

Generalized Linear Models

Sudipto Banerjee

sudiptob@biostat.umn.edu

University of Minnesota

Basics

- Generalized Linear Models (GLM) extend the theory of linear models to perform legitimate regression analysis to discrete outcomes.
- Assumption of normality is inappropriate; direct linear relationship between response and regressors is inappropriate.
- Modelling premise: data comes from an exponential *family*:

$$f(y | \theta, \phi) = \exp[(y\theta - b(\theta))/a(\phi) + c(y, \phi)],$$

where $a()$, $b()$ and $c()$ are monotonic. Also,

$$E[Y] = \mu = b'(\theta); \text{Var}(Y) = b''(\theta)a(\phi).$$

θ is called the *canonical parameter* and ϕ is a scale parameter.

- **Homework:** Show why a $N(\mu, \sigma^2)$ and a $Poi(\lambda)$ are special cases of the exponential family.
- Regression is carried out via a link function:

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}, \text{ or } \mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}).$$

Bernoulli or binomial outcomes

- Often response is in the binary scale. Let Y_i be coded 1 or 0 for $i = 1, \dots, n$. Let $\pi_i = P(Y_i = 1)$.
- Binary outcomes, when observed over n cases, yields the likelihood:

$$\prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}.$$

The assumption of a common $\pi_i = \pi$ across units results in a Binomial likelihood.

- How to specify link? It is assumed that $\pi_i = F(\mathbf{x}_i^T \boldsymbol{\beta})$ where $F(\cdot)$ is a distribution function. Thus, any distribution function can be taken as the inverse of the link function $g^{-1}(\cdot)$.
- Two popular versions are: normal cdf leading to probit link, and the logistic cdf $F(t) = e^t / (1 + e^t)$ that leads to the logistic link. Thus, regressors appear as:

$$\pi_i = \Phi^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}); \quad \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

- What do the regressors measure? In the former, they represent change in standard units of the normally distributed variable per unit change in x_i 's. In the latter, they represent change in the log-odds for a unit change in the x_i 's.