Exponential Family

- Suppose $Y_1, \cdots, Y_n$ are independent random variables.
- Let $f(y_i; \theta_i, \phi)$ be PMF or PDF of Y_i, where ϕ is a scale parameter.
- If we can write

$$f(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right],$$

then we call the PMF or the PDF $f(y_i; \theta_i, \phi)$ is an exponential family.

Normal Distribution

Assume $Y_i \sim N(\mu_i, \sigma^2)$. Then, $E(Y_i) = \mu_i$ and σ is a scale parameter. The PDF is

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}$$

$$= \exp\{\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} + [-\frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2}]\}.$$

We may use $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $\phi = \sigma^2$, $a(\phi) = \phi$, $c(y_i, \phi) = -(1/2)\log(2\pi\phi) - y_i^2/(2\phi)$.

Binomial Distribution

Assume $Y_i \sim Bin(n_i, p_i)$. Then, $E(Y_i) = n_i p_i$. The PMF is

$$\binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

$$= \exp\{y_i \log \frac{p_i}{1-p_i} + n_i \log(1-p_i) - \log \binom{n_i}{y_i}\}.$$

Thus, $\theta_i = \log[p_i/(1-p_i)]$, $b(\theta_i) = n_i \log(1 + e^{\theta_i})$, $\phi = 1$, $a(\phi) = 1$, $c(y, \phi) = -\log \binom{n_i}{y_i}$.

Poisson Distribution

Assume $Y_i \sim Poisson(\lambda_i)$. Then, $E(Y_i) = \lambda_i$. The PMF is

$$\begin{aligned} & \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i} \\ = & \exp\{y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\}. \end{aligned}$$

Thus, $\theta_i = \log(\lambda_i)$, $b(\theta_i) = e^{\theta_i}$, $\phi = 1$, $a(\phi) = 1$, $c(y_i, \phi) = -\log(y_i!)$.

Gamma Distribution

Assume $x_i \sim \Gamma(\alpha, \beta_i)$, β_i is unknown. Then, $E(x_i) = \alpha/\beta_i$. Then PMF is

$$\frac{\beta_i^\alpha x_i^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta_i x_i} = \exp\{\alpha \log x_i + \alpha \log(\beta_i) - \log(\Gamma(\alpha)) - \log(x_i) - \beta_i x_i\}.$$

Assume α is known. If we choose $y_i = x_i$, then $\theta_i = -\beta_i$ ($\theta_i < 0$), $b(\theta_i) = -\alpha \log(-\theta_i)$, $\phi = 1$ and $a(\phi) = 1$.

Remark: We can also choose $y_i = -x_i$ and $\theta_i = \beta_i$. Then, $b(\theta_i) = -\alpha \log \theta_i$.

Negative Binomial Distribution

Assume $X_i \sim NB(k, p_i)$. The PDF is

$$\binom{x_i - 1}{k - 1} p_i^k (1 - p_i)^{x_i}$$

$$= \exp\{x_i \log(1 - p_i) + k \log \frac{p_i}{1 - p_i} - \log \binom{x_i - 1}{k - 1}\},$$

for $x_i = 0, 1, \cdots$. We choose $y_i = x_i$. Then, $\theta_i = \log(1 - p_i)$, $b(\theta_i) = -k \log[(1 - e^{\theta_i})/e^{\theta_i}]$, $\phi = 1$, $a(\phi) = 1$ and $c(y_i, \phi) = -\log \binom{y_i - 1}{k - 1}$. Then,

$$E(Y_i) = E(X_i) = b'(\theta_i) = \frac{k}{1 - e^{\theta_i}} = \frac{k}{p_i}$$

and

$$V(Y_i) = V(X_i) = b''(\theta_i) = \frac{k e^{\theta_i}}{(1 - e^{\theta_i})^2} = \frac{k(1 - p_i)}{p_i^2}.$$

GLM

The definition of Generalized Linear Model (GLM) is based on exponential family. There are three components in GLM. They are

- Random component. Assume the distributions of the sample. Such as normal, binomial, Poisson and etc.
- Systematic component. Describe the form of predictor (independent) variables. Such as

  $$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{ip} x_{ip}.$$

- Link function. Connect the unknown parameters to model. Such as

  $$g[\mu_i(\theta_i)] = \eta_i$$

  for some $g(\cdot)$, where $\mu_i(\theta_i) = E(y_i)$ is the expected value.

Canonical Link

If $\theta_i = \eta_i$ (or simply write $\theta = \eta$), then the canonical link is derived.

- Normal: identity link $g(\mu_i) = \mu_i$ or simply write $g(\mu) = \mu$ (same as below).
- Binomial: logistic link $g(\mu) = \log \frac{\mu}{1-\mu}$.
- Poisson: log link $g(\mu) = \log(\mu)$.
- Gamma: negative inverse link $g(\mu) = -1/\mu$.
- Negative binomial: $g(\mu) = \log[\mu/k(1 + \mu/k)]$.

- The most important cases are binomial and Poisson.
- Canonical link is just one of the link functions.
- Estimation is based on the maximum likelihood approach.
- Except the normal case, numerical computation is needed.

Link for Binomial

There are three link functions for binomial.

- Logistic link.

$$\log \frac{p_i}{1-p_i} = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$$

called logistic linear model or logistic regression.

- Inverse CDF link.

$$F^{-1}(p_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j.$$

If $F = \Phi$, it is the probit link, called probit model.

- Complementary loglog link.

$$\log[-\log(1-p_i)] = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j.$$

Logistic Regression

Consider the simplest case. That is

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i.$$

Suppose $\hat{\beta}$ and $\hat{\beta}_1$ are the MLEs.

- Odds ratio: as x increases a units, the estimate of odds ratio is $e^{a\hat{\beta}_1}$.
- The significance of the odds ratio can be directly read by the p-value of β_1.
- Confidence interval can also be derived respectively.

An example

Table 1: Blood Pressure and Heart Disease

Blood	Heart Disease	
Pressure	Present	Absent
< 117	3	153
117 – 126	17	235
127 – 136	12	272
137 – 146	16	255
147 – 156	12	127
157 – 166	8	77
167 – 186	16	83
> 186	8	35

Goodness of Fit

Let $\hat{n}_{ij}$ be the predicted counts of the model.

- Pearson χ^2 is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}.$$

- Loglikehood ratio χ^2 is

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log(n_{ij}/\hat{n}_{ij}).$$