Bài giảng 2: Lý Thuyết Mô Hình Tuyến Tính Tổng Quát

TS. Tô Đức Khánh

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

Mô Hình Thống Kê Tuyến Tính Nâng Cao
–Cao học Khóa 33–

Mục lục

1 Ước lượng mô hình

2 Thống kê suy luận cho mô hình

3 Chuẩn đoán mô hình

1 Ước lượng mô hình

2 Thống kê suy luận cho mô hình

3 Chuẩn đoán mô hình

Thiết lập mô hình

GLM có ba thành phần chính:

- **biến phản hồi**: thành phần này chỉ định biến phản hồi Y và phân phối xác suất của nó. Bộ n quan sát $(y_1, \ldots, y_n)$ của Y được coi là quan sát thực tế của các biến ngẫu nhiên độc lập.
- **biến giải thích**: thành phần này chỉ định p biến giải thích cho một *linear predictor*, $\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}$, trong đó X_{ji} là giá trị của biến giải thích j cho quan sát thứ i.
- **hàm liên kết**: thành phần này là một hàm g được áp dụng cho kỳ vọng có điều kiện $\mu_i = \mathbb{E}(Y_i | X_{1i} = x_{1i}, \ldots, X_{pi} = x_{pi})$ của p biến phản hồi tại các giá trị biến giải thích $(x_{1i}, \ldots, x_{pi})$, liên hệ nó với *linear predictor*,

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} \equiv \eta_i.$$

Hàm $g(\cdot)$ là hàm đơn điệu, 1-1, khả vi, và do đó

$$\mu_i = g^{-1}(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}) \equiv g^{-1}(\eta_i).$$

Nếu $g(\mu_i) = \mu_i$, thì ta có hàm liên kết đồng nhất (*identity link function*).

Thiết lập mô hình

Họ phân phối mũ phân tán

Biến phản hồi Y trong GLM được giả định rằng có hàm phân phối thuộc vào **họ phân phối mũ phân tán** (**exponential-dispersion family**):

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

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.

Thiết lập mô hình

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, $\mathcal{B}(m,p)$ với m cố định và $p \in (0,1)$;
- phân phối multinomial, $\mathcal{M}(m; p_1, \ldots, p_k)$ với $p_i \in (0,1)$ 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 $\alpha, \beta > 0$.

Thiết lập mô hình

Một số hàm liên kết tương ứng với phân phối của biến phản hồi:

- 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.

Thiết lập mô hình

Một vài hàm liên kết thông dụng và hàm ngược của chúng:

Liên kết	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
log	$\log(\mu_i)$	$\exp(\eta_i)$
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
square-root	$\sqrt{\mu_i}$	η_i^2
logit	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$
probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
log-log	$-\log(-\log(\mu_i))$	$\exp(-\exp(-\eta_i))$
complementary log-log	$\log(-\log(1-\mu_i))$	$1-\exp(-\exp(\eta_i))$

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 $|t| < \delta$, $\delta > 0$:

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

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 của hàm log-likelihood là

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

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 $\dfrac{\partial \eta_i}{\partial \beta} = \dfrac{\partial \mathbf{X}_i^\top \beta}{\partial \beta} = \mathbf{X}_i$. 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, $W_i = \dfrac{1}{V(\mu_i)\left(g'(\mu_i)\right)^2}$ 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-Raphson 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 $\beta^{(0)}$

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.

Tiêu chuẩn hội tụ

Độ sai lệch tổng quát (Deviance)

Độ sai lệch tổng quát (*deviance*) giữa các giá trị quan sát $\mathbf{Y}$ và các giá trị trung bình $\boldsymbol{\mu}$, và được định nghĩa bởi

$$D(\mathbf{Y}, \boldsymbol{\mu}) = \sum_{i=1}^{n} d(Y_i, \mu_i),$$

với $d(Y_i, \mu_i)$ là độ sai lệch đơn vị (**unit deviance**) giữa một quan sát và trung bình μ_i.

Đối với một quan sát y của một biến $Y \sim \mathrm{EDM}(b(\theta), a(\phi))$, độ sai lệch đơn vị giữa y và $\mu = \mathbb{E}(Y)$ được định nghĩa bởi

$$d(y, \mu) = 2\left\{y\left[\theta(y) - \theta(\mu)\right] - \left[b(\theta(y)) - b(\theta(\mu))\right]\right\},$$

ở đây, $\theta(\mu)$ ký hiệu cho hàm ẩn của θ theo μ (ghi nhớ $\mu = b'(\theta)$).

Độ sai lệch đơn vị $d(y, \mu) = 0$ khi và chỉ khi $y = \mu$; và $d(y, \mu) > 0$ nếu $y > \mu$.

Tiêu chuẩn hội tụ

Ví dụ 1: Với phân phối chuẩn, $\theta = \mu = \mathbb{E}(Y)$, và $b(\theta) = \mu^2/2$, ta có độ sai lệch đơn vị là

$$d(y, \mu) = 2\left\{y\left[y - \mu\right] - \left[\frac{y^2}{2} - \frac{\mu^2}{2}\right]\right\} = (y - \mu)^2.$$

Ví dụ 2: Với phân phối Poisson, $\theta = \log(\mu)$ và $b(\theta) = \mu$, ta có độ sai lệch đơn vị là

$$d(y, \mu) = 2\left\{y \log\left(\frac{y}{\mu}\right) - (y - \mu)\right\},$$

với $y > 0$, và $d(0, \mu) = 2\mu$ (sử dụng giới hạn tại 0).

Tiêu chuẩn hội tụ

Bảng tổng hợp deviance đơn vị

Phân phối	deviance đơn vị
Normal	$(y-\mu)^2$
Binomial	$2\left\{y\log\left(\frac{y}{\mu}\right)+(1-y)\log\left(\frac{1-y}{1-\mu}\right)\right\}$
Negative Binomial	$2\left\{y\log\left(\frac{y}{\mu}\right)-(y+k)\log\left(\frac{y+k}{\mu+k}\right)\right\}$
Poisson	$2\left\{y\log\left(\frac{y}{\mu}\right)-(y-\mu)\right\}$
Gamma	$2\left\{-\log\left(\frac{y}{\mu}\right)+\frac{y-\mu}{\mu}\right\}$
Inverse Gaussian	$\frac{(y-\mu)^2}{\mu^2 y}$

Tiêu chuẩn hội tụ

Có nhiều tiêu chuẩn đánh giá khác nhau có thể áp dụng:

- so sánh sự khác biệt giữa giá trị lặp $\boldsymbol{\beta}^{(r+1)}$ và $\boldsymbol{\beta}^{(r)}$ với một số nhỏ $\delta > 0$, tức là, nếu
$$\left\|\boldsymbol{\beta}^{(r+1)} - \boldsymbol{\beta}^{(r)}\right\| \leq \delta,$$
trong đó, $\|\cdot\|$ là chuẩn Euclidean, $\delta = 10^{-6}$ hoặc có thể nhỏ hơn (nhưng không quá nhỏ);
- đánh giá độ sai lệch tổng quát (*deviance*) giữa quan sát $\mathbf{Y}$ và ước lượng trung bình $\boldsymbol{\mu}^{(r+1)}$ tại lần lặp thứ $r+1$:
$$\frac{\left|D(\mathbf{Y}, \boldsymbol{\mu}^{(r+1)}) - D(\mathbf{Y}, \boldsymbol{\mu}^{(r)})\right|}{|D(\mathbf{Y}, \boldsymbol{\mu}^{(r+1)})| + 0.1} \leq \delta.$$

Tiêu chuẩn đánh giá dựa trên sự sai lệch tổng quát được áp dụng trong các phần mềm thống kê.

Thuật toán lặp

Tổng kết lại, ước lượng ML, $\widehat{\beta}$, được tính bởi phương pháp lặp IWLS, với các bước tiến hành như sau:

1. Chọn một giá trị bắt đầu $\mu_i^{(0)}$ dựa vào Y_i tùy theo họ phân phối, tương ứng, tính hàm phương sai $V(\mu_i^{(0)})$ và đạo hàm $g'(\mu_i^{(0)})$.
2. Với $r = 0$, xác định

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

3. Đặt $r = r + 1$, và lặp lại bước 2 cho tới khi

$$\frac{\left|D(\mathbf{Y}, \boldsymbol{\mu}^{(r+1)}) - D(\mathbf{Y}, \boldsymbol{\mu}^{(r)})\right|}{|D(\mathbf{Y}, \boldsymbol{\mu}^{(r+1)})| + 0.1} \leq \delta,$$

   với δ có thể chọn là 10^{-8}, khi đó thuật toán hội tụ.
4. Nghiệm $\widehat{\beta}$ được xác định là giá trị lặp cuối cùng, tức là $\widehat{\beta} = \beta^{(r+1)}$.

Ước lượng hệ số phân tán

Nhắc lại rằng, đối với họ phân phối mũ phân tán, $a(\phi) = \phi/w$, với w là một trọng số đã biết.

Phân phối	ϕ	w
Normal	σ^2	1
Binomial	1	$\frac{1}{m}$
Negative Binomial	1	1
Poisson	1	1
Gamma	ϕ	1
Inverse Gaussian	ϕ	1

$\Rightarrow$ không cần thiết ước lượng ϕ trong các trường hợp Binomial, Negative Binomial và Poisson.

Ước lượng hệ số phân tán

Trong tổng quát, ϕ có thể được ước lượng bằng phương pháp ML:

$$U(\phi) = \frac{\mathrm{d}\,\ell}{\mathrm{d}\,\phi} = \sum_{i=1}^{n} \left\{ -\frac{Y_i\theta_i(\mu_i(\eta_i(\beta))) - b(\theta_i(\mu_i(\eta_i(\beta))))}{\phi^2/w_i} + \frac{\mathrm{d}\,c(Y_i,\phi)}{\mathrm{d}\,\phi} \right\},$$

với $c(Y_i, \phi)$ được xác định theo phân phối cụ thể.

Thay thế $\widehat{\beta}$ vào $U(\phi)$, ta có

$$U(\phi;\widehat{\beta}) = \frac{\mathrm{d}\,\ell}{\mathrm{d}\,\phi} = \sum_{i=1}^{n} \left\{ -\frac{Y_i\theta_i(\mu_i(\eta_i(\widehat{\beta}))) - b(\theta_i(\mu_i(\eta_i(\widehat{\beta}))))}{\phi^2/w_i} + \frac{\mathrm{d}\,c(Y_i,\phi)}{\mathrm{d}\,\phi} \right\}.$$

Giải phương trình $U(\phi;\widehat{\beta}) = 0$ bằng thuật toán IWLS, cho ra ước lượng $\widehat{\phi}_{\mathrm{ML}}$.

Tuy nhiên,

- ước lượng $\widehat{\phi}_{\mathrm{ML}}$ là chệch (biased);
- độ chệch lớn khi n nhỏ.

Ước lượng hệ số phân tán

Phương pháp moment thường được dùng để ước lượng ϕ.

Nhắc lại rằng, với $Y_i \sim \mathrm{EDM}(b(\theta_i), a(\phi))$, ta có $\mathbb{V}\mathrm{ar}(Y_i) = \dfrac{\phi}{w_i} V(\mu_i)$. Điều này dẫn tới

$$\phi = \mathbb{E}\left(w_i \frac{(Y_i - \mu_i)^2}{V(\mu_i)} \right).$$

Áp dụng phương pháp moment, ta có được ước lượng moment (*method-of-moment estimator*) cho ϕ là

$$\widehat{\phi}_{\mathrm{MM}} = \frac{1}{n} \sum_{i=1}^{n} \frac{w_i (Y_i - \mu_i)^2}{V(\mu_i)},$$

trong đó, $\mu_i = g^{-1}\left(\mathbf{X}_i^\top \beta\right)$ với giả định rằng hệ số hồi quy β là đã được biết.

Trong thực tế, ta chỉ có ước lượng $\widehat{\beta}$, do đó, khi thay thế $\widehat{\mu}_i = g^{-1}\left(\mathbf{X}_i^\top \widehat{\beta}\right)$ cho μ_i, ta có ước lượng cho ϕ là

$$\widehat{\phi} = \frac{1}{n - p - 1} \sum_{i=1}^{n} \frac{w_i (Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

Ví dụ

Ta xét dữ liệu từ nghiên cứu về bệnh bạch hầu (leukemia). Trong đó, ta quan tâm tới 2 biến:

- `REMISS`: có giá trị bằng 1, nếu có sự thuyên giảm bệnh bạch cầu xuất hiện ở một người bệnh, hoặc 0 nếu không xuất hiện;
- `LI`: chỉ số ghi nhãn tỷ lệ phần trăm tế bào bệnh bạch cầu tủy xương (percentage labeling index of the bone marrow leukemia cells).

Ta mong muốn xây dựng mô hình dự đoán xác suất thuyên giảm bệnh, dựa vào chỉ số `LI`.

Như vậy

- biến đáp ứng là `REMISS`,
- biến giải thích `LI`.

Bởi vì, `REMISS` là có dạng biến nhị phân, 0 và 1

$\Rightarrow$ phân phối phù hợp là phân phối nhị thức $\mathcal{B}(1, \mu)$, với μ là xác suất thành công (cũng là trung bình).

Ví dụ

Khi đó, $\texttt{REMISS}_i \sim \text{EDM}(b(\theta_i), 1)$, với

- $\theta_i = \log\left(\dfrac{\mu_i}{1-\mu_i}\right)$;
- $b(\theta_i) = -\log(1-\mu_i)$.

Thành phần linear predictor được xác định là

$$\eta_i = \beta_0 + \beta_1 \texttt{LI}_i.$$

Ta sử dụng hàm liên kết chuẩn tắc, tức là $\eta_i = g(\mu_i) = \theta_i = \log\left(\dfrac{\mu_i}{1-\mu_i}\right)$. Do đó, mô hình tuyến tính cần ước lượng sẽ là

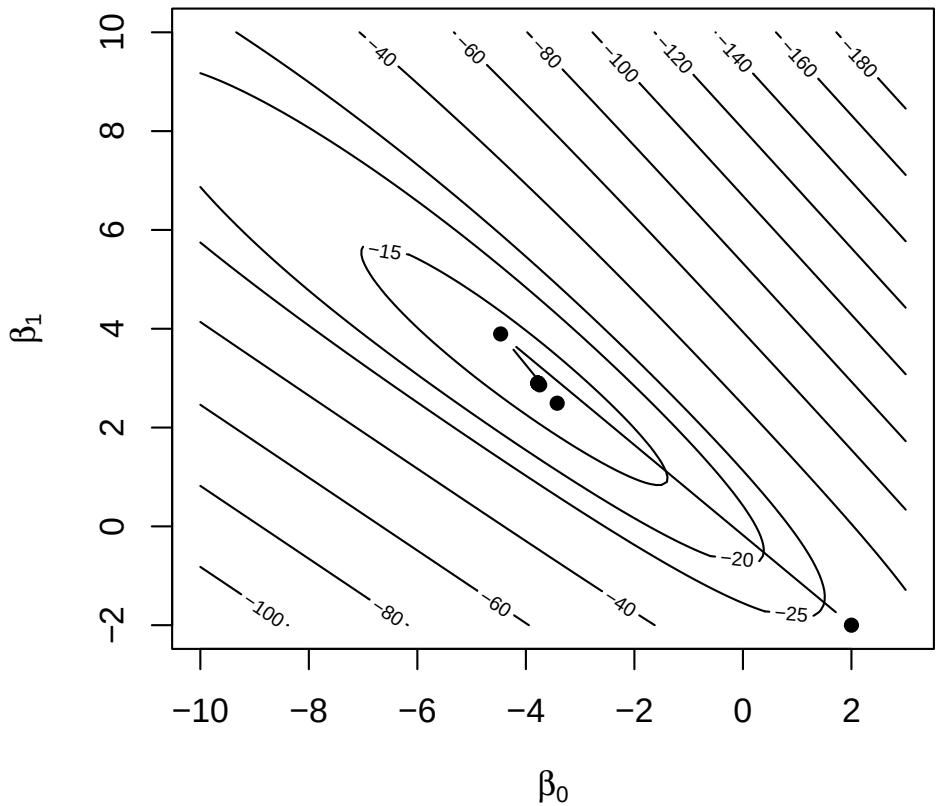
$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 \texttt{LI}_i.$$

Ta chọn giá trị bắt đầu

- $\beta_0^{(0)} = 2$;
- $\beta_1^{(0)} = -2$.

Ví dụ

Hình minh họa sự hội tụ của thuật toán IWLS, tương ứng với miền cực đại của hàm hợp lý của mô hình.



Ví dụ

Lần lặp r	$\beta_0^{(r)}$	$\beta_1^{(r)}$	Sai số $\boldsymbol{\beta}^{(r+1)}$	Sai số $D(\mathbf{Y}, \boldsymbol{\mu}^{(r+1)})$
0	2.00	−2.00		
1	−4.465	3.895	8.749	5.066×10^{-1}
2	−3.423	2.493	1.747	3.706×10^{-2}
3	−3.746	2.864	4.922×10^{-1}	5.897×10^{-3}
4	−3.777	2.897	4.520×10^{-2}	3.439×10^{-5}
5	−3.777	2.897	3.371×10^{-4}	1.767×10^{-9}
6	−3.777	2.897	1.799×10^{-8}	1.357×10^{-16}
7	−3.777	2.897	6.280×10^{-16}	1.357×10^{-16}
8	−3.777	2.897	4.441×10^{-16}	0.000
9	−3.777	2.897	4.441×10^{-16}	1.357×10^{-16}
10	−3.777	2.897	6.280×10^{-16}	1.357×10^{-16}
11	−3.777	2.897	4.441×10^{-16}	0.000

Kết quả ước lượng của mô hình là:

$$\log\left(\frac{\widehat{\mu}_i}{1-\widehat{\mu}_i}\right) = -3.777 + 2.897\,\mathtt{LI}_i.$$

Ước lượng hệ số phân tán $\widehat{\phi} = 0.957$, khá gần 1 (giá trị lý thuyết).

1 Ước lượng mô hình

2 Thống kê suy luận cho mô hình

3 Chuẩn đoán mô hình

Khoảng tin cậy cho hệ số - khi biết hệ số phân tán

Khoảng tin cậy cho hệ số - khi không biết hệ số phân tán

Khoảng tin cậy cho tiên đoán - khi biết hệ số phân tán

Khoảng tin cậy cho tiên đoán - khi không biết hệ số phân tán

Kiểm định giả thiết toàn cục

Kiểm định giả thiết từng hệ số

So sánh mô hình lồng nhau

Lựa chọn mô hình lồng nhau

1 Ước lượng mô hình

2 Thống kê suy luận cho mô hình

3 Chuẩn đoán mô hình

Chuẩn đoán mô hình