:

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

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 $\mathrm{Var}(Y \mid X)$.

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 :

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

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 2

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. □

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 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. □

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.////