Bài giảng 2: Một số ước lượng và ứng dụng

TS. Tô Đức Khánh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Trường hè Toán học sinh viên năm 2025

Nội dung buổi học

Giới thiệu

Giả sử, ta quan sát được giá trị y của biến ngẫu nhiên Y.

- Hàm mật độ xác suất của Y đã được biết với tham số θ, tức là $f(y;\theta)$

  $\hookrightarrow$ là một hàm của y và θ.
- Đặt $\mathcal{Y}$ là không gian mẫu $\Rightarrow y \in \mathcal{Y}$
- Đặt Θ là không gian tham số $\Rightarrow \theta \in \Theta$.

  tổng quát, y và θ có thể là hai vector.

Bài toán
Mục tiêu của chúng ta là đưa ra nhận định hoặc tuyên bố về phân phối của Y, dựa trên dữ liệu quan sát y.

Theo giả định, ta có:

- hàm mật độ xác suất f đã biết;
- quan sát y;

$\hookrightarrow$ ta cần đưa ra một nhận định về khoảng giá trị phù hợp của $\theta \in \Theta$, tương ứng với giá trị quan sát y.

Giới thiệu

Một phương pháp cơ bản là dựa trên hàm "hợp lý" (likelihood function) của θ:

$$L(\theta) = f(y; \theta),$$

với y cố định và $\theta \in \Theta$.

Diễn giải: dựa vào dữ liệu y, giá trị tham số $\theta \in \Theta$ là đáng tin hơn $\theta' \in \Theta$, như là một chỉ số của mô hình xác suất tạo ra dữ liệu, nếu $L(\theta) > L(\theta')$.

Tức là giá trị $L(\theta)$ sẽ tương đối lớn nếu như θ là gần so với giá trị thật θ_0, cái đã tạo ra dữ liệu.

- Khi Y là rời rạc, ta sử dụng hàm trọng lượng xác suất $\Pr(Y = y; \theta)$
- Khi Y là liên tục, ta sử dụng hàm mật độ xác suất $f(y; \theta)$.

Xem thêm trong cuốn sách Davison (2003).

Giới thiệu

Khi $y = (y_1, y_2, \ldots, y_n)$, với y_i là các quan sát độc lập nhau của Y_i, khi đó,

$$L(\theta) = f(y; \theta) = f(y_1, y_2, \ldots, y_n; \theta) = \prod_{i=1}^{n} f_{Y_i}(y_i; \theta).$$

Kết quả này có được do tính độc lập y_i.

Trong thực tế, sẽ thuận tiện hơn khi xét hàm log-likelihood:

$$\ell(\theta) = \log L(\theta) = \log f(y; \theta),$$

ta đặt $\ell(\theta) = -\infty$ nếu $L(\theta) = 0$.

Khi $y = (y_1, y_2, \ldots, y_n)$, thì

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \theta).$$

Tính bất biến

Hai hàm likelihood được gọi là **tương đương** nếu chúng chỉ sai khác nhau một hằng số nhân (không phụ thuộc tham số).

Tính bất biến của hàm likelihood
Hàm likelihood (hoặc hàm log-likelihood) là bất biến với phép biến đổi 1-1 của dữ liệu.

Thật vậy, gọi $Z = g(Y)$, với g là một làm đơn điệu.

Khi đó, hàm mật độ của Z là

$$f_Z(z;\theta) = f_Y(y;\theta)\left|\frac{\mathrm{d}y}{\mathrm{d}z}\right|$$

với $z = g(y)$, và $y = g^{-1}(z)$.

Suy ra,

$$L_Z(\theta) = \left|\frac{\mathrm{d}y}{\mathrm{d}z}\right| \times L_Y(\theta),$$

dễ thấy, $\left|\frac{\mathrm{d}y}{\mathrm{d}z}\right|$ không phụ thuộc tham số θ.

Ví dụ 1 - Phân phối Poisson

Giả sử y là một giá trị quan sát từ một phân phối Poisson:

$$\Pr(Y = y; \theta) = \frac{\theta^y \exp(-\theta)}{y!},$$

với $y \in \mathbb{Z}_+$ và $\theta > 0$.

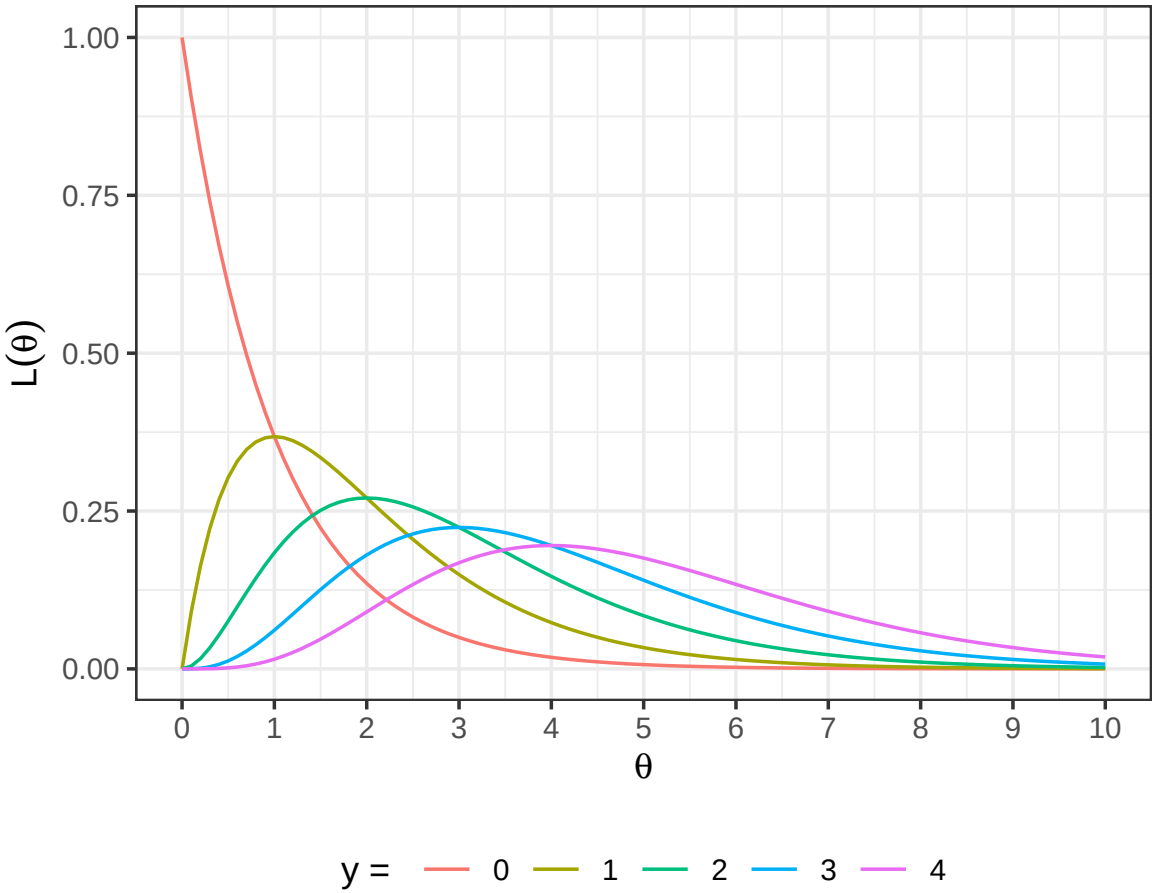
Khi đó, hàm likelihood là

$$L(\theta) = \frac{\theta^y \exp(-\theta)}{y!}.$$

Nếu

- $y = 0$, thì $L(\theta) = \exp(-\theta)$, hàm đồng điệu giảm của θ;
- $y > 0$, thì $L(\theta)$ sẽ đặt cực đại tại $\theta = y$, và có giới hạn 0 khi θ tiệm cận 0 hoặc ∞.

Ví dụ 1 - Phân phối Poisson



Ví dụ 2: phân phối mũ

Xét y là một mẫu ngẫu nhiên $y_1, y_2, \ldots, y_n$, độc lập, từ phân phối mũ với hàm mật độ xác suất

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

với $y > 0$ và $\theta > 0$.

Khi đó, hàm likelihood là

$$L(\theta) = \prod_{i=1}^{n} \theta^{-1} \exp\left(-y_i/\theta\right) = \theta^{-n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} y_i\right)$$

và hàm log-likelihood là

$$\ell(\theta) = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} y_i$$

Hàm likelihood và hàm log-likelihood đạt cực đại tại $\theta = \frac{1}{n}\sum_{i=1}^{n} y_i$

Ví dụ 3: phân phối chuẩn

Xét y là một mẫu ngẫu nhiên $y_1, y_2, \ldots, y_n$, độc lập, từ phân phối chuẩn với hàm mật độ xác suất

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

với $y \in \mathbb{R}$, $\mu \in \mathbb{R}$ và $\sigma > 0$.

Khi đó, hàm likelihood là

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

và hàm log-likelihood là

$$\ell(\theta) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2,$$

với $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$.

Thông tin của hàm log-likelihood

Xét hàm log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \theta)$$

Thông tin Fisher (Fisher information) được định nghĩa bởi:

$$\mathcal{I}(\theta) = -\mathbb{E}\left\{ \frac{\mathrm{d}^2 \ell(\theta)}{\mathrm{d}\theta^2} \right\}$$

Thông tin quan sát (observed information) được định nghĩa bởi:

$$\mathcal{J}(\theta) = -\frac{\mathrm{d}^2 \ell(\theta)}{\mathrm{d}\theta^2}$$

Thông tin của hàm log-likelihood

Ví dụ 1: xét hàm log-likelihood cho phân phối Poisson

$$\ell(\theta) = \log(\theta) \sum_{i=1}^{n} y_i - n\theta,$$

với $\theta > 0$.

Ta dễ dàng tính được

- $\mathcal{J}(\theta) = \frac{1}{\theta^2} \sum_{i=1}^{n} y_i$
- $\mathcal{I}(\theta) = \frac{n}{\theta}$

Ví dụ 2: xét hàm log-likelihood cho phân phối mũ

$$\ell(\theta) = -n \log(\theta) - \frac{1}{\theta} \sum_{i=1}^{n} y_i,$$

với $\theta > 0$. Hãy tìm $\mathcal{J}(\theta)$ và $\mathcal{I}(\theta)$.

Thông tin của hàm log-likelihood

Tổng quát, khi θ là một vecto p chiều, thì ta có

- ma trận thông tin Fisher (Fisher information matrix):

$$\mathcal{I}(\theta) = -\mathbb{E}\left\{\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^\top}\right\}$$

- ma trận thông tin quan sát (observed information matrix):

$$\mathcal{J}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^\top}$$

Chúng đều là các ma trận cỡ $p \times p$, với các phần tử thứ (r, s) lần lượt là

$$-\mathbb{E}\left\{\frac{\partial^2 \ell(\theta)}{\partial\theta_r\partial\theta_s}\right\}, \qquad -\frac{\partial^2 \ell(\theta)}{\partial\theta_r\partial\theta_s}.$$

Nhận xét:

- ma trận thông tin Fisher $\mathcal{I}(\theta)$ có thể xác định không cần dữ liệu;
- ma trận thông tin quan sát $\mathcal{J}(\theta)$ cần dữ liệu để xác định.

Thông tin của hàm log-likelihood

Ví dụ 3: xét hàm log-likelihood cho phân phối chuẩn

$$\ell(\theta) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2,$$

với $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$.

Ta dễ dàng tính được

- ma trận thông tin quan sát

$$\mathcal{J}(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{2n}{\sigma^3}(\overline{y} - \mu) \\ \frac{2n}{\sigma^3}(\overline{y} - \mu) & -\frac{n}{\sigma^2} + \frac{3}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 \end{pmatrix}$$

- ma trận thông tin Fisher

$$\mathcal{I}(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix}$$

Ước lượng hợp lý cực đại

Như đã giới thiệu ở phần định nghĩa, một giá trị "hợp lý" cho θ là giá trị sao cho $L(\theta) > L(\theta')$ hoặc tương đương $\ell(\theta) > \ell(\theta')$.

Ta cần tìm θ sao cho $L(\theta)$ hoặc $\ell(\theta)$ đạt cực đại:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta),$$

hay tương đương

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta).$$

Ta gọi $\widehat{\theta}$ là **ước lượng hợp lý cực đại - maximum likelihood estimator (MLE)**.

Để thuận tiện, ta sẽ xét phương trình thứ 2, với trường hợp tổng quát, θ là một vector p chiều.

Tính bất biến của MLE

Cho

- $\widehat{\theta}$ là MLE của tham số θ;
- $g(\cdot)$ là một hàm đơn điệu, 1-1 của θ, tức là $\psi = g(\theta)$.

Khi đó, $\widehat{\psi} = g(\widehat{\theta})$ cũng là MLE của ψ.

Điều này có được là bởi tính chất 1-1 của hàm $g(\cdot)$, tức là $\theta = g^{-1}(\psi)$, khi đó

$$\ell(\theta) = \ell\left(g^{-1}(\psi)\right) \equiv \ell^*(\psi).$$

Hơn nữa

$$\sup_{\psi} \ell^*(\psi) = \sup_{\psi} \ell\left(g^{-1}(\psi)\right) = \sup_{\theta} \ell(\theta).$$

Do đó, cực đại của $\ell^*(\psi)$ xác định tại $\psi = g(\theta) = g(\widehat{\theta})$, chứng minh rằng MLE của ψ là $g(\widehat{\theta})$.

Từ kết quả này, ta có thể viết $\widehat{\theta} = g^{-1}(\widehat{\psi})$.

Tính chất này được sử dụng trong các bài toán với miền xác định Θ của θ bị chặn.

Ước lượng hợp lý cực đại

Ước lượng hợp lý cực đại $\widehat{\theta}$ có thể được tìm bằng cách giải phương trình đạo hàm bậc 1. Tức là, $\widehat{\theta}$ là nghiệm của phương trình

$$U(\theta) \equiv \frac{\partial \ell(\theta)}{\partial \theta} = 0,$$

ta gọi $U(\theta)$ là hàm score (score function).

Để kiểm tra $\widehat{\theta}$ là một cực trị địa phương, ta kiểm tra điều kiện ma trận thông tin quan sát $\mathcal{J}(\theta)$ là xác định dương tại $\widehat{\theta}$.

Ước lượng hợp lý cực đại

Ví dụ 1 (tiếp theo): xét hàm log-likelihood cho phân phối Poisson

$$\ell(\theta) = \log(\theta) \sum_{i=1}^{n} y_i - n\theta,$$

với $\theta > 0$.

Hàm score được xác định bởi:

$$U(\theta) = \frac{1}{\theta} \sum_{i=1}^{n} y_i - n.$$

Giải phương trình $U(\theta) = 0$, ta thu được:

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Ta kiểm tra được rằng $\mathcal{J}(\widehat{\theta}) = n/\widehat{\theta} > 0$.

Ước lượng hợp lý cực đại

Ví dụ 3 (tiếp theo): xét hàm log-likelihood cho phân phối chuẩn

$$\ell(\theta) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2,$$

với $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$.

Hàm score được xác định bởi:

$$U(\theta) = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) \\ -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \mu)^2 \end{pmatrix}$$

Giải phương trình $U(\theta) = 0$, ta thu được:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu})^2}.$$

Ta kiểm tra được rằng $\mathcal{J}(\widehat{\mu}, \widehat{\theta})$ là xác định dương.

Phương pháp giải lặp Newton

Một phương pháp khác để xác định ước lượng hợp lý cực đại $\widehat{\theta}$ đó là giải lặp phương trình đạo hàm.

Cho trước một giá trị $\theta^{\dagger}$, áp dụng khai triển Taylor (bậc 1) cho hàm score tại $\theta^{\dagger}$, ta được

$$U(\theta) = U\left(\theta^{\dagger}\right) + \frac{\partial U\left(\theta^{\dagger}\right)}{\partial \theta}\left(\theta - \theta^{\dagger}\right)$$

Mặt khác, do $\widehat{\theta}$ là nghiệm của phương trình $U(\theta) = 0$, nên $U(\widehat{\theta}) = 0$ và

$$0 = U(\widehat{\theta}) = U\left(\theta^{\dagger}\right) + \frac{\partial U\left(\theta^{\dagger}\right)}{\partial \theta}\left(\widehat{\theta} - \theta^{\dagger}\right).$$

Suy ra,

$$\widehat{\theta} = \theta^{\dagger} + \mathcal{J}^{-1}\left(\theta^{\dagger}\right) U\left(\theta^{\dagger}\right),$$

với $\mathcal{J}^{-1}\left(\theta^{\dagger}\right)$ là ma trận nghịch đảo của $\mathcal{J}\left(\theta^{\dagger}\right)$.

Đây là một biến thể của phương pháp giải lặp Newton-Raphson.

Phương pháp giải lặp Newton

Thuật toán giải lặp

1. Chọn một giá trị bắt đầu $\theta^{(0)}$
2. Với bước lặp $t = 0$, ta tính

$$\theta^{(t+1)} = \theta^{(t)} + \mathcal{J}^{-1}\left(\theta^{(t)}\right) U\left(\theta^{(t)}\right),$$

3. Đặt $t = t + 1$, lặp lại bước 2 cho tới khi nào thuật toán hội tụ, có thể là

$$\left\|\theta^{(t+1)} - \theta^{(t)}\right\| < \varepsilon.$$

4. ước lượng hợp lý cực đại $\widehat{\theta} = \theta^{(t+1)}$.

Phương pháp giải lặp Newton

Ngoài ra, ta có thể thay thế $\mathcal{J}(\theta)$ bằng $\mathcal{I}(\theta)$, khi đó, thuật toán có tên Fisher scoring

$$\widehat{\theta} = \theta^{\dagger} + \mathcal{I}^{-1}\left(\theta^{\dagger}\right) U\left(\theta^{\dagger}\right).$$

Phương pháp thường được áp dụng khi:

- ma trận $\mathcal{J}(\theta)$ không được định nghĩa tốt;
- ma trận $\mathcal{J}(\theta)$ có công thức phức tạp.

Phương pháp giải lặp Newton

Ví dụ 4: 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 $\theta, \alpha > 0$.

Hàm log-likelihood được xác định là

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

Từ đây có xác định được hàm score là

$$U(\theta,\alpha) = \begin{pmatrix} -n\alpha/\theta + \alpha\theta^{-1}\sum_{i=1}^{n}(y_i/\theta)^{\alpha} \\ n/\alpha + \sum_{i=1}^{n}\log(y_i/\theta) - \sum_{i=1}^{n}(y_i/\theta)^{\alpha}\log(y_i/\theta) \end{pmatrix}$$

ta không thể giải phương trình này bằng phương pháp giải tích.

Phương pháp giải lặp Newton

Từ hàm score, ta xác định ma trận thông tin quan sát $\mathcal{J}(\theta, \alpha)$:

$$\mathcal{J}(\theta, \alpha) = \begin{pmatrix} j_{\theta,\theta}(\theta, \alpha) & j_{\theta,\alpha}(\theta, \alpha) \\ j_{\theta,\alpha}(\theta, \alpha) & j_{\alpha,\alpha}(\theta, \alpha) \end{pmatrix},$$

trong đó,

$$\begin{aligned} j_{\theta,\theta}(\theta, \alpha) &= -\frac{n\alpha}{\theta^2} + \frac{\alpha(\alpha+1)}{\theta^2} \sum_{i=1}^{n} \left(\frac{y_i}{\theta}\right)^{\alpha}, \\ j_{\theta,\alpha}(\theta, \alpha) &= \frac{n}{\theta} - \sum_{i=1}^{n} \frac{y_i^{\alpha}}{\theta^{\alpha+1}} \left(1 + \alpha \log\left(\frac{y_i}{\theta}\right)\right), \\ j_{\alpha,\alpha}(\theta, \alpha) &= \frac{n}{\alpha^2} + \sum_{i=1}^{n} \left(\frac{y_i}{\theta}\right)^{\alpha} \log\left(\frac{y_i}{\theta}\right), \end{aligned}$$

với $\theta, \alpha > 0$.

Phương pháp giải lặp Newton

Đặt $\beta = (\theta, \alpha)$, khi đó, nghiệm giải lặp là

$$\widehat{\beta} = \beta^{\dagger} + \mathcal{J}^{-1}\left(\beta^{\dagger}\right) U\left(\beta^{\dagger}\right),$$

Để đảm bảo ước lượng $\widehat{\theta}, \widehat{\alpha} > 0$, ta sử dụng biến đổi $\psi = (\log(\theta), \log(\alpha))$. Khi đó,

$$\widehat{\psi} = \psi^{\dagger} + \mathcal{J}^{-1}\left(\psi^{\dagger}\right) U\left(\psi^{\dagger}\right).$$

Sau đó, với phép biến đổi ngược, exp(), ta thu được kết quả $\widehat{\theta}, \widehat{\alpha} > 0$.

Áp dụng

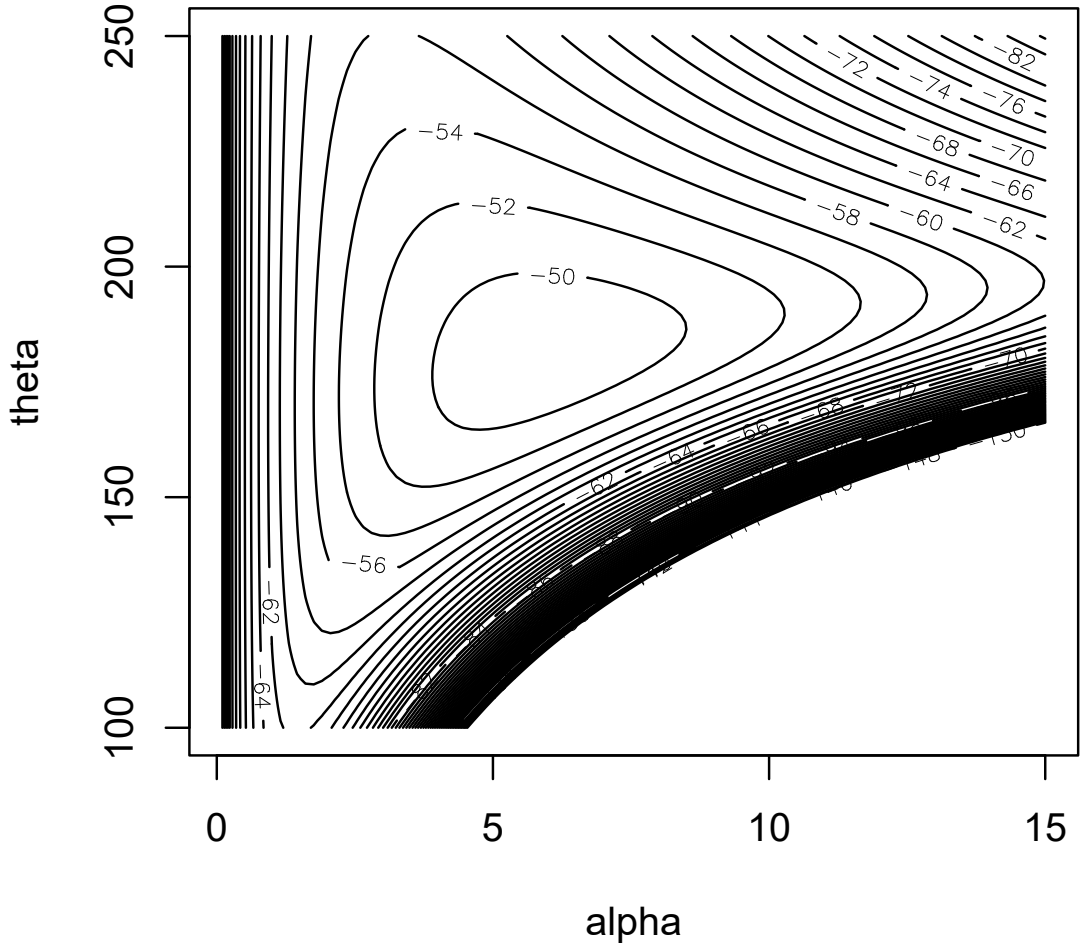
Ta áp dụng mô hình phân phối Weibull để mô hình hóa dữ liệu về thời gian hỏng của lò xo trong thí nghiệm với mức ứng suất 950 N/mm^2:

225, 171, 198, 189, 189, 135, 162, 135, 117, 162

Dựa vào công thức hàm log-likelihood của phân phối Weibull, ta có thể biểu diễn đồ thị như sau.

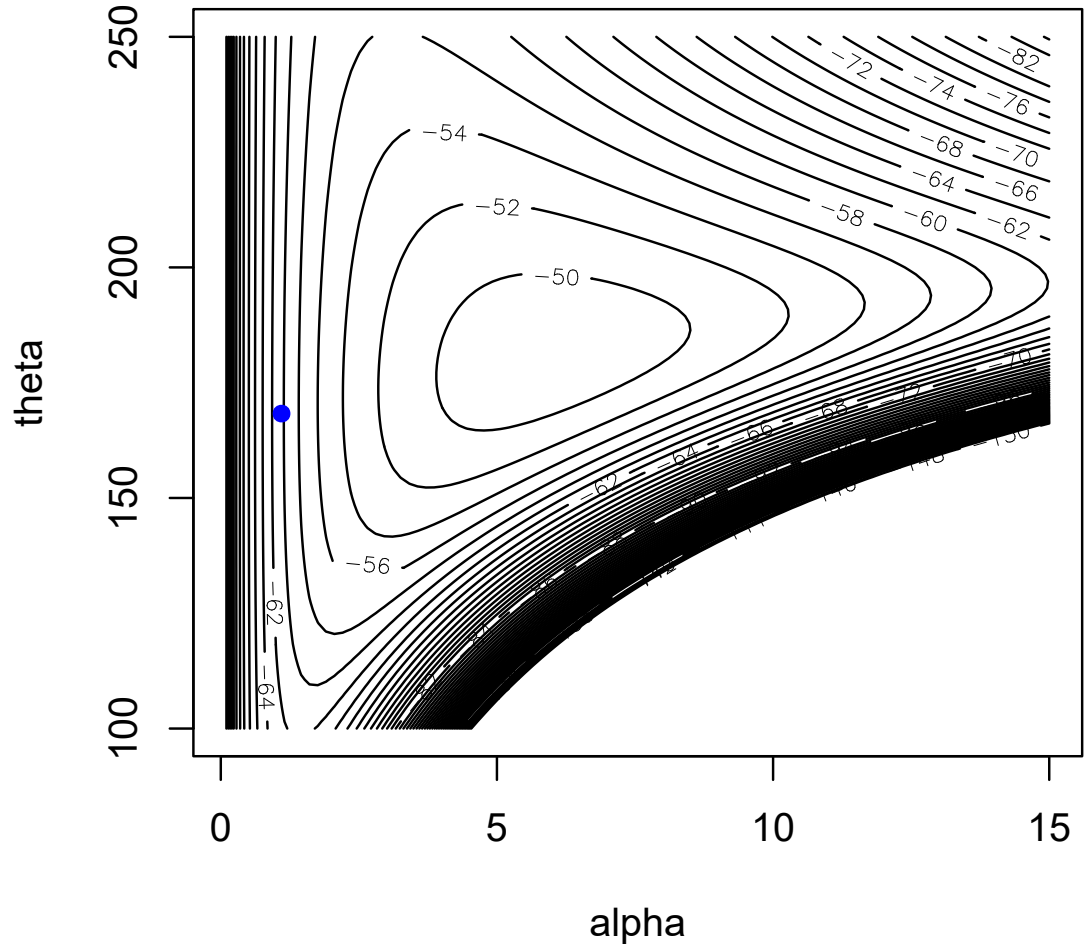
Phương pháp giải lặp Newton

Hình chiếu của hàm log-likelihood của phân phối Weibull.



Phương pháp giải lặp Newton

Thực hiện giải lặp, với sai số chặn là 10^{-9}.



Phương pháp giải lặp Newton

Bảng tổng hợp kết quả

Lần lặp t	$\theta^{(t)}$	$\alpha^{(t)}$	Sai số
0	168.3000	1.1000	
1	172.1154	2.1111	6.5226×10^{-1}
2	175.1454	3.6845	5.5723×10^{-1}
3	180.4374	5.2457	3.5452×10^{-1}
4	181.1004	5.9082	1.1899×10^{-1}
5	181.4075	5.9764	1.1604×10^{-3}
6	181.4056	5.9769	8.1710×10^{-5}
7	181.4056	5.9769	3.6697×10^{-9}
8	181.4056	5.9769	2.2204×10^{-16}

$\Rightarrow$ ước lượng MLE của θ và α là $\widehat{\theta} = 181.4056$, $\widehat{\alpha} = 5.9769$.

Thông tin của hàm likelihood và MLE

Nhắc lại rằng, thông tin quan sát của hàm log-likelihood

$$\mathcal{J}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^\top}.$$

Nhận xét:

- về mặt hình học, với $p = 1$, $\mathcal{J}(\theta)$ đo độ cong của $\ell(\theta)$;
- $\mathcal{J}(\theta)$ là một hàm tuyến tính theo n, khi $p = 1$;
- độ cong của $\ell(\theta)$ tại giá trị cực đại, sẽ tăng khi n tăng lên.
- khi thông tin quan sát tại một điểm $\theta^\dagger$, $\mathcal{J}(\theta^\dagger)$ càng lớn, thì $\theta^\dagger$ càng được gim chặt vào vùng cực trị của $\ell(\theta)$.

Mô hình thống kê

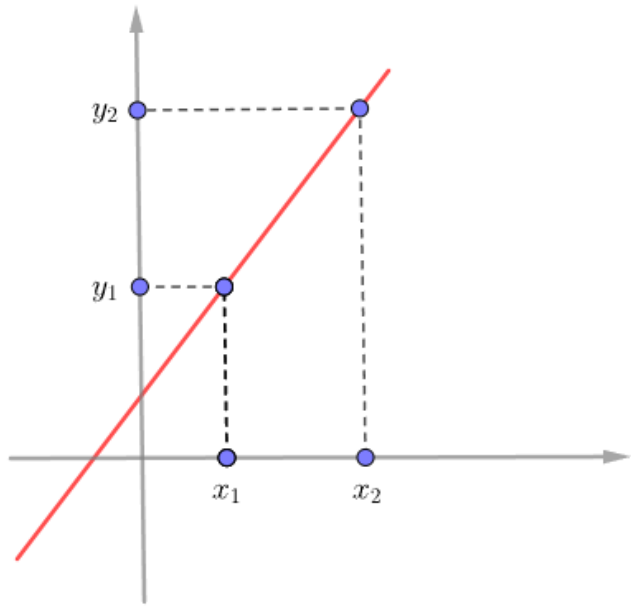
Xét phương trình đường thẳng:

$$y = a + bx$$

với a và b là các hằng số được biết trước.

- Cho trước một giá trị x_1 ta dễ dàng tính được tương ứng một giá trị y_1.

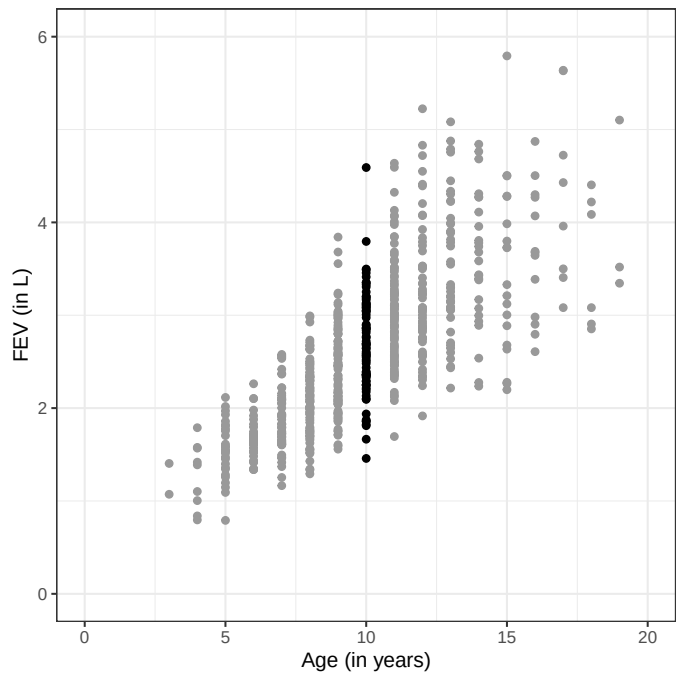
↪ Phương trình $y = a + bx$ là một dạng **mô hình toán** mô tả sự thay đổi giá trị y bởi giá trị của x.



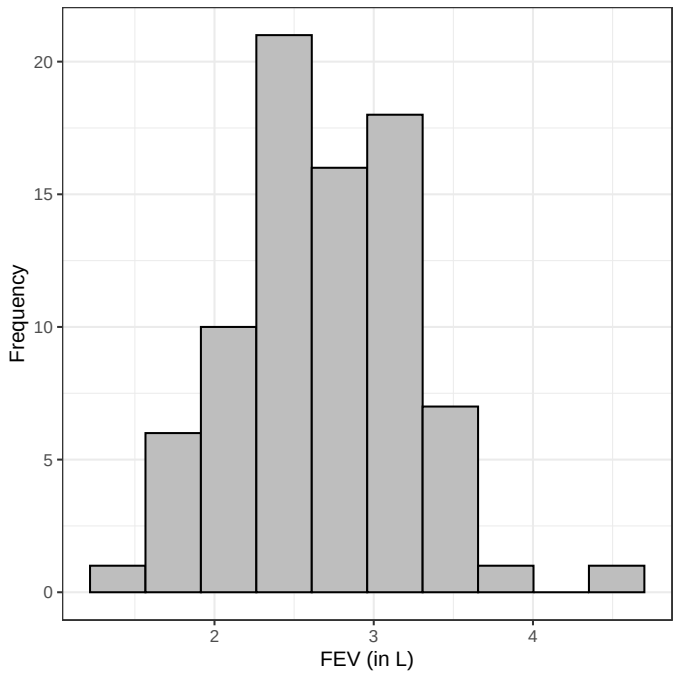
Về mặt mô hình toán, với một giá trị của x ta sẽ quan sát được một giá trị của y, tuy nhiên, đối với dữ liệu quan sát thực tế thì không như vậy.

Mô hình thống kê

Một giá trị của Age (x) có thể ghi nhận nhiều giá trị khác nhau của FEV (y). Ví dụ Age $= 10$ (các điểm màu đen).



Những quan sát FEV của các đối tượng 10 tuổi là ngẫu nhiên, và do đó, chúng có phân phối (*phân phối có điều kiện*).



Mô hình thống kê

Do đó, một mô hình toán cho FEV dựa vào Age, chẳng hạn:

$$\text{FEV} = a + b \times \text{Age},$$

sẽ chỉ có thể mô tả được giá trị trung bình của FEV tương ứng của một giá trị Age, mà không thể mô tả được phân phối của FEV.

$\hookrightarrow$ Điều này đòi hỏi phải có thêm một thành phần trong mô hình để mô tả cho phân phối của FEV.

Mô hình thống kê

Mô hình thống kê

Một mô hình thống kê (*statistical model*) là một mô hình gồm hai thành phần:

- thành phần hệ thống (*systematic component*);
- thành phần ngẫu nhiên (*random component*),

mô tả lần lượt hai đặc trưng của biến đáp ứng: trung bình và phân phối, dựa vào giá trị của một hoặc nhiều biến giải thích.

Mô hình thống kê

Ví dụ 1: Mô hình thống kê cho FEV dựa vào các biến Age, Ht, Gender và Smoke có thể là:

- thành phần hệ thống:

$$\mu = \mathbb{E}(\text{FEV}) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Ht} + \beta_3 \text{Gender} + \beta_4 \text{Smoke},$$

- thành phần ngẫu nhiên: $\text{FEV} \sim \mathcal{N}(\mu, \sigma^2)$;

trong đó, các tham số $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ và σ là chưa biết, cần phải ước lượng từ dữ liệu.

Cách viết mô hình như trên là cách viết dạng tổng quát. Khi ta có bộ dữ liệu với n quan sát độc lập, mô hình thống kê sẽ được biểu diễn cho quan sát thứ i, ví dụ:

- thành phần hệ thống:

$$\mu_i = \mathbb{E}(\text{FEV}_i) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Ht}_i + \beta_3 \text{Gender}_i + \beta_4 \text{Smoke}_i,$$

- thành phần ngẫu nhiên: $\text{FEV}_i \sim \mathcal{N}(\mu_i, \sigma^2)$ (chúng ta đang giả định các quan sát FEV_i có phương sai đồng nhất).

Mô hình thống kê

Ví dụ 2: Một mô hình thống kê cho FEV có thể là,

- thành phần hệ thống:
$$\mu_i = \mathbb{E}(\text{FEV}_i) = \beta_0,$$
- thành phần ngẫu nhiên: $\text{FEV}_i \sim \mathcal{N}(\mu_i, \sigma^2)$.

Thông thường sẽ có nhiều dạng khác nhau cho thành phần hệ thống cũng như thành phần ngẫu nhiên.

Mô hình hồi quy

Mô hình hồi quy

Nếu ta giả định rằng thành phần hệ thống, hay trung bình μ_i là một hàm f của p biến giải thích với các tham số chưa biết, tức là

$$\mu_i = \mathbb{E}(y_i) = f(x_{1i}, \ldots, x_{pi}; \beta_0, \beta_1, \ldots, \beta_p),$$

khi đó mô hình thống kê sẽ được gọi là một mô hình hồi quy (*regression model*).

Về mặt toán học, sẽ có rất nhiều sự kết hợp khác nhau của $x_{1i}, \ldots, x_{pi}$ và $\beta_0, \beta_1, \ldots, \beta_p$. Thông thường, ta giả sử rằng sự kết hợp này là tuyến tính, tức là

$$\mu_i = \mathbb{E}(y_i) = f(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}),$$

khi này, ta có mô hình hồi quy tuyến tính theo tham số.

Mô hình hồi quy

Ta có hai dạng mô hình tuyến tính như sau:

Mô hình hồi tuyến tính - Linear regression model: là một mô hình thống kê với

- thành phần hệ thống

$$\mu_i = \mathbb{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi},$$

- thành phần ngẫu nhiên $\mathbb{V}\mathrm{ar}(y_i) = \sigma^2$. Chú ý, chúng ta không cần đưa ra giả định cụ thể nào về phân phối.

Mô hình hồi tuyến tính tổng quát - Generalized linear model (GLM): là một mô hình thống kê với

- thành phần hệ thống

$$\mu_i = \mathbb{E}(y_i) = g^{-1}(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}),$$

  trong đó, $g(\cdot)$ là một hàm được xác định sao cho đồng biến và khả vi, và được gọi là hàm liên kết (*link function*),
- thành phần ngẫu nhiên: y_i tuân theo một phân phối xác định F với trung bình μ_i.

Xem thêm về mô hình thống kê trong các sách: Agresti (2015), Dobson and Barnett (2018).

Một số dạng có thể của thành phần hệ thống

Giả sử ta có biến đáp ứng là y với trung bình μ, các biến giải thích lần lượt là x_1, x_2, x_3 và x_4. Các dạng có thể của thành phần hệ thống là:

$$\begin{aligned}
\mu &= \beta_0 + \beta_1 x_1 + \beta_4 x_4 && (1)\\
\mu &= \beta_0 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_4 && (2)\\
\mu &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 && (3)\\
\mu &= \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_4 x_4 && (4)\\
\mu &= \beta_0 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_4 && (5)\\
1/\mu &= \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 && (6)\\
\log(\mu) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 && (7)\\
\mu &= \beta_0 + \exp(\beta_1 x_1) - \exp(\beta_2 x_2) + \beta_4 x_4^2 && (8)
\end{aligned}$$

- Các phương trình từ (1) - (7) đều có dạng tuyến tính theo tham số.
- Các phương trình từ (1) - (5) có thể được sử dụng để chỉ định một mô hình quy tuyến tính.

Thành phần ngẫu nhiên

Thành phần ngẫu nhiên của GLM

Thành phần ngẫu nhiên (*random component*) của GLM bao hàm 1 biến phản hồi Y với các quan sát độc lập nhau $(y_1, y_2, \ldots, y_n)$ có hàm mật độ xác suất hoặc hàm xác suất của một phân phối thuộc họ phân phối mũ phân tán (*exponential dispersion family*):

$$f_Y(y_i|\theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right),$$

trong đó,

- θ_i được gọi là tham số tự nhiên (*natural parameter*);
- $\phi > 0$ được gọi là tham số phân tán (*dispersion parameter*).

Thông thường,

- $a(\phi) = 1$ và $c(y_i, \phi) = c(y_i) \Rightarrow$ họ phân phối mũ tự nhiên (*natural exponential family*);
- $a(\phi) = \phi$ hoặc $a(\phi) = \phi/\omega_i$, với ω_i là trọng số đã biết.

Thành phần ngẫu nhiên

Một số phân phối thuộc họ phân phối mũ phân tán:

- phân phối Bernoulli, $\mathcal{B}(p)$, với $p \in (0,1)$;
- phân phối nhị thức, $\mathcal{B}(n,p)$ với n cố định và $p \in (0,1)$;
- phân phối multinomial, $\mathcal{M}(n; p_1, \ldots, p_k)$ với n cố định, $p_i \in (0,1)$ và $\sum_{i=1}^{k} p_i = 1$;
- phân phối Poisson, $\mathcal{P}(\lambda)$, $\lambda > 0$;
- phân phối chuẩn, $\mathcal{N}(\mu, \sigma^2)$, $\sigma > 0$;
- phân phối Gamma, $\mathcal{G}(\alpha, \beta)$

$$f_Y(y|\alpha,\beta) = \frac{1}{\Gamma(\alpha)} y^{\alpha-1} \exp(-y\beta)\, \beta^{\alpha},$$

với $y > 0$, và $\alpha, \beta > 0$;

- phân phối Beta, $\mathcal{B}e(\alpha,\beta)$

$$f_Y(y|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},$$

với $y \in [0,1]$, và $\alpha, \beta > 0$.

Hàm liên kết - Link function

Hàm liên kết - link function

Hàm liên kết (*Link function*) là một hàm đồng biến, được sử dụng để liên kết thành phần ngẫu nhiên với thành phần tuyến tính (*linear predictor*) của GLM:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij},$$

với $i = 1, \ldots, n$.

Cụ thể, với thành phần ngẫu nhiên, ta có $\mu_i = \mathbb{E}(y_i) \Rightarrow$ liên kết giữa η_i và μ_i được biểu diễn bởi $\eta_i = g(\mu_i) \Rightarrow g(\cdot)$ được gọi là hàm liên kết (*link function*) với tính chất:

- đồng biến (monotonic);
- khả vi (differentiable).

Hàm liên kết - Link function

Một số hàm liên kết tương ứng với thành phần ngẫu nhiên:

- phân phối chuẩn, $\eta_i = \mu_i$, hay $g(\cdot)$ là hàm đồng nhất (identity link);
- phân phối nhị thức, $\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$, logit link;
- phân phối nhị thức, $\eta_i = -\log\left(-\log(\mu_i)\right)$, log-log link;
- phân phối Poisson, $\eta_i = \log(\mu_i)$, log link;
- phân phối Gamma, $\eta_i = \mu_i^{-1}$, inverse link.

Canonical link

Trong một số trường hợp khi phân phối mũ phân tán có trung bình trùng với tham số tự nhiên (*natural parameter*) thì hàm liên kết $g(\cdot)$ được gọi là liên kết chính tắc (*canonical link*). Ví dụ:

- phân phối chuẩn, $\eta_i = \mu_i$, hay $g(\cdot)$ là hàm đồng nhất (identity link);
- phân phối nhị thức, $\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$, logit link;
- phân phối Poisson, $\eta_i = \log(\mu_i)$, log link;
- phân phối Gamma, $\eta_i = \mu_i^{-1}$, inverse link.

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

biến đáp ứng (response variable) biến ngẫu nhiên được quan tâm trong nghiên cứu, giá trị và sự biến động được diễn tả bởi mô hình thông qua những biến khác;

biến giải thích (explanatory variable) biến được dùng để giải thích sự thay đổi của biến đáp ứng bởi mô hình;

biến nhân tố (factor) là biến giải thích nhưng với dạng phân loại (định danh hoặc thứ bậc), biến này có tên khác là giả biến (*dummy variable*);

biến gây nhiễu (compounding variable) là một dạng đặc biệt của biến giải thích, một biến giải thích được coi là gây nhiễu nếu sự xuất hiện của nó làm thay đổi sự tác động của một hoặc nhiều biến giải thích khác.

Biến ngẫu nhiên trong mô hình

Giả sử rằng ta muốn xây dựng một mô hình thống kê để giải thích sự thay đổi của FEV bởi sự tách động của Age (độ tuổi), Ht (chiều cao), Gender (giới tính) và Smoke (trạng thái hút thuốc). Khi đó,

- FEV là biến đáp ứng;
- Age, Ht, Gender và Smoke là các biến giải thích;
- Gender và Smoke là biến nhân tố.

Diễn giải mô hình

Các mô hình hữu ích nhất khi chúng có những diễn giải hợp lý.

So sánh hai thành phần hệ thống sau:

$$\mu = \beta_0 + \beta_1 x \tag{9}$$
$$\log(\mu) = \beta_0 + \beta_1 x \tag{10}$$

Ta có nhận xét

- mô hình (9): x tăng 1 đơn vị thì μ tăng β_1 đơn vị;
- mô hình (10): x tăng 1 đơn vị thì $\log(\mu)$ tăng β_1 đơn vị;

  $\hookrightarrow$ x tăng 1 đơn vị thì μ tăng $\exp(\beta_1)$ lần. Tại sao?

Trong ứng dụng, ta cần xem xét lựa chọn mô hình (thành phần hệ thống) sao cho phù hợp với thực tế, hơn là tập trung quá nhiều vào công thức toán. Ví dụ:

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i},$$

trong đó,

- x_{1i}: số năm kinh nghiệm trong công việc;
- x_{2i}: giới tính (1 = nữ, 0 = nam).

Diễn giải mô hình

Hai mục tiêu chính của mô hình hồi quy là

Dự đoán: để tạo ra những dự đoán chính xác từ dữ liệu mới hoặc dữ liệu trong tương lai.

Hiểu và diễn giải: để hiểu các biến có liên quan với nhau như thế nào.

Ví dụ, hãy xem xét nghiên cứu về dung tích phổi.

- Ta bắt đầu với một câu hỏi về xác định mối liên hệ giữa trạng thái hút thuốc, giới tính, độ tuổi, chiều cao với dung tích phổi (FEV).
  $\hookrightarrow$ các hệ số trong mô hình sẽ miêu tả mối liên hệ (tuyến tính), và ta cần diễn giải triệt để các hệ số này.
- Mặt khác, trong tương lai, nếu ta thu thập được dữ liệu về trạng thái hút thuốc, giới tính, độ tuổi và chiều cao của một nhóm người, ta có thể dự đoán dung tích phổi (FEV) dựa vào mô hình, mà không cần yêu cầu thực hiện thí nghiệm.

Hàm trung bình và hàm phương sai

Từ hàm mật độ

$$f_{Y_i}(y_i|\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right),$$

ta tính được hàm sinh mô-men $M_Y(t)$ với $|t| < \delta$, $\delta > 0$:

$$\begin{aligned} M_{Y_i}(t) &= \mathbb{E}\left\{\exp(tY_i)\right\} \\ &= \int_\Omega \exp(t\,y_i)\exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}\mathrm{d}y_i \\ &= \exp\left\{\frac{b(t\,a(\phi)+\theta_i) - b(\theta_i)}{a(\phi)}\right\}, \end{aligned}$$

ở đây, Ω là miền xác định của Y_i.

Nhắc lại rằng, nếu hàm sinh mô-men tồn tại và khả vi, thì mô-men bậc k của Y được tính bởi

$$\mathbb{E}\left(Y^k\right) = \frac{\mathrm{d}^k}{\mathrm{d}t^k} M_Y(t)\bigg|_{t=0}$$

Hàm trung bình và hàm phương sai

Trung bình μ_i (hay kỳ vọng) của Y_i chính là mô-men bậc $k = 1$:

$$\mu_i = \mathbb{E}(Y_i) = \frac{\mathrm{d}}{\mathrm{d}t} M_{Y_i}(t)\bigg|_{t=0} = b'(\theta_i).$$

Mô-men bậc $k = 2$ của Y_i là

$$\mathbb{E}\left(Y_i^2\right) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} M_{Y_i}(t)\bigg|_{t=0} = \left(b'(\theta_i)\right)^2 + a(\phi)b''(\theta_i).$$

Do đó, ta có phương sai của Y_i là

$$\mathbb{V}\mathrm{ar}(Y_i) = \mathbb{E}\left(Y_i^2\right) - \left(\mathbb{E}(Y_i)\right)^2 = \phi b''(\theta_i) = a(\phi)V(\mu_i),$$

với $V(\mu) = b''(\theta_i)$, và ta gọi $V(\mu_i)$ là hàm phương sai (*variance function*).

$\Rightarrow$ về mặt tổng quát $\mathbb{V}\mathrm{ar}(Y_i)$ có thể được thay đổi theo μ_i.

Hàm trung bình và hàm phương sai

Ví dụ 1: Với phân phối chuẩn, $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, ta có $b(\theta_i) = \mu_i^2/2$ và $a(\phi) = \sigma^2$, do đó:

$$\mathbb{E}(Y_i) = b'(\theta_i) = \mu_i, \qquad \text{và} \qquad \mathbb{V}\text{ar}(Y_i) = \sigma^2 b''(\theta_i) = \sigma^2,$$

điều này dẫn tới hàm phương sai $V(\mu_i) = 1$.

Ví dụ 2: Với phân phối Poisson, $Y_i \sim \mathcal{P}(\mu_i)$, ta có $b(\theta_i) = \exp(\theta_i) = \mu_i$ và $a(\phi) = 1$, do đó:

$$\mathbb{E}(Y_i) = b'(\theta_i) = \exp(\theta_i) = \mu_i \qquad \text{và} \qquad \mathbb{V}\text{ar}(Y_i) = a(\phi) b''(\theta_i) = \mu_i$$

điều này dẫn tới hàm phương sai $V(\mu_i) = \mu_i$.

Ví dụ 3: Với phân phối nhị thức, $Y_i \sim \mathcal{B}(m_i, p_i)$, ta chứng minh được rằng, $Y_i/m_i \sim \text{EDM}(b(\theta_i), a(\phi))$ với $b(\theta_i) = \log(1 + \exp(\theta_i))$ và $a(\phi) = 1/m_i$, do đó,

$$\mathbb{E}\left(\frac{Y_i}{m_i}\right) = b'(\theta_i) = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = p_i \equiv \mu_i$$

và

$$\mathbb{V}\text{ar}\left(\frac{Y_i}{m_i}\right) = a(\phi) b''(\theta_i) = \frac{p_i(1 - p_i)}{m_i}$$

điều này dẫn tới hàm phương sai $V(\mu_i) = \mu_i(1 - \mu_i)$.

Hàm hợp lý cho GLM

Đặt

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$,
- $\mathbf{X}_i = (1, X_{1i}, \ldots, X_{pi})^\top$,
- $$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & \ldots & X_{p1} \\ 1 & X_{12} & X_{22} & \ldots & X_{p2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \ldots & X_{pn} \end{pmatrix},$$
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$,
- $\eta_i \equiv \eta_i(\boldsymbol{\beta}) = \mathbf{X}_i^\top \boldsymbol{\beta}$,
- $\eta \equiv \eta(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$.

Hàm hợp lý cho GLM

Do $Y_i \sim \text{EDM}(b(\theta_i), a(\phi)) \Rightarrow \mathbb{E}(Y_i) = \mu_i = b'(\theta_i)$.

Mặt khác, ta cũng có

- $\mu_i = g^{-1}(\eta_i)$;
- $b'(\theta_i) = \mu_i$;

$\Rightarrow$ ta có thể viết θ_i dưới dạng hàm ẩn của β, như sau $\theta_i \equiv \theta_i(\mu_i(\eta_i(\beta)))$.

Từ đây, ta có hàm mật độ xác suất của Y_i là

$$f(Y_i; \beta, a(\phi)) = \exp\left\{\frac{Y_i\theta_i(\mu_i(\eta_i(\beta))) - b(\theta_i(\mu_i(\eta_i(\beta))))}{a(\phi)} + c(Y_i, \phi)\right\},$$

và do đó, hàm log-likelihood là

$$\ell(\beta, a(\phi)) = \sum_{i=1}^{n} \ell_i(\beta, a(\phi)) = \sum_{i=1}^{n} \left\{\frac{Y_i\theta_i(\mu_i(\eta_i(\beta))) - b(\theta_i(\mu_i(\eta_i(\beta))))}{a(\phi)} + c(Y_i, \phi)\right\}.$$

Bởi vì các hàm $b(\cdot)$ và $g(\cdot)$ là liên tục và khả vi nên ta có thể tính được đạo hàm của chúng, cũng như là hàm ẩn $\theta_i(\mu_i(\eta_i(\beta)))$.

Hàm score cho GLM

Hàm score cho GLM, $\mathbf{U}(\beta)$ được xác định bởi

$$\mathbf{U}(\beta) = \frac{\partial \ell(\beta, a(\phi))}{\partial \beta} = \sum_{i=1}^{n} \frac{\mathrm{d}\ell_i}{\mathrm{d}\theta_i} \frac{\mathrm{d}\theta_i}{\mathrm{d}\mu_i} \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} \frac{\partial \eta_i}{\partial \beta}.$$

Thành phần đạo hàm thứ nhất được tính như sau

$$\frac{\mathrm{d}\ell_i}{\mathrm{d}\theta_i} = \frac{\mathrm{d}}{\mathrm{d}\theta_i} \left(\frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i; \phi) \right) = \frac{Y_i - \mu_i}{a(\phi)},$$

kết quả trên có được là do $b'(\theta_i) = \mu_i$.

Mặt khác, ta lại có $\theta_i = (b')^{-1}(\mu_i)$, áp dụng quy tắc đạo hàm của hàm ngược, ta có

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}\mu_i} = \frac{\mathrm{d}(b')^{-1}(\mu_i)}{\mathrm{d}\mu_i} = \frac{1}{b''((b')^{-1}(\mu_i))} = \frac{1}{V(\mu_i)}.$$

Hàm score cho GLM

Bởi vì $\mu_i = g^{-1}(\eta_i)$, áp dụng quy tắc đạo hàm của hàm ngược, ta có được

$$\frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} = \frac{1}{g'(\mu_i)}.$$

Ta dễ dàng tính được $\frac{\partial \eta_i}{\partial \beta} = \frac{\partial \mathbf{X}_i^\top \beta}{\partial \beta} = \mathbf{X}_i$. Từ các kết quả này, ta suy ra

$$\mathbf{U}(\beta) = \frac{1}{a(\phi)} \sum_{i=1}^{n} W_i g'(\mu_i)(Y_i - \mu_i)\mathbf{X}_i,$$

trong đó, $W_i = \frac{1}{V(\mu_i)\left(g'(\mu_i)\right)^2}$ và được gọi là *working weights*.

Đặc biệt, khi sử dụng hàm liên kết chính tắc (tức là $g(\mu_i) = \theta_i$), thì

$$g'(\mu_i) = \frac{\mathrm{d}(b')^{-1}(\mu_i)}{\mathrm{d}\mu_i} = \frac{1}{V(\mu_i)},$$

lúc đó, hàm score $\mathbf{U}(\beta)$ rút gọn thành

$$\mathbf{U}(\beta) = \frac{1}{a(\phi)} \sum_{i=1}^{n} (Y_i - \mu_i)\mathbf{X}_i.$$

Hàm score cho GLM

Hàm score $\mathbf{U}(\beta)$ được viết lại dưới dạng ma trận như sau:

$$\mathbf{U}(\beta) = \frac{1}{a(\phi)}\mathbf{X}^\top \mathbf{W}\mathbf{G}\,(\mathbf{Y} - \mu)\,,$$

trong đó

- $\mathbf{W} = \begin{pmatrix} W_1 & 0 & \dots & 0 \\ 0 & W_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & W_n \end{pmatrix}$,
- $\mathbf{G} = \begin{pmatrix} g'(\mu_1) & 0 & \dots & 0 \\ 0 & g'(\mu_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & g'(\mu_n) \end{pmatrix}$,
- $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^\top$.

Ma trận thông tin

Với biểu thức tổng quát của $\mathbf{U}(\beta)$, ta tìm được ma trận thông tin quan sát:

$$\begin{aligned}\mathcal{J}(\beta) &= -\frac{\partial \mathbf{U}(\beta)}{\partial \beta} \\ &= -\frac{1}{a(\phi)}\sum_{i=1}^{n}\left\{\left[\frac{\partial}{\partial \beta}\left(W_i g'(\mu_i)\right)\right](Y_i-\mu_i) - W_i g'(\mu_i)\frac{\partial \mu_i}{\partial \beta}\right\}\mathbf{X}_i.\end{aligned}$$

Nhận xét rằng,

- $\mathbb{E}(Y_i) = \mu_i$;
- $\dfrac{\partial \mu_i}{\partial \beta} = \dfrac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i}\dfrac{\partial \eta_i}{\partial \beta} = \dfrac{1}{g'(\mu_i)}\mathbf{X}_i.$

Khi đó, ta tính được ma trận thông tin Fisher

$$\mathcal{I}(\beta) = \mathbb{E}\left(\mathcal{J}(\beta)\right) = \frac{1}{a(\phi)}\sum_{i=1}^{n} W_i \mathbf{X}_i \mathbf{X}_i^\top = \frac{1}{a(\phi)}\mathbf{X}^\top \mathbf{W}\mathbf{X}.$$

Đặc biệt, khi sử dụng hàm liên kết chính tắc:

$$\mathcal{J}(\beta) = \mathcal{I}(\beta) = \frac{1}{a(\phi)}\mathbf{X}^\top \mathbf{G}\mathbf{X}.$$

Công thức nghiệm lặp cho β

Ước lượng hợp lý cực đại (MLE), $\widehat{\beta}$, được xác định bởi việc giải hệ phương trình đạo hàm (*score equations*):

$$\mathbf{U}(\beta) = \frac{1}{a(\phi)} \sum_{i=1}^{n} W_i g'(\mu_i)(Y_i - \mu_i)\mathbf{X}_i = \mathbf{0},$$

với $\mu_i = g^{-1}(\mathbf{X}_i^\top \beta)$.

Về mặt tổng quát, hệ phương trình này không có nghiệm giải tích.

$\Rightarrow$ ta cần tìm nghiệm qua phương pháp giải lặp.

Công thức nghiệm lặp cho β

Áp dụng công thức nghiệm lặp Newton cho MLE, ta có:

$$\beta^{(r+1)} = \beta^{(r)} + \mathcal{J}^{-1}\left(\beta^{(r)}\right)\mathbf{U}\left(\beta^{(r)}\right).$$

Công thức này là khá phức tạp trừ khi sử dụng liên kết chính tắc.

Do đó, trong tổng quát, ta sử dụng phương pháp Fisher Scoring:

$$\beta^{(r+1)} = \beta^{(r)} + \mathcal{I}^{-1}\left(\beta^{(r)}\right)\mathbf{U}\left(\beta^{(r)}\right).$$

Áp dụng các công thức biểu diễn, ta có:

$$\beta^{(r+1)} = \beta^{(r)} + \left(\mathbf{X}^\top \mathbf{W}^{(r)}\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{W}^{(r)}\mathbf{G}^{(r)}\left(\mathbf{Y} - \boldsymbol{\mu}^{(r)}\right),$$

trong đó, $\mathbf{W}^{(r)}$, $\boldsymbol{\mu}^{(r)}$ và $\mathbf{G}^{(r)}$ lần lượt được tính dựa vào hệ số $\beta^{(r)}$.

Công thức nghiệm lặp cho β

Công thức lặp Fisher Scoring có thể được viết lại là

$$\beta^{(r+1)} = \left(\mathbf{X}^\top \mathbf{W}^{(r)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(r)} \boldsymbol{z}^{(r)},$$

trong đó,

- $\boldsymbol{z}^{(r)} = \boldsymbol{\eta}^{(r)} + \mathbf{G}^{(r)} \left(\mathbf{Y} - \boldsymbol{\mu}^{(r)}\right)$,
- $\boldsymbol{\eta}^{(r)} = \mathbf{X}\beta^{(r)}$.

$\Rightarrow$ công thức lặp có dạng của ước lượng bình phương nhỏ nhất (least square), chỉ khác là nó có thêm trọng số, và trọng số này được tính lại sau mỗi bước lặp.

$\Rightarrow$ công thức lặp được gọi là **lặp bình phương nhỏ nhất với trọng số được tính lại** (**iteratively re-weighted least square**), hay viết tắt bởi **IWLS**.

Do đó, phương pháp Fisher scoring cho GLM còn được là thuật toán IWLS.

Chú ý: trong biểu thức lặp của $\beta^{(r+1)}$, không có sự xuất hiện của tham số $a(\phi)$, vì vậy ước lượng $\widehat{\beta}$ được tính mà không cần thông tin của $a(\phi)$.

Chọn giá trị bắt đầu $\boldsymbol{\beta}^{(0)}$

Ở thời điểm bắt đầu,

- ta không bất kỳ có thông tin về sự liên hệ giữa biến đáp ứng Y và các biến giải thích $X_1, \ldots, X_p$;
- thông tin duy nhất là trung bình mẫu của Y, tức là $\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i$.

Do đó, cách đơn giản để bắt đầu là ta giả định biến phản hồi Y không có sự liên hệ tuyến tính với các biến giải thích $X_1, \ldots, X_p$, tức là

- $\beta_0^{(0)} = \overline{Y}$;
- $\beta_j^{(0)} = 0$ với mọi $j = 1, \ldots, p$,

và do đó $\boldsymbol{\beta}^{(0)} = \left(\overline{Y}, 0, \ldots, 0\right)$.

$\Rightarrow \mu_i^{(0)} = g^{-1}\left(\overline{Y}\right)$ với mọi $i = 1, \ldots, n$.

Một điểm hạn chế của cách chọn này thuật toán IWLS có thể hội tụ chậm, và cũng có thể không hội tụ trong một số trường hợp mô hình phức tạp.

Chọn giá trị bắt đầu $\beta^{(0)}$

Nhận thấy rằng, ở lần lặp đầu tiên của IWLS,

- $\mathbf{W}^{(0)}$
- $\boldsymbol{z}^{(0)}$

chỉ phụ thuộc vào hệ số $\beta^{(0)}$ thông qua giá trị trung bình $\mu_i^{(0)}$.

Do đó, ta có thể chọn trực tiếp giá trị bắt đầu $\mu_i^{(0)}$ thay vì phải tính.

Mặt khác, mục tiêu của việc sử dụng GLM là để ước lượng $\widehat{\mu}_i$ sao cho gần nhất có thể với giá trị quan sát Y_i.

$\Rightarrow$ ta có thể bắt đầu thuật toán bằng cách chọn giá trị bắt đầu $\mu_i^{(0)} = Y_i$.

Chú ý kỹ thuật: khi hàm liên kết của mô hình có dạng logarithm hoặc nghịch đảo $\Rightarrow$ thuật toán không thể khởi động nếu $Y_i = 0$.

$\Rightarrow$ ta cần phải áp dụng một số tinh chỉnh nhỏ, ví dụ:

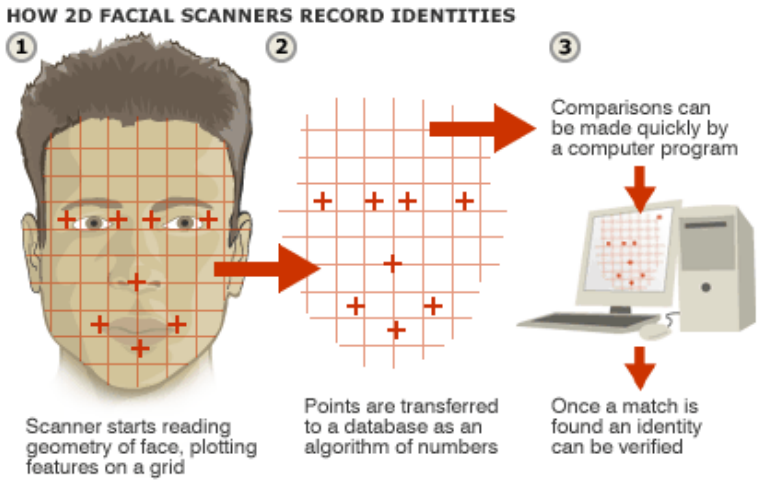
- $\mu_i^{(0)} = Y_i + 0.1$;
- $\mu_i^{(0)} = (mY_i + 0.5)/(m+1)$, thay vì đặt $\mu_i^{(0)} = 0$ hoặc 1, đối với trường hợp của phân phối nhị thức.

Cách chọn điểm bắt đầu như thế này giúp thuật toán IWLS hội tụ nhanh, và nó cũng được áp dụng trong các phần mềm thống kê.

Ví dụ: Human Face Recognition

Thuật toán nhận diện khuôn mặt người thường tập trung 3 yếu tố

- sự khác biệt về cường độ điểm ảnh ở vùng mắt (`eyediff`);
- sự khác biệt về cường độ điểm ảnh ở vùng mũi/má (`nosecheekdiff`);
- sự thay đổi cường độ điểm ảnh so sánh giữa hai hình ảnh của cùng một người (`variabilityratio`).



Dựa trên các yếu tố này, thuật toán sẽ so sánh 1 bức hình của 1 người với các hình còn lại (có duy nhất 1 hình của cùng 1 người), kết quả trả về:

- khớp (xác định đúng hình của người đang xét) - 1;
- không khớp - 0.

Ví dụ: Human Face Recognition

Để đánh giá sự ảnh hưởng của các yếu tố này tới khả năng nhận diện đúng của thuật toán, một thí nghiệm được tiến hành:

	match	eyediff	nosecheekdiff	variabilityratio
1	1	0.0096827670	0.027914422	1.0862036
2	1	0.0276138930	0.017821209	0.9898404
3	1	0.0153665190	0.025831893	1.0109830
4	1	0.0133714650	0.024047631	1.0127243
5	1	0.0114276410	0.017800852	1.0443961
6	1	0.0686048350	0.050388073	0.9508176
7	1	0.0092405937	0.016042421	1.0067264
8	1	0.0303490000	0.014707760	1.0186794
9	1	0.0310754380	0.069049962	1.0307903
10	0	0.0072111477	0.099909148	0.9289499

Ví dụ: Human Face Recognition

Ta quan tâm tới yếu tố eyediff và muốn đánh giá sự ảnh hưởng của yếu tố này tới xác suất nhận diện đúng $\Pr(\texttt{match} = 1|\texttt{eyediff})$.

Một mô hình được đề xuất là mô hình logistic

$$\Pr(\texttt{match}_i = 1|\texttt{eyediff}_i) = \frac{\exp\left(\beta_0 + \beta_1 \texttt{eyediff}_i\right)}{1 + \exp\left(\beta_0 + \beta_1 \texttt{eyediff}_i\right)},$$

$i = 1, \ldots, n$. Đặt $\pi_i = \Pr(\texttt{match}_i = 1|\texttt{eyediff}_i)$ và $\eta_i(\beta) = \beta_0 + \beta_1 \texttt{eyediff}_i$.

Đối với mô hình logistic

- $\mu_i = \pi_i$;
- hàm liên kết $g(u) = \dfrac{\exp(u)}{1 + \exp(u)}$;
- $a(\phi) = 1$
- hàm log-likelihood

$$\ell(\beta) = \sum_{i=1}^{n} \{Y_i \eta_i(\beta) - \log\left(1 + \exp(\eta_i(\beta))\right)\}$$

Ví dụ: Human Face Recognition

Ta tìm công thức nghiệm lặp IRLS

$$\beta^{(t+1)} = \beta^{(t)} + \left(\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \left(\mathbf{Y} - \boldsymbol{\pi}^{(t)}\right),$$

với

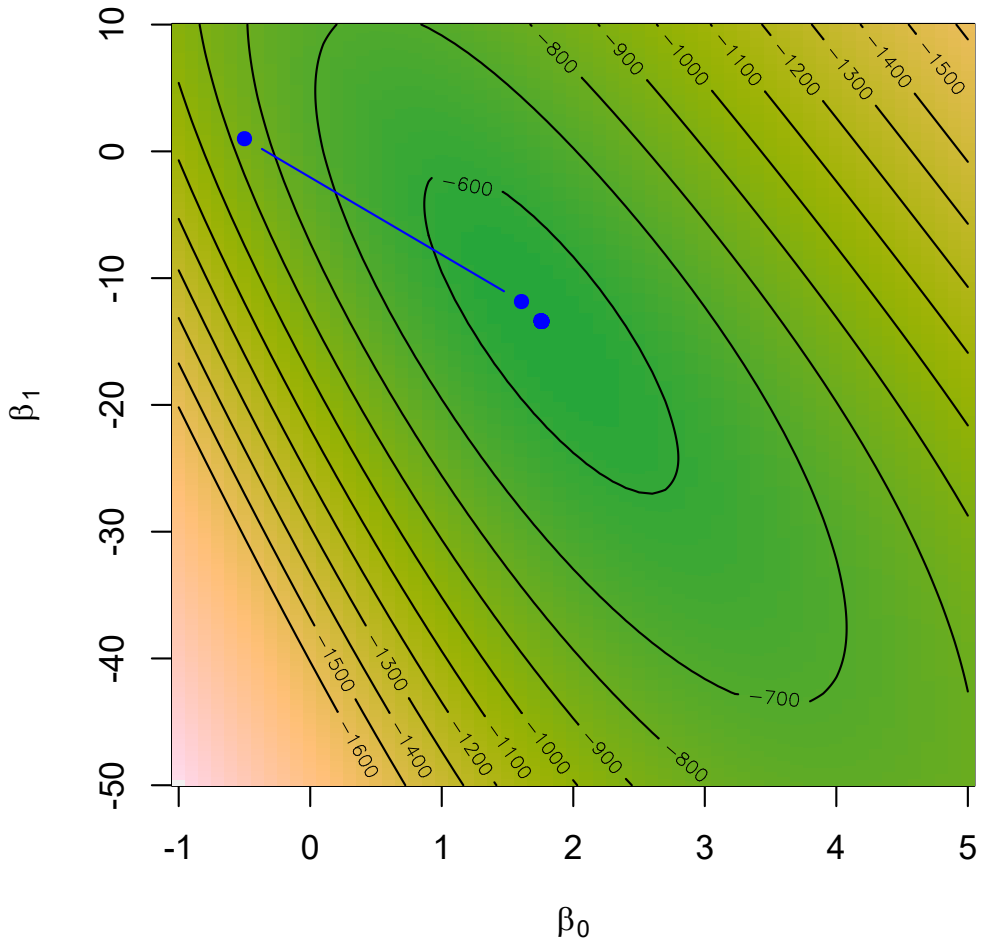
$$\mathbf{Y} = \begin{pmatrix} \texttt{match}_1 \\ \texttt{match}_2 \\ \vdots \\ \texttt{match}_n \end{pmatrix}, \quad \pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & \texttt{eyediff}_1 \\ 1 & \texttt{eyediff}_2 \\ \vdots & \vdots \\ 1 & \texttt{eyediff}_n \end{pmatrix}$$

và

$$\mathbf{W} = \begin{pmatrix} \pi_1(1-\pi_1) & 0 & \dots & 0 \\ 0 & \pi_2(1-\pi_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \pi_n(1-\pi_n) \end{pmatrix}$$

Ví dụ: Human Face Recognition

Với điểm bắt đầu $\boldsymbol{\beta}^{(0)} = \left(\beta_0^{(0)}, \beta_1^{(0)}\right) = (-0.5, 1)$, quá trình lặp được biểu diễn bởi hình dưới đây:



Ví dụ: Human Face Recognition

Kết quả của 40 lần lặp

t	$\beta_0^{(t)}$	$\beta_1^{(t)}$	error	relative error
0	-0.5000	1.0000	–	–
1	1.6063	-11.8460	13.0175	11.6432
2	1.7535	-13.3465	1.5078	0.1261
3	1.7587	-13.3400	0.0537	0.0040
4	1.7587	-13.4000	6.7048×10^{-5}	4.9611×10^{-6}
5	1.7587	-13.4000	1.0302×10^{-10}	7.6228×10^{-12}
6	1.7587	-13.4000	3.6621×10^{-15}	2.7096×10^{-16}
⋮	⋮	⋮	⋮	⋮
18	1.7587	-13.4000	1.7763×10^{-15}	1.3144×10^{-16}
19	1.7587	-13.4000	1.7902×10^{-15}	1.3246×10^{-16}
⋮	⋮	⋮	⋮	⋮
40	1.7587	-13.4000	3.5804×10^{-15}	2.6492×10^{-16}

$\Rightarrow \widehat{\beta}_0 = 1.7587$, $\widehat{\beta}_1 = -13.4000$, sau 5 lần lặp.

Chỉ số AIC

Khoảng cách Kullback-Leibler

Khoảng cách Kullback-Leibler (*Kullback-Leibler distance*) là thước đo sự gần nhau giữa hai phân phối (thường là phân phối ước lượng và phân phối chính xác):

$$D\left(f_\theta, g\right) = \int \log\left(\frac{g(y)}{f(y;\theta)}\right) g(y)\mathrm{d}y = \mathbb{E}_Y\left\{\log\left(\frac{g(Y)}{f(Y;\theta)}\right)\right\},$$

trong đó,

- $g(y)$ là mật độ chính xác (không biết),
- $f(y;\theta)$ là 1 mật độ *ứng viên*.

trong đó, θ được ước lượng bởi ML:

$$\widehat{\theta} = \arg\max_{\theta\in\mathbb{R}} \sum_{i=1}^{n} \log\left(f(Y_i;\theta)\right).$$

Rõ ràng, ta có thể có nhiều mô hình ứng viên $f(y;\theta)$, và mô hình tốt nhất là mô hình cho $f(y;\theta)$ gần $g(y)$ nhất, tức là cực tiểu hóa $D\left(f_\theta, g\right)$.

Chỉ số AIC

Chỉ số AIC (Akaike Information Criterion[1]) là một chỉ số được dùng để so sánh hai mô hình hồi quy mà được xây đựng đựa trên log-likelihood.

Xét GLM, với $Y_i \sim \mathrm{EDM}(b(\theta_i, a(\phi)))$ và

$$\mu_i = g^{-1}\left(\mathbf{X}_i^\top \beta\right).$$

Khi đó, từ mô hình, ta có 1 hàm mật độ cho Y_i là $f(Y_i; \beta)$ và hàm mật độ cho toàn bộ dữ liệu $Y_1, Y_2, \ldots, Y_n$,

$$f(\mathbf{Y}; \beta) = \prod_{i=1}^{n} f(Y_i; \beta)$$

Gọi $g(Y_i)$ là hàm mật độ chính xác của Y_i, và

$$g(\mathbf{Y}) = \prod_{i=1}^{n} g(Y_i)$$

[1] Hirotugu Akaike (1927 - 2009)

Chỉ số AIC

Khi đó,

$$\mathbb{E}_{\mathbf{Y}}\left\{\log\left(\frac{g(\mathbf{Y})}{f(\mathbf{Y};\beta)}\right)\right\} = \sum_{i=1}^{n}\mathbb{E}_{Y_i}\left\{\log\left(\frac{g(Y_i)}{f(Y_i;\beta)}\right)\right\}.$$

Dẫn tới, khoảng cách Kullback-Leibler là

$$\mathbf{D}\left(f_{\widehat{\beta}}, g\right) = nD\left(f_{\widehat{\beta}}, g\right),$$

Áp dụng khai triển Taylor, ta thu được

$$n\mathbb{E}_Y\left\{D\left(f_{\widehat{\beta}}, g\right)\right\} \doteq nD\left(f_{\beta_0}, g\right) + \frac{1}{2}\mathrm{tr}\left\{\mathcal{I}_g^{-1}(\beta_0)\mathcal{K}(\beta_0)\right\},$$

với

$$\mathcal{I}_g(\beta_0) = -n\int \frac{\partial^2 \log f(y;\beta_0)}{\partial\beta\partial\beta^\top} g(y)\mathrm{d}y,$$

và

$$\mathcal{K}(\beta_0) = -n\int \frac{\partial \log f(y;\beta_0)}{\partial\beta}\frac{\partial \log f(y;\beta_0)}{\partial\beta^\top} g(y)\mathrm{d}y,$$

Chỉ số AIC

Nếu mô hình là "tốt" $\Leftrightarrow \widehat{\beta} \approx \beta_0$ và $f \approx g$, khi đó $n\mathbb{E}_Y\left\{D\left(f_{\widehat{\beta}}, g\right)\right\}$ là nhỏ nhất.

Tuy nhiên, trong thực tế, ta không thể ước lượng được $n\mathbb{E}_Y\left\{D\left(f_{\widehat{\beta}}, g\right)\right\}$ do không có thông tin của hàm g.

Ta chứng minh được dạng biểu diễn $n\mathbb{E}_Y\left\{D\left(f_{\widehat{\beta}}, g\right)\right\}$ là

$$\begin{aligned} n\mathbb{E}_Y\left\{D\left(f_{\widehat{\beta}}, g\right)\right\} &\doteq \mathbb{E}_Y\left(-\ell(\widehat{\beta})\right) + \operatorname{tr}\left\{\mathcal{I}_g^{-1}(\beta_0)\mathcal{K}(\beta_0)\right\} \\ &\quad + n\int \log\left(g(y)\right) g(y)\mathrm{d}y. \end{aligned}$$

Từ đây, ta có 1 ước lượng cho $n\mathbb{E}_Y\left\{D\left(f_{\widehat{\beta}}, g\right)\right\}$ là

$$-\ell(\widehat{\beta}) + (p+1).$$

Dựa trên ước lượng này, Hirotugu Akaike định nghĩa chỉ số:

$$\mathrm{AIC} = -2\ell(\widehat{\beta}) + (p+1).$$

Chỉ số AIC

Chỉ số AIC của 1 mô hình là nhỏ nhất $\Rightarrow$ khoảng cách Kullback-Leibler nhỏ nhất $\Leftrightarrow$ mô hình đó là tốt nhất.

Áp dụng:

- So sánh mô hình A với mô hình rỗng: nếu AIC của mô hình A là lớn hơn AIC của mô hình rỗng $\Rightarrow$ mô hình A tệ hơn.
- So sánh mô hình A với mô hình B: nếu AIC của mô hình A là nhỏ hơn AIC của mô hình B $\Rightarrow$ mô hình A tốt hơn.

Chỉ số BIC

Chỉ số BIC (Bayes Information Criterion) là một chỉ số đánh giá mô hình được xây dựng dựa trên thống kê Bayes.

$$\mathrm{BIC} = -2\ell(\widehat{\beta}) + (p+1)\log(n).$$

Chỉ số BIC của 1 mô hình là nhỏ nhất $\Leftrightarrow$ mô hình đó là tốt nhất.

Áp dụng:

- So sánh mô hình A với mô hình rỗng: nếu BIC của mô hình A là lớn hơn BIC của mô hình rỗng $\Rightarrow$ mô hình A tệ hơn.
- So sánh mô hình A với mô hình B: nếu BIC của mô hình A là nhỏ hơn BIC của mô hình B $\Rightarrow$ mô hình A tốt hơn.

Hàm kernel

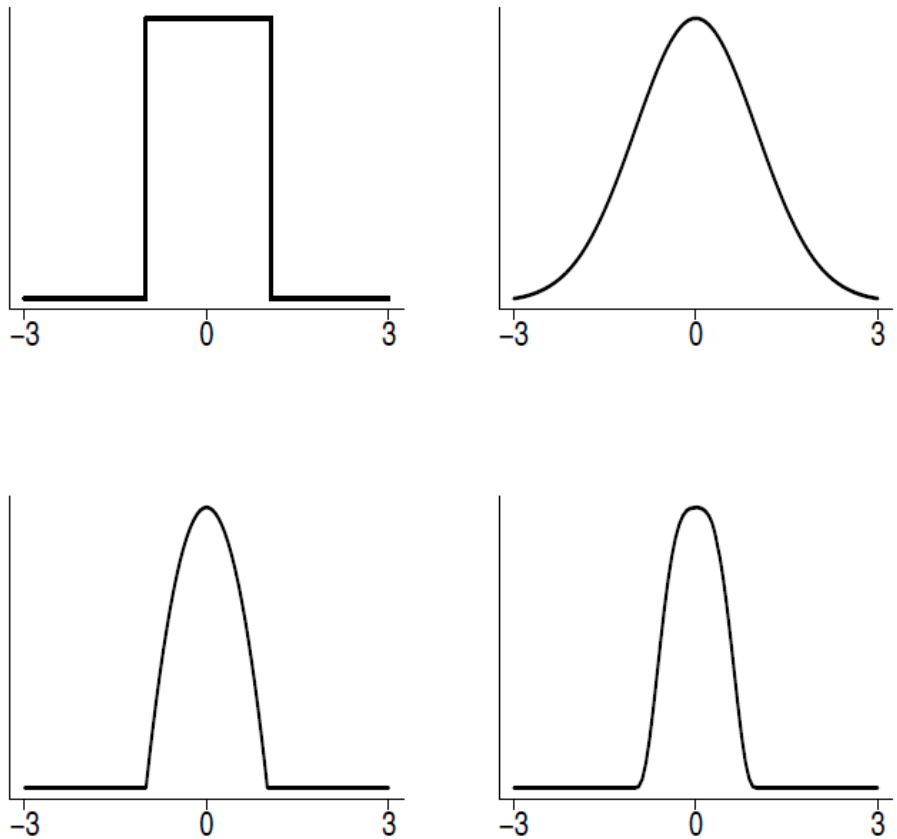
Hàm kernel $K(x)$ là một hàm thực dương, đối xứng, sao cho:

- $\int_{\mathbb{R}} K(x)\mathrm{d}x = 1;$
- $\int_{\mathbb{R}} xK(x)\mathrm{d}x = 0;$
- $\int_{\mathbb{R}} x^2K(x)\mathrm{d}x > 0.$

Một số hàm kernel thường được sử dụng:

- boxcar kernel: $K(u) = \frac{1}{2}\mathrm{I}(|x| \leq 1);$
- Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-0.5u^2);$
- Epanechnikov kernel: $K(u) = 0.75\left(1 - u^2\right)$ nếu $|u| \leq 1;$
- tricube kernel: $K(u) = \frac{70}{81}\left(1 - |u|^3\right)^3 \mathrm{I}(|x| \leq 1).$

Hàm kernel



Ước lượng hàm mật độ xác suất

Để ước lượng hàm mật độ xác suất, ta có một kỹ thuật được gọi tên là **kernel density estimation**:

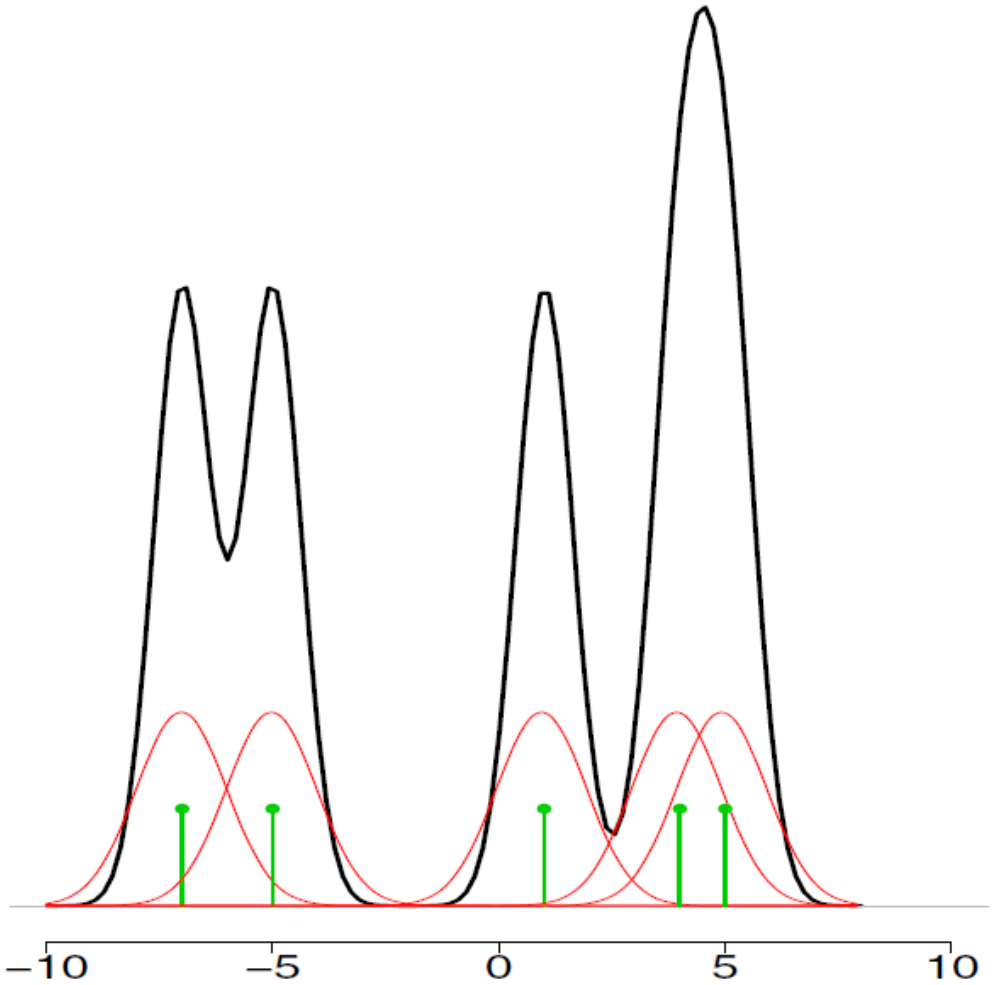
$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right),$$

trong đó,

- $h > 0$ là bandwidth, được dùng để hiệu chỉnh độ trơn của hàm mật độ ước lượng;
- $K(\cdot)$ là hàm kernel.

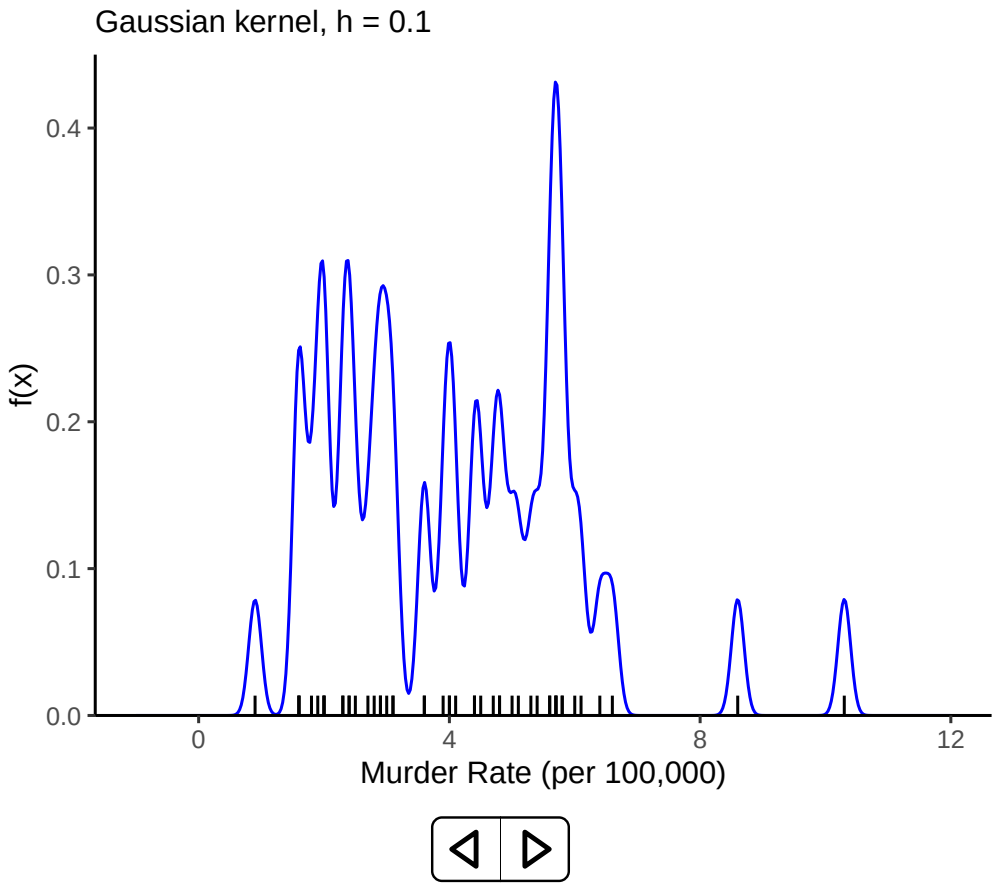
Xem thêm về ước lượng kernel trong sách Wasserman (2006).

Ước lượng hàm mật độ xác suất



Với mỗi một điểm x, giá trị $\widehat{f}(x)$ được xác định bằng trung bình giá trị hàm kernel tính tại các điểm dữ liệu lân cận.

Ước lượng hàm mật độ xác suất



Ước lượng hàm mật độ xác suất

Nhận xét:

- h càng nhỏ thì hàm mật độ ước lượng càng bị phân mảnh (quá chi tiết) - undersmooth;
- h càng lớn thì hàm mật độ ước lượng càng trơn - oversmooth.

Tìm h "tối ưu" là vấn đề then chốt trong ước lượng hàm mật độ.

Trong thống kê, có nhiều cách tìm h "tối ưu":

- rule-of-thumb
- unbiased cross validation
- biased cross validation
- Sheather & Jones method

Ước lượng hàm mật độ xác suất

Trong đó, phương pháp rule-of-thumb được sử dụng rộng rãi nhất:

$$h_{opt} = C \times \min\{s, IQR/1.34\} \times n^{-0.2},$$

với $C = 0.9$ hoặc 1.06.

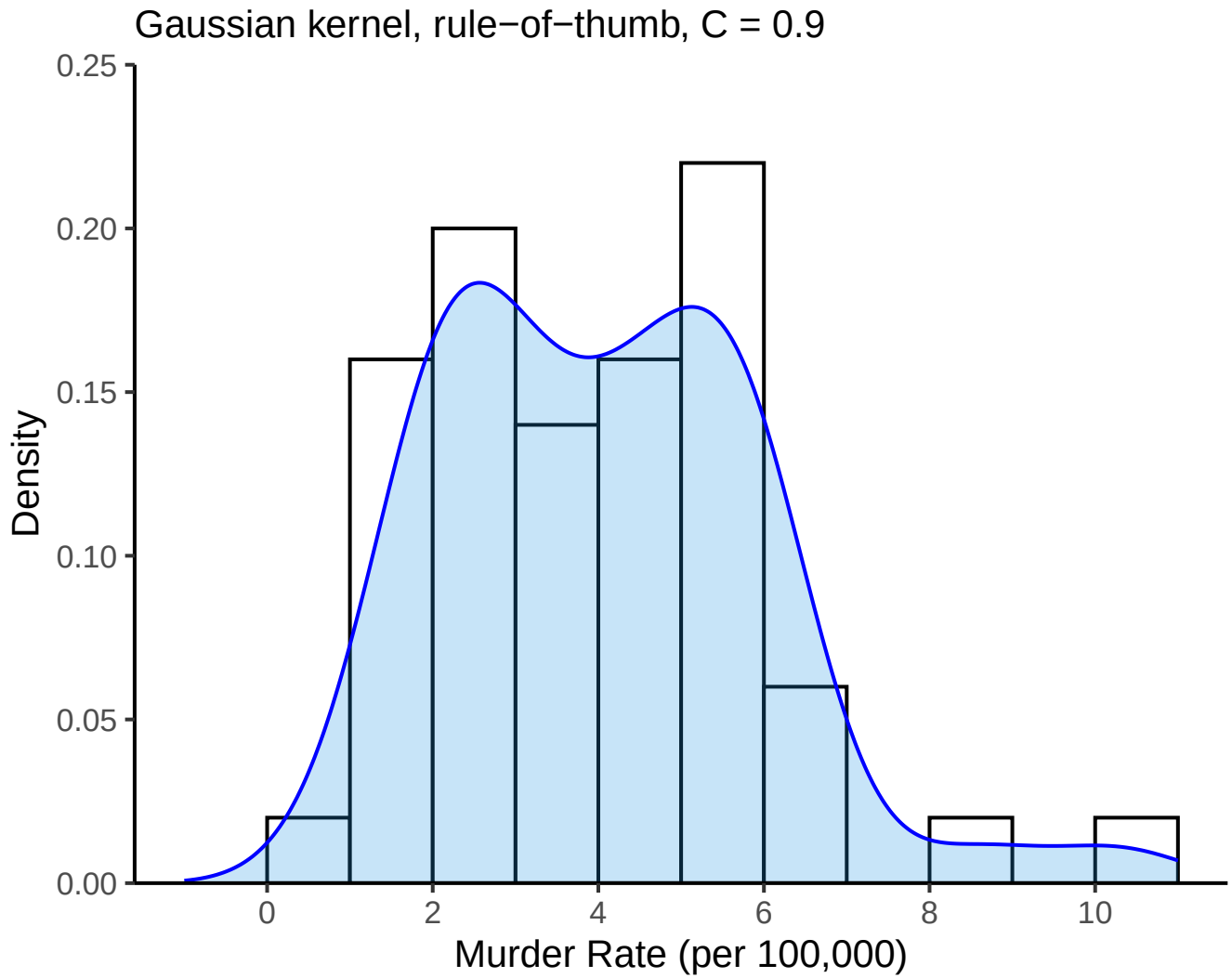
Thông thường, cách chọn h này sẽ đi kèm với Gaussian kernel.

Phương pháp cross-validation xác định h bởi việc cực tiểu hóa hàm sai số:

$$\hat{h} = \arg\min_h \int \widehat{f}^2(x)\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(x),$$

trong đó, $\widehat{f}_{-i}(x)$ là ước lượng kernel xác định trên dữ liệu bỏ đi điểm i.

Ước lượng hàm mật độ xác suất



Phân loại Bayes - Bayes classifier

Trong bộ dữ liệu, ta có:

- biến phản hồi $Y = j$ với $j = 1, \ldots, K$ (đại diện cho K nhóm cần phân loại);
- $X_1, X_2, \ldots, X_p$ là p biến giải thích, độc lập nhau, cùng có mối liên hệ với biến phản hồi Y.

Xác suất phân một đối tượng vào nhóm $Y = j$:

$$p_j(\boldsymbol{x}) = \Pr(Y = j | X_1 = x_1, \ldots, X_p = x_p) = \frac{f_j(\boldsymbol{x})\pi_j}{\sum_{i=1}^{k} f_i(\boldsymbol{x})\pi_i}$$

với $\boldsymbol{x} = (x_1, \ldots, x_p)$. Ta gọi

- $f_j(\boldsymbol{x})$ là mật độ xác suất đồng thời của $X_1, \ldots, X_p$ trên nhóm j.
- $\pi_j = \Pr(Y = j)$ là xác suất tiên nghiệm (prior probability);
- $p(\boldsymbol{x}) = \Pr(Y = j | X_1 = x_1, \ldots, X_p = x_p)$ là xác suất hậu nghiệm (posterior probability).

Phân loại Bayes - Bayes classifier

Dựa vào dữ liệu, ta định nghĩa ước lượng $\widehat{p}(\boldsymbol{x})$

$$\widehat{p}_j(\boldsymbol{x}) = \frac{\widehat{f}_j(\boldsymbol{x})\widehat{\pi}_j}{\sum_{i=1}^{k} \widehat{f}_i(\boldsymbol{x})\widehat{\pi}_i},$$

với

- $\widehat{f}_j(\boldsymbol{x})$ là ước lượng mật độ xác suất đồng thời của $X_1, \ldots, X_p$ trên nhóm j.
- $\widehat{\pi}_j$ là ước lượng xác suất tiên nghiệm

$$\widehat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(Y_i = j).$$

Vấn đề: ước lượng $\widehat{f}_j(\boldsymbol{x}) = \widehat{f}_j(x_1, x_2, \ldots, x_p)$?

Hàm mật độ p-chiều (của p biến $X_1, \ldots, X_p$), tính toán phức tạp.

Phương pháp Naive Bayes

Giả sử $X_1, \ldots, X_p$ là độc lập nhau, trong nhóm j, khi đó hàm mật độ p-chiều

$$f_j(\boldsymbol{x}) = f_{j1}(x_1) \times f_{j2}(x_2) \times \ldots \times f_{jp}(x_p)$$

và do đó

$$\widehat{f}_j(\boldsymbol{x}) = \widehat{f}_{j1}(x_1) \times \widehat{f}_{j2}(x_2) \times \ldots \times \widehat{f}_{jp}(x_p).$$

$\hookrightarrow$ ta cần ước lượng từng hàm mật độ thành phần $f_{jl}(x_l)$ trên nhóm j.

Xác suất hậu nghiệm

$$\widehat{p}_{j,NB}(\boldsymbol{x}) = \frac{\widehat{f}_{j1}(x_1) \times \widehat{f}_{j2}(x_2) \times \ldots \times \widehat{f}_{jp}(x_p) \times \widehat{\pi}_j}{\sum_{i=1}^{k} \widehat{f}_{i1}(x_1) \times \widehat{f}_{i2}(x_2) \times \ldots \times \widehat{f}_{ip}(x_p) \times \widehat{\pi}_i}$$

Các phương pháp ước lượng

- MLE: $X_l | Y = j \sim \mathcal{N}(\mu_{jl}, \sigma^2_{jl})$ - phân phối chuẩn;
- ước lượng kernel;
- ước lượng tỷ lệ (khi X_l là biến định tính).

Phương pháp Naive Bayes

Dựa trên ước lượng xác suất hậu nghiệm $\widehat{p}_{j,NB}(\boldsymbol{x})$, ta có ước lượng nhóm như sau:

- $K = 2$

$$\widehat{Y}(\boldsymbol{x}) = \begin{cases} 1 & \text{nếu } \widehat{p}_{1,NB}(\boldsymbol{x}) > p_0, \\ 0 & \text{nếu } \widehat{p}_{1,NB}(\boldsymbol{x}) \leq p_0, \end{cases}$$

  chú ý, $\widehat{p}_{0,NB}(\boldsymbol{x}) = 1 - \widehat{p}_{1,NB}(\boldsymbol{x})$

- $K \geq 3$

$$\widehat{Y}(\boldsymbol{x}) = \underset{j=1,\ldots,K}{\arg\max}\ \widehat{p}_{j,NB}(\boldsymbol{x}),$$

  tức là nhóm j có xác suất hậu nghiệm lớn nhất. Cách làm này được gọi tên là phương pháp MAP (maximum a posterior).

Ví dụ phân loại tín dụng

Phân loại 1 người có thể bị phá sản tín dụng (`Default`) dựa trên thông tin

- dư nợ tín dụng (`Balance`) - X_1;
- thu nhập hàng tháng (`Income`) - X_2;

biến đáp ứng $Y = 0$ (không phá sản), $Y = 1$ (phá sản).

Áp dụng phân loại Bayes

$$\widehat{p}_{j,NB}(\boldsymbol{x}) = \frac{\widehat{f}_{j1}(x_1) \times \widehat{f}_{j2}(x_2) \times \widehat{\pi}_j}{\sum_{i=1}^{2} \widehat{f}_{i1}(x_1) \times \widehat{f}_{i2}(x_2) \times \widehat{\pi}_i},$$

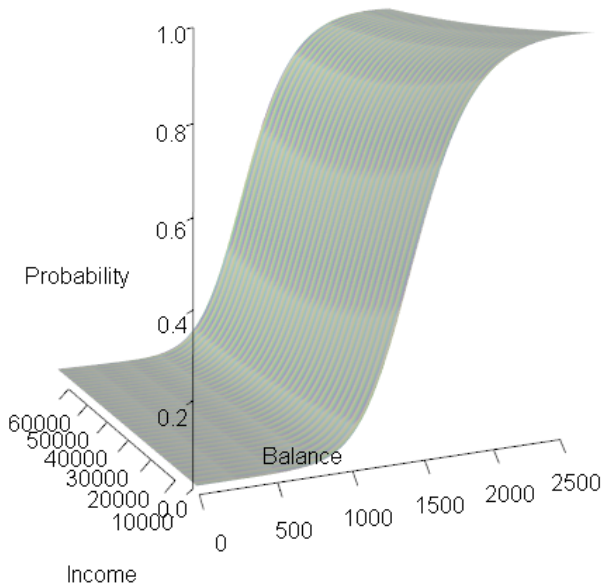
trong đó, $\widehat{f}_{jl}(x_l)$ là ước lượng hàm mật độ kernel:

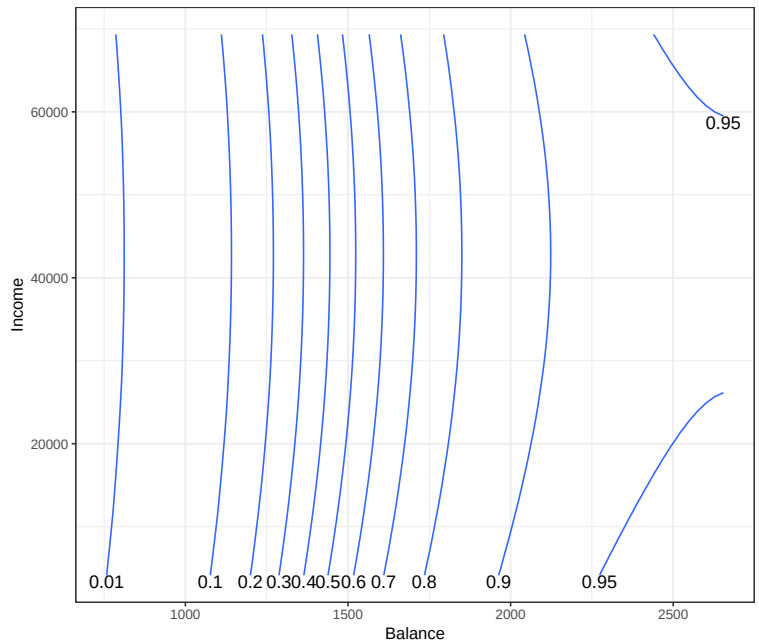
$$\widehat{f}_{jl}(x_l) = \frac{1}{n_{jl} h_{jl}} \sum_{i=1}^{n} K\left(\frac{x_{il} - x_l}{h_{jl}}\right).$$

Ví dụ phân loại tín dụng

Kết quả của quá trình ước lượng cho ta:

- mặt ước lượng $\widehat{p}_{1,NB}(x)$ (hình bên trái);
- đường đồng mức của $\widehat{p}_{1,NB}(x)$ (hình bên phải).

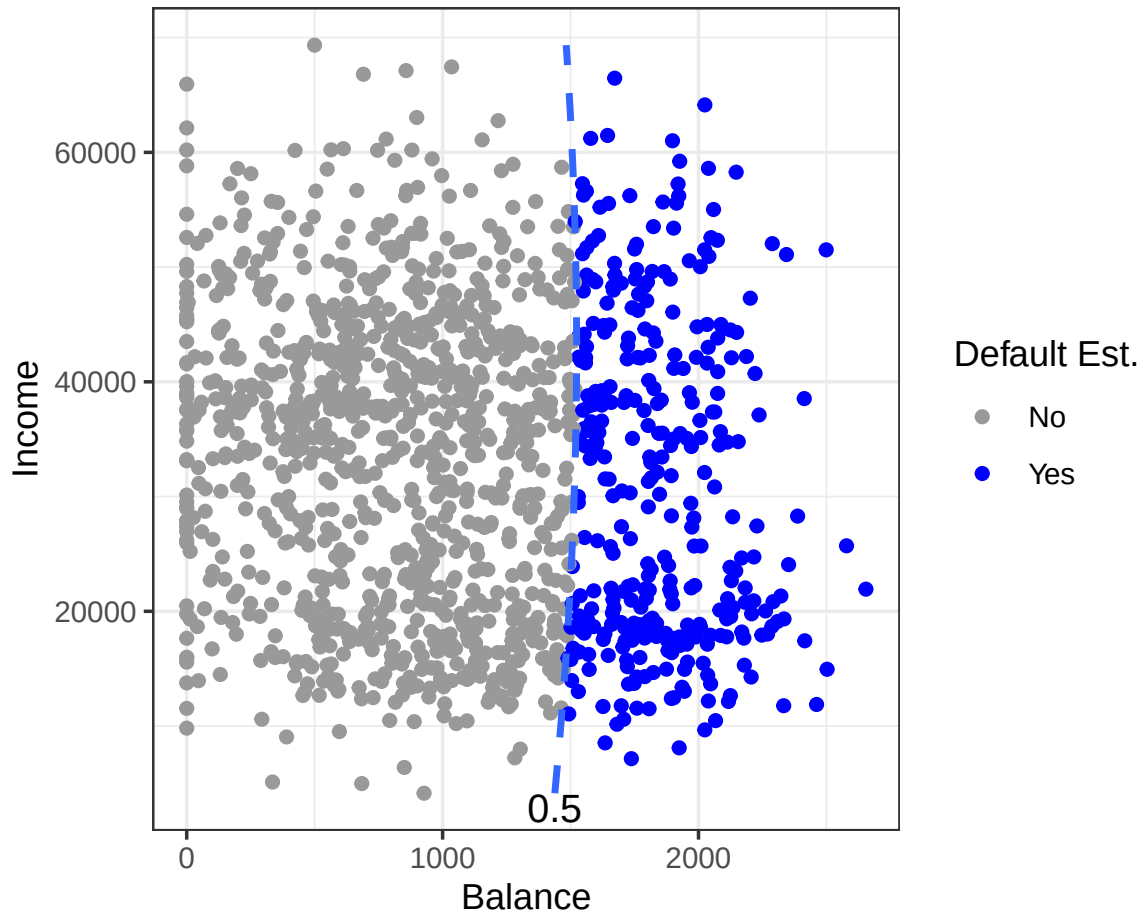




Các đường đồng mức này, gợi ý tới ngưỡng đưa ra quyết định.

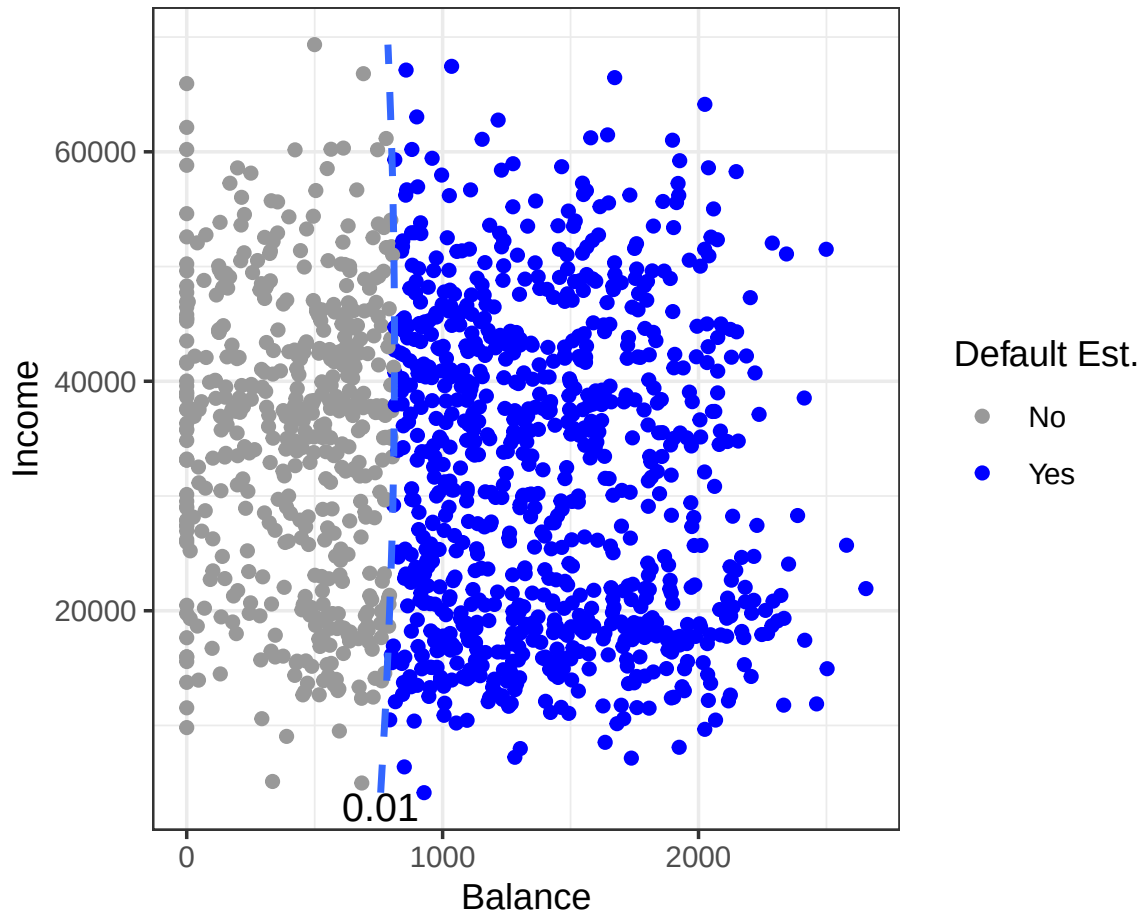
Ví dụ phân loại tín dụng

Ví dụ, chọn $p_0 = 0.5$



Ví dụ phân loại tín dụng

Thay đổi ngưỡng $p_0 \in (0, 1) \Rightarrow \widehat{Y}$ sẽ thay đổi



Ví dụ phân loại loài hoa iris

Phân loại 1 bông hoa iris vào 1 trong 3 loài:

- Versicolor ($j = 1$);
- Setosa ($j = 2$);
- Virginica ($j = 3$).

dựa trên thông tin

- chiều dài lá đài (`Sepal Length`) - X_1;
- chiều rộng lá đài (`Sepal Width`) - X_2.

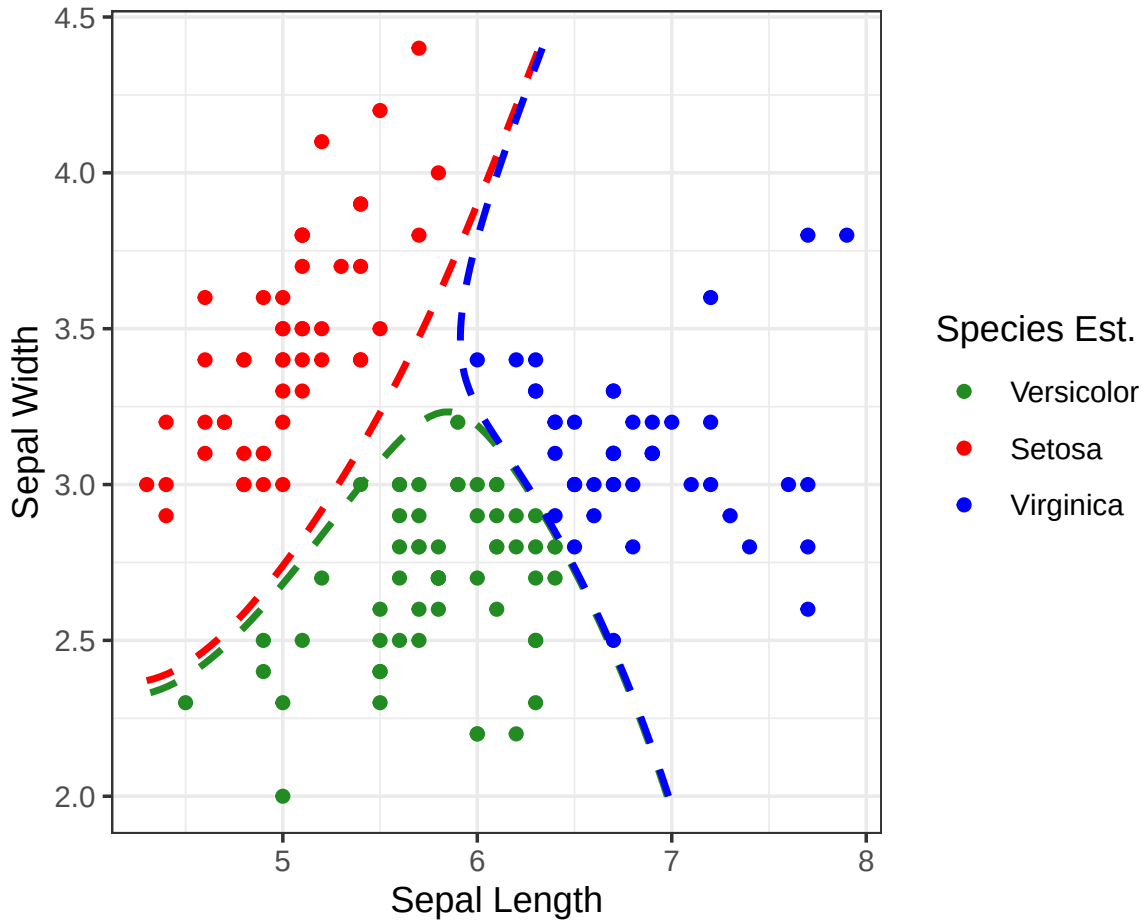
Áp dụng phân loại Bayes

$$\widehat{p}_{j,NB}(\boldsymbol{x}) = \frac{\widehat{f}_{j1}(x_1) \times \widehat{f}_{j2}(x_2) \times \widehat{\pi}_j}{\sum_{i=1}^{3} \widehat{f}_{i1}(x_1) \times \widehat{f}_{i2}(x_2) \times \widehat{\pi}_i},$$

trong đó, $\widehat{f}_{jl}(x_l)$ là ước lượng hàm mật độ của phân phối chuẩn $\mathcal{N}(\widehat{\mu}_{jl}, \widehat{\sigma}^2_{jl})$.

Ví dụ phân loại loài hoa iris

Áp dụng quy tắc MAP, ta có kết quả ước lượng nhóm được minh họa bởi hình



Main references I

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Davison, A. C. (2003). *Statistical Models*, volume 11. Cambridge University Press.

Dobson, A. J. and Barnett, A. G. (2018). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.