# GLMs

Are a general class of model that allow us to model data with various kinds of response variables (within reason - see below)

For example, if we have a continuous response we might use regression.
If we have a 0/1 or successes/trials response, we might use logistic regression.
If we have count responses (like in a table) we might use log-linear analysis.

GLMs give us a mechanism to model many different response types within a general framework.
Note that each example above is a special case of a GLM.

As long as the response variable has a theoretical distribution that is a member of the exponential class of pdfs (or pmfs), the GLM framework can be applied.

[See lecture note handout]

The exponential class of pdfs has the general mathematical form:

$$f(y|\theta) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\phi}\right\} \cdot c(y, \phi)$$

$\theta$ usually turns out to be some function of the expected value of the response, $E(Y)$, and is called the canonical parameter.

$\phi$ is known as the scale, or dispersion, parameter.

$b(\theta)$ is a function of $\theta$ and is known as the cumulant generating function.

$c(y, \phi)$ is a function of $y$ and $\phi$ only (not of $\theta$)

If we are able to express the pdf or pmf of $Y$ in this form, then we can use the GLM framework to model this response.

Example: $Y \sim Bin(n, p)$, Assume $n$ is known.

$$f_Y = \binom{n}{y} p^y (1-p)^{n-y}$$

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

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

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

So we have that $\theta = \ln\left(\frac{p}{1-p}\right)$ $c(y, \phi) = \binom{n}{y}$

and $\phi = 1$. What about $b(\theta)$?

$b(\theta) = -n\ln(1-p)$ but this expression does not contain $\theta$... or does it?

$\theta = \ln\left(\frac{p}{1-p}\right) \quad \therefore e^{\theta} = \frac{p}{1-p} \Rightarrow p = \frac{e^{\theta}}{1+e^{\theta}}$.

$\therefore 1-p = \frac{1+e^{\theta}-e^{\theta}}{1+e^{\theta}} = \frac{1}{1+e^{\theta}}$.

$\therefore$ we have that $b(\theta) = -n\ln\left(\frac{1}{1+e^{\theta}}\right)$

or, using ln rules,

$b(\theta) = n\ln(1+e^{\theta})$.

$\therefore Y \sim Bin(n,p)$ can be written as an exponential distribution family member. $\therefore$ There is a GLM for Binomial responses.

Example: $Y \sim N(\mu, \sigma^2)$

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

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

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

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

So we have that $\theta = \mu$, $\phi = \sigma^2$,

$b(\theta) = \mu^2/2$ and $c(y, \phi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-y^2/2\sigma^2\}$
$= \theta^2/2$

Example: $Y \sim \text{Poisson}(\lambda)$

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

$$= \exp\{y \ln \lambda - \lambda - \ln(y!)\}$$

$$= \exp\{y \ln \lambda - e^{\ln \lambda}\} \cdot \frac{1}{y!}$$

$\therefore$ We have that $\theta = \ln \lambda$, $b(\theta) = e^\theta$, $\phi = 1$

and $c(y, \theta) = \frac{1}{y!}$.

## PROPERTIES OF EXPONENTIAL FAMILIES:

It can be shown mathematically that the following properties hold if a r.v. has an exponential family distribution:

The mean, $E(Y) = b'(\theta) = \frac{d\, b(\theta)}{d\theta}$

The variance, $Var(Y) = b''(\theta)\phi = \frac{d^2 b(\theta)}{d\theta^2}\phi$

We call $b''(\theta)$ the variance function, because in general the variance can depend on the mean.

Example: $Y \sim \text{bin}(m, p)$

$$b(\theta) = m \ln(1+e^{\theta})$$

$$\therefore E(Y) = b'(\theta) = m \frac{e^{\theta}}{1+e^{\theta}} \cdot 1 = \underline{mp}.$$

$$\text{Var}(Y) = b''(\theta)\phi = m\left[e^{\theta}(-1)(1+e^{\theta})^{-2}e^{\theta} + (1+e^{\theta})^{-1}e^{\theta}\right] \cdot 1$$

$$= m\left[\frac{e^{\theta}}{1+e^{\theta}} - \left(\frac{e^{\theta}}{1+e^{\theta}}\right)^2\right]$$

$$= m(p - p^2)$$

$$= \underline{mp(1-p)}.$$

Example: $Y \sim N(\mu, \sigma^2)$

$$b(\theta) = \frac{\theta^2}{2}$$

$$\therefore E(Y) = b'(\theta) = \theta = \underline{\mu}.$$

$$\text{Var}(Y) = b''(\theta)\cdot\phi = 1 \cdot \sigma^2 = \underline{\sigma^2}.$$

Example: $Y \sim \text{Pois}(\lambda)$

$$b(\theta) = e^{\theta}$$

$$\therefore E(Y) = b'(\theta) = e^{\theta} = \underline{\lambda}.$$

$$\text{Var}(Y) = b''(\theta)\cdot\phi = e^{\theta} \cdot 1 = \underline{\lambda}.$$

## LINK FUNCTION AND $\theta$, THE CANONICAL PARAMETER:

Every GLM follows the same rule: we must "link" the expected value of the response to the linear predictor via a link function, $g(\cdot)$.

ie:

$$g(\mu) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

We are free to choose any link function we wish (as long as it makes sense - see logistic regression notes).

However, there is a natural link function for each type of GLM (Normal, Binomial etc.) called the canonical link function. This is the link function that is described by the canonical parameter, $\theta$, when we write the dist$^n$ of $Y$ as an exponential family.

Example: $Y \sim Bin(n, p)$

$\theta = \ln\left(\frac{p}{1-p}\right)$ ← this is canonical link. (logit)

Example: $Y \sim N(\mu, \sigma^2)$

$\theta = \mu$ ← this is canonical link (identity)

Example: $Y \sim Poisson(\lambda)$

$\theta = \ln \lambda$ ← canonical link (log)

Each type of GLM has a canonical link function that can be found in this way, but there do exist other, non-canonical, links for each GLM too (eg: probit for binomial).

The help pages of SAS proc genmod lists a few acceptable link functions for each type of GLM.