Mô hình thống kê tuyến tính nâng cao

Bài tập 01: Hàm hợp lý và Ước lượng hợp lý cực đại

Phân bài cho các nhóm:

- Nhóm A: bài 9, 11, 13, 14
- Nhóm B: bài 3, 4, 6, 15
- Nhóm C: bài 2, 7, 10, 16
- Nhóm D: bài 1, 5, 8, 12

Bài tập 1. Xác định phân phối xác suất nào trong các phân phối sau đây là thuộc vào họ phân phối mũ phân tán, nếu thuộc thì hãy xác định tham số tự nhiên θ, tham số phân tán ϕ, các hàm $b(\theta)$, $a(\phi)$ và $c(\phi; y)$:

1. Phân phối Beta
$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1},$$
với $y \in (0,1)$, $a > 0$, $b > 0$, $\Gamma(\cdot)$ là hàm gamma, tức là $\Gamma(x) = \displaystyle\int_0^{+\infty} t^{x-1}\exp(-t)\,\mathrm{d}\,t$.

2. Phân phối hình học:
$$f(y; p) = p(1-p)^{y-1},$$
với $y = 1, 2, 3, \ldots$, và $0 < p < 1$.

3. Phân phối Cauchy:
$$f(y; c, s) = \frac{1}{\pi c \left\{1 + \left(\frac{y-a}{s}\right)^2\right\}},$$
với $-\infty < y < -\infty$, $-\infty < c < +\infty$, và $s > 0$.

Bài tập 2. Xem xét phân phối Gamma, với hàm mật độ xác suất như sau:

$$f(y; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} \exp\left(-y/\beta\right),$$

với $y > 0$, $\alpha > 0$ (tham số hình dạng - shape parameter), $\beta > 0$ (tham số tỷ lệ - scale parameter), và $\Gamma(\cdot)$ là hàm gamma.

1. Chứng minh rằng phân phối Gamma thuộc vào họ phân phối mũ phân tán, xác định tham số tự nhiên θ, tham số phân tán ϕ, các hàm $b(\theta)$, $a(\phi)$ và $c(\phi; y)$.

2. Xác định hàm liên kết chính tắc.

Bài tập 3. Xem xét phân phối Gamma ngược, với hàm mật độ xác suất như sau:

$$f(y;\mu,\phi) = (2\pi y^3\phi)^{-1/2}\exp\left\{-\frac{1}{2\phi}\frac{(y-\mu)^2}{y\mu^2}\right\},$$

với $y > 0$, $\mu > 0$ và $\phi > 0$.

1. Chứng minh rằng phân phối Gamma ngược thuộc vào họ phân phối mũ phân tán, xác định tham số tự nhiên θ, tham số phân tán ϕ, các hàm $b(\theta)$, $a(\phi)$ và $c(\phi; y)$.

2. Xác định hàm liên kết chính tắc.

Bài tập 4. Xác định xem các hàm dưới đây có là các hàm liên kết phù hợp cho một mô hình hồi quy tuyến tính tổng quát. Với những trường hợp không phù hợp, hãy giải thích tại sao.

1. $g(\mu) = -1/\mu^2$ với $\mu > 0$.

2. $g(\mu) = |\mu|$ với $-\infty < \mu < +\infty$.

3. $g(\mu) = \log(\mu)$ với $\mu > 0$.

4. $g(\mu) = \mu^2$ với $-\infty < \mu < +\infty$.

5. $g(\mu) = \mu^2$ với $0 < \mu < +\infty$.

Bài tập 5. Xem xét phân phối Weibull với hàm mật độ xác suất như sau:

$$f(y;\alpha,\gamma) = \frac{\alpha}{\gamma}\left(\frac{y}{\gamma}\right)^{\alpha-1}\exp\left\{-\left(\frac{y}{\gamma}\right)^{\alpha}\right\},$$

với $y > 0$, $\alpha > 0$ và $\gamma > 0$.

1. Chứng minh rằng, về mặt tổng quát phân phối Weibull không thuộc vào họ phân phối mũ phân tán.

2. Chứng minh rằng nếu $\alpha = 1$, phân phối Weibull thuộc vào họ phân phối mũ phân tán. Khi này, xác định tham số tự nhiên θ, tham số phân tán ϕ, các hàm $b(\theta)$, $a(\phi)$ và $c(\phi; y)$.

3. Xác định hàm liên kết chính tắc.

Bài tập 6. Xem xét phân phối mũ với hàm mật độ xác suất như sau:

$$f(y;\lambda) = \lambda\exp(-\lambda y),$$

với $y > 0$ và $\lambda > 0$.

1. Xác định tham số tự nhiên θ, tham số phân tán ϕ, các hàm $b(\theta)$, $a(\phi)$ và $c(\phi; y)$.

2. Xác định hàm liên kết chính tắc.
3. Hãy chỉ ra điểm khó khăn trong thực tế có thể xảy ra khi sử dụng hàm liên kết chính tắc này.

Bài tập 7. Xây dựng hàm hợp lý, hàm log-likelihood cho một mẫu ngẫu nhiên $y_1, y_2, \ldots, y_n$, độc lập, từ phân phối hình học (geometric distribution), với hàm trọng lượng xác suất $\Pr(Y = y) = p(1-p)^y$, với $y = 0, 1, 2, \ldots$ và $0 < p < 1$. Và tìm:

(a) ước lượng hợp lý cực đại cho tham số p;

(b) tìm thông tin quan sát;

(c) tìm thông tin Fisher;

(d) sai số chuẩn của ước lượng hợp lý cực đại (tức là căn bậc hai của phương sai của ước lượng).

Bài tập 8. Xây dựng hàm hợp lý, hàm log-likelihood cho một mẫu ngẫu nhiên, $y_1, y_2, \ldots, y_n$, độc lập, từ phân phối có hàm mật độ

$$f(y; \lambda) = \lambda \exp(-\lambda y)$$

với $y > 0$ và $\lambda > 0$. Chứng minh rằng hàm log-likelihood là bất biến với phép biến đổi $\psi = 1/\lambda$. Và tìm:

(a) ước lượng hợp lý cực đại cho tham số λ;

(b) tìm thông tin quan sát;

(c) tìm thông tin Fisher;

(d) sai số chuẩn của ước lượng hợp lý cực đại (tức là căn bậc hai của phương sai của ước lượng);

từ đó, suy luận ra kết quả cho ψ.

Bài tập 9. Xây dựng hàm hợp lý, hàm log-likelihood cho một mẫu ngẫu nhiên, $y_1, y_2, \ldots, y_n$, độc lập, từ phân phối Weibull có hàm mật độ

$$f(y; \theta, \alpha) = \frac{\alpha}{\theta}\left(\frac{y}{\theta}\right)^{\alpha-1} \exp\left\{-\left(\frac{y}{\theta}\right)^{\alpha}\right\},$$

với $y > 0$ và $\theta, \alpha > 0$. Chứng minh rằng hàm log-likelihood là bất biến với phép biến đổi $\beta = \log(\theta)$, $\psi = \log(\alpha)$. Và tìm:

(a) ma trận thông tin quan sát;

(b) ma trận thông tin Fisher.

từ đó, suy luận ra kết quả cho β và ψ.

Bài tập 10. Giả sử rằng ta có dãy biến ngẫu nhiên $Y_0, Y_1, \ldots, Y_n$ sao cho phân phối điều kiện của Y_{i+1} khi biết $Y_i = y_i$ là một phân phối Poisson với trung bình θy_i, tức là

$$\Pr(Y_{i+1} = y_{i+1}; \theta | y_i) = \frac{(\theta y_i)^{y_{i+1}}}{y_{i+1}!} \exp\left(-\theta y_i\right),$$

với $y_{i+1} = 0, 1, 2, \ldots,$ và $\theta > 0$. Điều này có nghĩa là Y_{i+1} chỉ phụ thuộc vào Y_i. Giả sử Y_0 tuân theo phân phối Poisson với trung bình $\theta > 0$:

(a) tìm hàm mật độ xác suất đồng thời của một mẫu ngẫu nhiên, $y_0, y_1, \ldots, y_n$ tương ứng của các biến ngẫu nhiên $Y_0, Y_1, \ldots, Y_n$;

(b) xác định hàm hợp lý, hàm log-likelihood tương ứng;

(c) ước lượng hợp lý cực đại cho tham số θ;

(d) tìm thông tin quan sát;

(e) tìm thông tin Fisher.

Bài tập 11. Xét một mô hình location-scale với hai tham số μ và σ, với hàm mật độ

$$f(y; \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{y-\mu}{\sigma}\right),$$

với $-\infty < y < \infty$, $\mu \in \mathbb{R}$, $\sigma > 0$ và $g(\cdot)$ là một hàm mật độ xác suất.

(a) Chứng minh rằng thông tin Fisher cho một quan sát đơn lẻ có dạng

$$\mathcal{I}(\mu, \sigma) = \sigma^{-2} \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

trong đó, a, b và c có dạng một hàm $h(\cdot) = \log g(\cdot)$. Chứng minh rằng $b = 0$ nếu g đối xứng qua 0.

(b) Tìm a, b và c trong trường hợp:

- $g(u) = (2\pi)^{-1/2} \exp\left(-u^2/2\right)$ - mật độ phân phối chuẩn tắc;
- $g(u) = \exp\left(\kappa u - \exp(u)\right) / \Gamma(\kappa)$, với $\kappa > 0$ và được biết; $\Gamma(\kappa)$ là hàm gamma. (Hàm mật độ xác suất của phân phối log-gamma).

Bài tập 12. Giả sử rằng ta có dãy biến ngẫu nhiên $Y_1, \ldots, Y_n$, độc lập và $Y_i \sim \mathcal{B}(m, p_i)$ (phân phối nhị thức), với m cố định và p_i được mô hình bởi:

$$p_i = \frac{\exp\left(\beta_0 + \beta_1 X_i\right)}{1 + \exp\left(\beta_0 + \beta_1 X_i\right)},$$

$\beta_0, \beta_1 \in \mathbb{R}$ là các tham số chưa biết, X_i là biến ngẫu nhiên ghép cặp với Y_i, tức là (X_i, Y_i).

(a) tìm hàm mật độ xác suất đồng thời của một mẫu ngẫu nhiên, $(x_1, y_1), \ldots, (x_n, y_n)$ tương ứng của các biến ngẫu nhiên $(X_1, Y_1), \ldots, (X_n, Y_n)$;

(b) xác định hàm hợp lý, hàm log-likelihood cho (β_0, β_1);

(c) tìm ma trận thông tin quan sát;

(d) tìm ma trận thông tin Fisher;

Bài tập 13. Xét hàm log-likelihood có dạng như sau:

$$\ell(\lambda, \kappa) = n\kappa \log \lambda + (\kappa - 1) \sum_{i=1}^{n} \log y_i - \lambda \sum_{i=1}^{n} y_i - n \log \Gamma(\kappa).$$

(a) Chứng minh ma trận thông tin quan sát là

$$\mathcal{J}(\lambda, \kappa) = n \begin{pmatrix} \kappa/\lambda^2 & -1/\lambda \\ -1/\lambda & \mathrm{d}^2 \log \Gamma(\kappa)/\mathrm{d}\kappa^2 \end{pmatrix}$$

(b) Tìm ma trận thông tin Fisher.

(c) Giả sử rằng ta ghi $\lambda = \kappa/\mu$, với μ là trung bình của phân phối. Tìm hàm log-likelihood của (μ, κ), chứng minh rằng $\mathcal{J}(\mu, \kappa)$ là ngẫu nhiên, và ma trận thông tin Fisher là

$$\mathcal{I}(\mu, \kappa) = n \begin{pmatrix} 2\kappa/\mu^2 & 0 \\ 0 & \mathrm{d}^2 \log \Gamma(\kappa)/\mathrm{d}\kappa^2 - 1/\kappa \end{pmatrix}$$

Bài tập 14. Một đối tượng có thời gian sống T, nhưng các thử nghiệm để ước tính thời gian sống của đối tượng kết thúc sau một thời gian đã biết c, khi người thực hiện thử nghiệm nghỉ giải lao. Dữ liệu có sẵn là thời gian sống quan sát được $Y = \min(T, c)$ và $D = \mathrm{I}(T \leq c)$, trong đó D cho biết liệu T đã được quan sát hay chưa. Nếu $T > c$ thì T được cho là bị che khuyết bên phải (*right-censored*), tức là, chúng ta chỉ biết rằng giá trị của nó vượt quá c. Nếu T có hàm mật độ và hàm phân phối lần lượt là $f(t; \theta)$ và $F(t; \theta)$, khi đó, hàm mật độ của (Y, D) được xác định bởi:

$$f(y, d; \theta) = f(y; \theta)^d \left\{1 - F(y; \theta)\right\}^{1-d}.$$

(a) Hãy xây dựng hàm hợp lý cho dữ liệu độc lập $(y_1, d_1), \ldots, (y_n, d_n)$. (Gồm 2 phần: quan sát được, và bị che khuyết).

(b) Xác định hàm hợp lý khi T có phân phối mũ với tham số λ.

(c) Tìm ma trận thông tin quan sát và ma trận thông tin Fisher.

Bài tập 15. Xét y là một mẫu ngẫu nhiên $y_1, y_2, \ldots, y_n$, độc lập, từ phân phối Weibull với hàm mật độ xác suất:

$$f(y; \theta, \alpha) = \frac{\alpha}{\theta} \left(\frac{y}{\theta}\right)^{\alpha-1} \exp\left\{-\left(\frac{y}{\theta}\right)^{\alpha}\right\},$$

với $y > 0$ và $\theta, \alpha > 0$.

(a) Hãy xác định các score functions.

(b) Chứng minh rằng, MLE của θ là

$$\widehat{\theta}(\alpha) = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^{\alpha}\right)^{1/\alpha},$$

với $y_i > 0$, và $\alpha > 0$.

(c) Chứng minh rằng MLE của α là nghiệm của phương trình:

$$\alpha - \left(\frac{\sum_{i=1}^{n} y_i^{\alpha}\log(y_i)}{\sum_{i=1}^{n} y_i^{\alpha}} - \frac{1}{n}\sum_{i=1}^{n}\log(y_i)\right)^{-1} = 0,$$

với $y_i > 0$.

Bài tập 16. Nếu log-likelihood cho vectơ tham số θ với $p \times 1$ số chiều, là

$$\ell(\theta) = a + b^{\top}\theta - \frac{1}{2}\theta^{\top}C\theta,$$

trong đó, các hằng số a, b và C lần lượt là số, vectơ $p \times 1$ và ma trận xác định dương đối xứng $p \times p$.

(a) Hãy chứng minh rằng score function có thể được viết là $U(\theta) = b - C\theta$.

(b) Hãy tìm ma trận thông tin quan sát $\mathcal{J}(\theta)$.

(c) Thiết lập công thức nghiệm bằng phương pháp lặp với bất kỳ giá trị ban đầu nào của θ.