Generalized Linear Models

$i = 1, \dots, n \quad \{x_i, y_i\}$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$g(Y_i) = \beta_0 + \beta_1 x_i$$

Y_i = color $\qquad X_i$ = Wing span

$$g\left(Y_i = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix}\right) = \beta_0 + \beta_1 X_i$$

Y_i = species

$$g\left(Y_i = \begin{Bmatrix} \text{condor} \\ \text{duck} \\ \text{seagull} \end{Bmatrix}\right) = \beta_0 + \beta_1 X_i$$

Normal , Binary , Poisson

Learning Outcomes

- Exponential Family of Distributions
- Generalized Linear Models

Exponential Family of Distributions

Exponential Family of Distributions

An exponential family of distributions are random variables that allow their probability density function to have the following form: $y \sim F(\theta)$

$$f(y; \theta, \phi) = a(y, \phi) \exp\left\{ \frac{y\theta - \kappa(\theta)}{\phi} \right\}$$

θ : canonical function (parameter)

$\kappa(\theta)$: Cumulant log function

ϕ : dispersion parameter

$a(y; \phi)$: normalizing constant

Canonical Parameter

The canonical parameter represents the relationship between the random variable and the $E(Y) = \mu$

$$\eta = \beta_0 + \beta_1 X$$

$$\theta = \eta = g(\mu)$$

Normal Distribution

$$f_x = a(x;\phi)\, e^{\left\{\frac{x\theta - k(\theta)}{\phi}\right\}}$$

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-x^2}{2\sigma^2}}\, e^{\frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}}$$

$$f(x) = a(x;\phi)\, e^{\frac{x\theta - k(\theta)}{\phi}}$$

$$\frac{e^{\frac{-x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \quad e^{\frac{x\mu - \mu^2/2}{\sigma^2}}$$

$$a(x, \phi) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

$$\theta = \mu \qquad \phi = \sigma^2$$

$$k(\theta) = \mu^2/2$$

$$\eta = \mu$$

$$E(Y) = \beta_0 + \beta_1 X_1$$

Binomial Distribution

$$f(x) = a(x;\phi)\, e^{\left(\frac{x\theta - h(\theta)}{\phi}\right)}$$

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

$$y = e^{\ln y}$$

$$\binom{n}{x} \approx a(x;\phi)$$

$$\binom{n}{x} e^{\ln\left(p^x(1-p)^{n-x}\right)}$$

$$\binom{n}{x} e^{x\ln(p) + (n-x)\ln(1-p)}$$

$$f(x) = a(x;\phi)\, e^{\frac{\cancel{x}\theta - k(\theta)}{\phi}}$$

$$\binom{n}{x} e^{x \ln p + n \ln(1-p) - x \ln(1-p)}$$

$$\binom{n}{x} e^{x(\ln p - \ln(1-p)) + n \ln(1-p)}$$

$$\binom{n}{x} e^{\frac{x \ln\left(\frac{p}{1-p}\right) + n \ln(1-p)}{1}}$$

$$Y = \begin{matrix} 1 \\ 0 \end{matrix}$$

$$\theta = \ln\left(\frac{p}{1-p}\right) \qquad \phi = 1$$

$$K(\theta) = -n \ln(1-p)$$

$$a(x; \phi) = \binom{n}{x}$$

~~Binary~~
Binomial
Multinomial

$$\eta = \ln\left(\frac{p}{1-p}\right) \qquad E(Y=1) = \frac{e^{\eta}}{1+e^{\eta}}$$

logit

$P(x=1)$

success

$\ln\left(\frac{p}{1-p}\right) = \text{logit}$ $\eta = \beta_0 + \beta_1 X$

$$\frac{P(y=1)}{P(y=0)} = \frac{\text{success}}{\text{failure}} = \text{odds}$$

Logistic Regression

Binary/Binomial logit link function

Poisson Distribution

$$f(x) = a(x;\phi)\, e^{\frac{x\theta - k(\theta)}{\phi}}$$

$E(X) = \lambda$

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

$a(x;\phi) \approx \frac{1}{x!}$

$$\frac{1}{x!} e^{-\lambda} e^{\ln \lambda^x} = \frac{1}{x!} e^{-\lambda + \ln \lambda^x}$$

$$= \frac{1}{x!} e^{\frac{x \ln \lambda - \lambda}{1}}$$

$\theta = \ln(\lambda)$ $\quad \phi = 1$

$k(\theta) = \lambda$ $\quad a(x;\phi) = \frac{1}{x!}$

$Y \sim \text{Pois}$

$$\eta = \ln(\lambda)$$

$$e^{\eta} = E(Y)$$

Poisson Regression   log link function

Common Distributions and Canonical Parameters

Random Variable	Canonical Parameter
Normal	μ
Binomial	$\log\left(\frac{\mu}{1-\mu}\right)$
Negative Binomial	$\log\left(\frac{\mu}{\mu+k}\right)$
Poisson	$\log(\mu)$
Gamma	$-\frac{1}{\mu}$
Inverse Gaussian	$-\frac{1}{2\mu^2}$

Random Variable	Canonical Parameter

Generalized Linear Models

Generalized Linear Models

A generalized linear model (GLM) is used to model the association between an outcome variable (of any data type) and a set of predictor values. We estimate a set of regression coefficients $\boldsymbol{\beta}$ to explain how each predictor is related to the expected value of the outcome.

$$X^T\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$X^T\beta = g(E(Y))$$

Generalized Linear Models

A GLM is composed of a systematic and random component.

Systematic $X^T\beta$

random component Distribution

Random Component

The random component is the random variable that defines the randomness and variation of the outcome variable.

Systematic Component

The systematic component is the linear model that models the association between a set of predictors and the expected value of Y:

$$g(\mu) = \eta = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}$$