Lecture 5

§1 GLM 的 systematic component

1. **Definition:** GLM 的 systematic component

在 GLM 的 systematic component 中, linear predictor $\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ 与 mean μ_i 通过一个 link function $g(\cdot)$ 连接, 即

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

注: ① Systematic component 表明 GLM 为 linear in parameters

② link function $g(\cdot)$ 为一个 monotonic & differentiable function.

- 单调性用于保证一个 η 的值只与一个 μ 的值 map
- 可微性用于对 β 的 estimation.

e.g. 一系列常用的 link functions:

① Identity link: $\eta_i = g(\mu_i) = \mu_i$

② Logit link: $\eta_i = g(\mu_i) = \log \frac{\mu_i}{1-\mu_i}$

③ Probit link: $\eta_i = g(\mu_i) = \Phi^{-1}(\mu_i)$, 其中 Φ 为 $N(0,1)$ 的 CDF

④ Complementary log-log link: $\eta_i = g(\mu_i) = \log(-\log(1-\mu_i))$

⑤ Log link: $\eta_i = g(\mu_i) = \log(\mu_i)$

⑥ Inverse link: $\eta_i = g(\mu_i) = \frac{1}{\mu_i}$

2. **Definition:** Canonical link function (规范链接函数)

对于 GLM:

$$Y_i \sim f(y_i; \theta_i, \phi) = a(y_i, \phi) \exp\left\{ \frac{y_i\theta_i - \kappa(\theta_i)}{\phi} \right\}$$

记 $\mu_i = E[Y_i]$, η_i 为 linear predictor $\eta_i = x_i^T \beta$.

则 $g(\cdot)$ 被称为 canonical link function corresponding to the distribution of Y_i, 若

$$\eta_i = g(\mu_i) = \theta_i$$ (canonical parameter θ 恰好等于 linear predictor η)

注: ① 由于对于 EDM, 有 $\mu = E[Y_i] = \kappa'(\theta)$, 因此 canonical link 为

$$g(\cdot) = \kappa'^{-1}(\cdot)$$

② 对于非 canonical link 的情况, 有

$$\begin{cases} \mu = \kappa'(\theta) \\ g(\mu) = \eta \end{cases} \iff \mu = g^{-1}(\eta)$$ (对于 canonical link, $\kappa'(\cdot) = g'(\cdot)$, 故 $\theta = \eta$)

因此 $\theta = [\kappa'^{-1} \circ g^{-1}](\eta)$, 相较 canonical form 会增大计算量

例 1: 求 Normal distribution 的 canonical link function

将 distribution function 写成 EDM 的形式:

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

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\} \cdot \exp\left\{\frac{\mu y - \frac{1}{2}\mu^2}{\sigma^2}\right\}$$

其中，$\theta = \mu$

$\phi = \sigma^2$

$K(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2$

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

因此，$g(\mu) = \mu$ （由于 $\theta = \mu$，且 $E[y] = \mu$，为了使 $g(\mu) = \theta$，取 $g(\mu) = \mu$）

这个 μ 代表 $E[y]$

例2：求 Poisson distribution 的 canonical link function

将 distribution function 写成 EDM 的形式：

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!}$$

$$= \frac{1}{y!} \exp\{y \log\mu - \mu\}$$

其中，$\theta = \log\mu$ （$\mu = e^\theta$）

$\phi = 1$

$K(\theta) = \mu = e^\theta$

$a(y,\phi) = \frac{1}{y!}$

因此，$g(\mu) = \log\mu$ （由于 $\theta = \log\mu$，且 $E[y] = \mu$，为了使 $g(\mu) = \theta$，取 $g(\mu) = \log\mu$）

例3：求 Binomial distribution 的 canonical link function

将 distribution function 写成 EDM 的形式：

$$f(y;n,p) = \binom{n}{y} \cdot p^y (1-p)^{n-y}$$

$$= \binom{n}{y} \cdot \exp\left\{y \log\frac{p}{1-p} + n\log(1-p)\right\}$$

其中，$\theta = \log\frac{p}{1-p}$ （$p = \frac{e^\theta}{1+e^\theta}$）

$\phi = 1$

$K(\theta) = -n\log(1-p) = n\log(1+e^\theta)$

$a(y,\phi) = \binom{n}{y}$

因此，$\mu = K'(\theta) = \frac{ne^\theta}{1+e^\theta}$

$\Rightarrow e^\theta = \frac{\mu}{n-\mu}$

$\Rightarrow \theta = \log\frac{\mu}{n-\mu}$

因此，$g(\mu) = \log\frac{\mu}{n-\mu}$

注：① 若选用 canonical link function，则 $f(y;n,p)$ 可化为 $\binom{n}{y} \cdot \exp\{yX'\beta - n\log(1+e^{X'\beta})\}$

② 若选用 Probit link：$\eta_i = g(\mu_i) = \Phi^{-1}(\mu_i)$，则 $f(y;n,p)$ 可化为

$\binom{n}{y} \cdot \exp\{ y \log \frac{\Phi(X\beta)}{n-\Phi(X\beta)} + n \log(1-\frac{\Phi(X\beta)}{n}) \}$ (因为 $\eta = \Phi^{-1}(np) \Rightarrow p = \frac{\Phi(X\beta)}{n}$)

注：求 canonical link function 的方法：

① 将 distribution function 写成 EDM 的形式

② 找出 θ 关于 parameters 的表达式

κ 关于 parameters 的表达式

并利用前者将 κ 化为关于 θ 的表达式 $\kappa(\theta)$

③ 求出 $\mu = \kappa'(\theta)$

④ 求出反函数 $\theta = g(\mu) = \kappa'^{-1}(\mu)$

3. Definition: offset

Offset 为一个事先已知的 "structural" predictor. 其 coefficient 不是通过 model 估计的，而是默认值为 1.

注：大多数 GLMs 的 linear predictor 形式通常为 $\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$，但在某些情况下 (如 Poisson regressions for rate data)，需要引入 offset，形式变为 $\eta_i = o_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

offset o_i 可被视作一个值事先已知的 "predictor"，其 coefficient 值为 1.

e.g.

Example 4.2

Consider modelling the annual hospital birth rate in various cities to facilitate resource planning. To model the rate, we know

- the annual number of hospital births in each city, Y_i
- the population size of each city, P_i

Denote $\mu = E[Y]$ in general. We can model the number of births per unit of population, assuming a logarithmic link function, using the following systematic component

$$\log(\mu/P) = \eta,$$

for the linear predictor $\eta = X\beta$. If rearranging the model, we will have:

$$\log(\mu) = \log P + X\beta.$$

The first term in above systematic component $\log P$ is completely known: nothing needs to be estimated. The term $\log P$ is called an *offset*.

§2 GLM 的定义

1. GLM 的组成部分

Individual components of a generalized linear model (GLM) have been discussed. Here we formally define a GLM:

- Random component: The observations y_i come independently from a specified EDM such that $y_i \sim EDM(\mu_i, \phi)$ for $i = 1, 2, ..., n$.
- Systematic component:
  - A linear predictor $\eta_i = o_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ where the o_i are offsets that are often equal to zero,
  - and $g(\mu) = \eta$ is a known, monotonic, differentiable link function.

2. Definition: GLM

GLM 被定义为：

$$\begin{cases} y_i \sim EDM(\mu_i, \phi) \\ g(\mu_i) = o_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \end{cases}$$

其中，GLM 的 core structure 由 EDM distribution 和 link function 的选取决定，因此可被表示为 GLM (EDM ; Link function)

总结：

Structure of GLM

- Random component
  - $Y \sim EDMs$
  - MGF and CGF for EDMs
  - Variance function: $var(Y) = \phi V(\mu)$, $V(\mu)$ uniquely determines the distribution function within EDMs
  - Deviance and its asymptotic distribution: $D(y, \mu) = \sum_{i=1} d(y_i, \mu_i)$, $\frac{D(y,\mu)}{\phi} \sim \chi_n^2$ under exact saddlepoint approximation
- Systematic component
  - Linear predictor
    - offsets in Poisson-regression
  - link function and canonical links