

### 1.3 Model estimation: Background:

Let

$$l(\beta) = \log[f(y, \theta, \phi)] = \log\{f[y, h(x\beta), \phi]\},$$

where  $\theta = h(x\beta)$ . Let

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \begin{bmatrix} \frac{\partial l(\beta)}{\partial \beta_1} \\ \frac{\partial l(\beta)}{\partial \beta_2} \\ \vdots \\ \frac{\partial l(\beta)}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} U_1(\beta) \\ U_2(\beta) \\ \vdots \\ U_p(\beta) \end{bmatrix}$$

and

$$\begin{aligned} A(\beta) &= -\frac{\partial^2 l(\beta)}{\partial \beta^t \partial \beta} = -\frac{\partial U(\beta)}{\partial \beta} \\ &= \begin{bmatrix} -\frac{\partial^2 l(\beta)}{\partial \beta_1^2} & -\frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_2} & \cdots & -\frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_p} \\ -\frac{\partial^2 l(\beta)}{\partial \beta_2 \partial \beta_1} & -\frac{\partial^2 l(\beta)}{\partial \beta_2^2} & \cdots & -\frac{\partial^2 l(\beta)}{\partial \beta_2 \partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_1} & -\frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_2} & \cdots & -\frac{\partial^2 l(\beta)}{\partial \beta_p^2} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}(\beta) & A_{12}(\beta) & \cdots & A_{1p}(\beta) \\ A_{21}(\beta) & A_{22}(\beta) & \cdots & A_{2p}(\beta) \\ \vdots & \vdots & \ddots & \vdots \\ A_{p1}(\beta) & A_{p2}(\beta) & \cdots & A_{pp}(\beta) \end{bmatrix}. \end{aligned}$$

#### Intuition:

If  $\hat{\beta}_{1 \times 1}$  is the maximum likelihood estimate (MLE), then

$$U(\hat{\beta}) = 0.$$

Further, by mean value theorem,

$$-U(\beta_0) = U(\hat{\beta}) - U(\beta_0) = \frac{\partial U(\beta^*)}{\partial \beta} (\hat{\beta} - \beta_0) = -A(\beta^*)(\hat{\beta} - \beta_0),$$

where  $\beta^* \in [\beta_0, \hat{\beta}]$ . Thus,

$$\begin{aligned} \hat{\beta} - \beta_0 &= A^{-1}(\beta^*) U(\beta_0) \\ \Leftrightarrow \hat{\beta} &= \beta_0 + A^{-1}(\beta^*) U(\beta_0). \end{aligned}$$

Motivated by the last equation, two algorithms can be used to obtain the maximum likelihood estimate  $\hat{\beta}$ .

Let  $\hat{\beta}_t = \begin{bmatrix} \hat{\beta}_{t1} \\ \hat{\beta}_{t2} \\ \vdots \\ \hat{\beta}_{tp} \end{bmatrix}$  and  $\hat{\beta}_{t+1} = \begin{bmatrix} \hat{\beta}_{(t+1)1} \\ \hat{\beta}_{(t+1)2} \\ \vdots \\ \hat{\beta}_{(t+1)p} \end{bmatrix}$  be the maximum likelihood estimate at the  $t - th$  and  $(t + 1)th$  iterations, respectively.

**1. Newton-Raphson method:**

$$\hat{\beta}_{t+1} = \hat{\beta}_t + A^{-1}(\hat{\beta}_t)U(\hat{\beta}_t), t = 0, 1, 2, \dots$$

or

$$A(\hat{\beta}_t)\hat{\beta}_{t+1} = A(\hat{\beta}_t)\hat{\beta}_t + U(\hat{\beta}_t), t = 0, 1, 2, \dots$$

**2. Fisher's scoring method:**

$$\hat{\beta}_{t+1} = \hat{\beta}_t + I^{-1}(\hat{\beta}_t)U(\hat{\beta}_t), t = 0, 1, 2, \dots$$

or

$$I(\hat{\beta}_t)\hat{\beta}_{t+1} = I(\hat{\beta}_t)\hat{\beta}_t + U(\hat{\beta}_t), t = 0, 1, 2, \dots$$

where

$$I(\beta) = \begin{bmatrix} I_{11}(\beta) & I_{12}(\beta) & \cdots & I_{1p}(\beta) \\ I_{21}(\beta) & I_{22}(\beta) & \cdots & I_{2p}(\beta) \\ \vdots & \vdots & \ddots & \vdots \\ I_{p1}(\beta) & I_{p2}(\beta) & \cdots & I_{pp}(\beta) \end{bmatrix} = E[A(\beta)] = -E\left[\frac{\partial^2 l(\beta)}{\partial \beta^t \partial \beta}\right].$$

The converge criteria are:

1.  $\|\hat{\beta}_{N+1} - \hat{\beta}_N\| < \varepsilon_1$ ,  $\varepsilon_1$  is some pre-specified small number.
2.  $\|U(\hat{\beta}_N)\| < \varepsilon_2$ ,  $\varepsilon_2$  is some pre-specified small number.

**Note:**

$U(\beta)$  is called the score function while  $I(\beta)$  is called the information matrix.