Crash course in Generalized Linear Models (GLMs)

Christopher L. Cahill

- Questions from last week
- This week's workshop is focused on extending the general linear model from last week to the generalized linear model (GLM)
- This will help us account for data where the response variable is non-normally distributed
- Vast scope of GLM in applied statistics
- Unify a large number of seemingly unrelated models and techniques
- Building blocks for more complex models we will talk about later
  - Many ecological models for inference about populations or communities can be viewed as a sequence of coupled GLMs

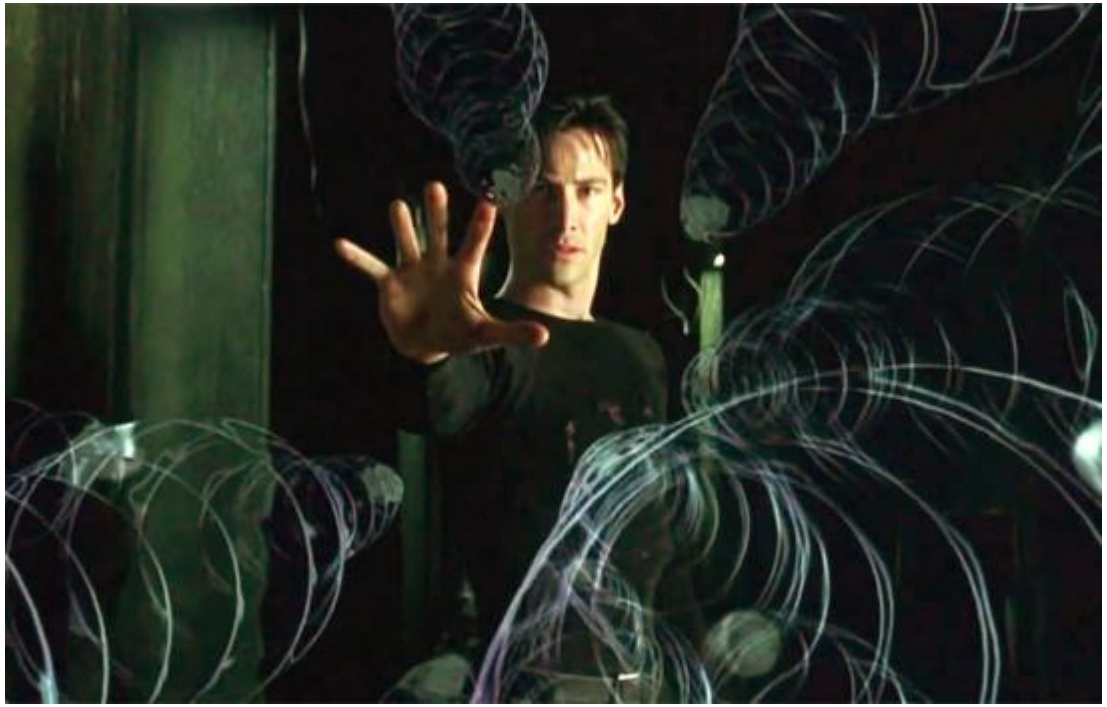
A refresher on the general linear model

$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \epsilon_i$ where $\epsilon_i \sim \mathrm{N}(0, \sigma)$

- Normal likelihood to account for random error
- Linear in the predictor or systematic component
- i.e., this is not a nonlinear model

- We can re-write this in terms of matrices for convenience:
- $\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 \\ . \\ \beta_p \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$



- *n* is number of data points, *p* is number of predictors

The general form of a GLM

1. The random part of the response (or the error structure of the model):

$$y_i \sim f(\theta)$$

2. A link function g, which we apply to the expected response $E(y) = \mu_i$ with η_i known as the *linear predictor*

$$g\big(E(y)\big) = g(\mu_i) = \eta_i$$

3. Systematic or linear part of the response (think of this like the mean prediction or structure of the model): the linear predictor (η_i), which contains a linear model:

$$\eta_i = \beta_0 + \beta_1 x_1$$

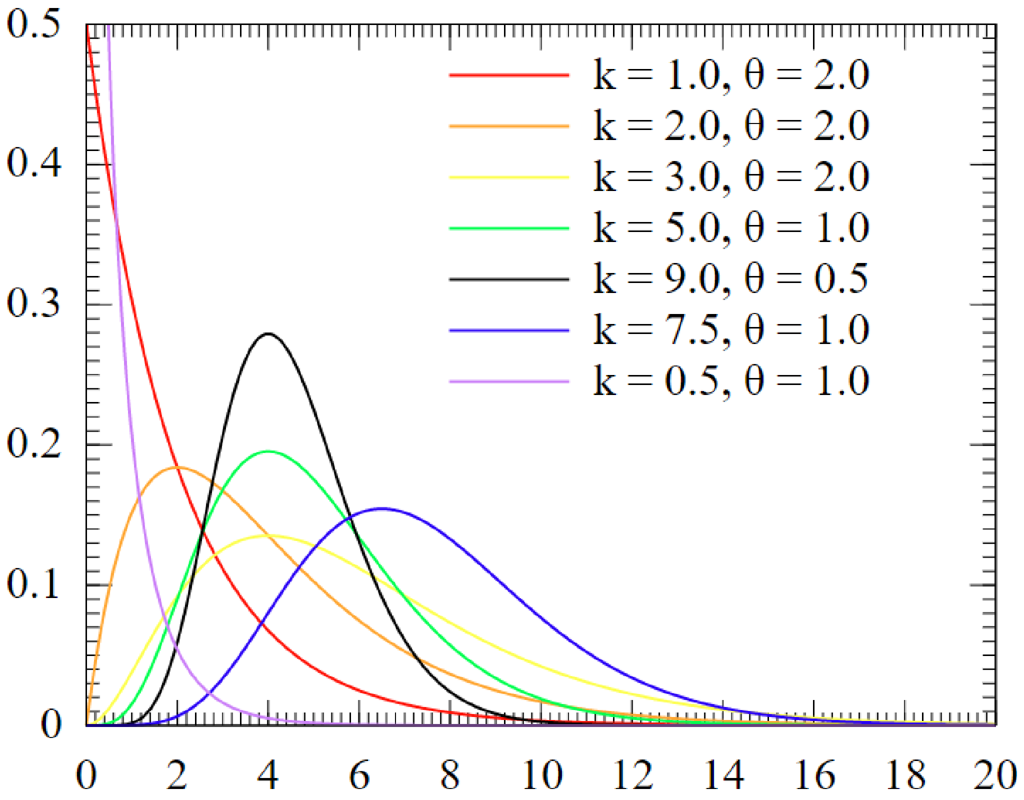
See section 3.3 in Royle and Kery 2008

Generalized Linear Models (GLMs)

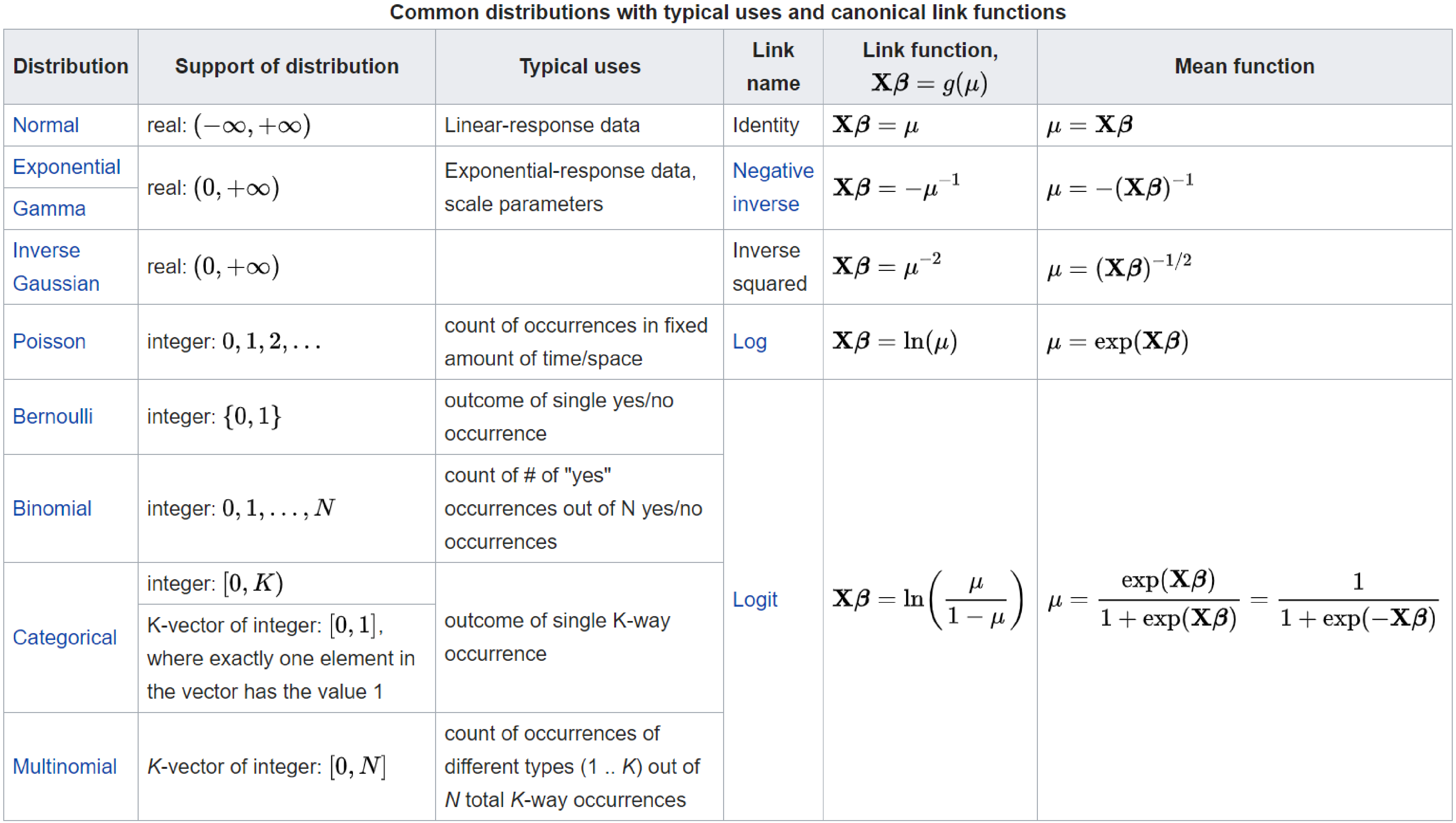
- This will help us account for data where the response variable is non-normally distributed
- The most common types of data that ecologists have are probably count and binary, so will emphasize these (Royle and Dorazio 2008)
- GLMs are defined for all members of statistical distributions belonging to the exponential family (Royle and Kery 2016)
- The GLM concept gives you considerable creative freedom in combining the three components of a GLM—but there are typically pairs of response distributions and link functions that go well together (called *canonical* links)
  - "Identity Link" = μ for the normal and sometimes gamma
  - "Log Link" = log(μ) for integers (think count data)
  - "Logit Link" = $\log(\frac{\mu}{1-\mu})$ for transforming [0,1] data to (-∞, +∞)
  - All of these can be inverted using algebra

Generalized Linear Models (GLMs)

- Common distributions
  - Bernoulli (0,1)
  - Binomial (0,1)
  - Beta (0,1)
  - Poisson (0, +∞), integers
  - Negative Binomial (0, +∞), integers
  - Gamma (0, +∞), continuous
  - Normal (-∞, +∞), continuous
  - Lognormal(0, +∞), continuous
  - t-distribution (-∞, +∞), continuous



Gamma distribution



Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences			
Categorical	integer: $[0, K)$	outcome of single K-way occurrence			
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

Some notes on overdispersion

- Neither the Poisson nor binomial have a dispersion parameter, which means that the magnitude of the variance is a fixed function of the mean.
    - Variance equations: Poisson = λ, binomial = $Np(1-p)$, Bernoulli = $p(1-p)$
- Often, we observe that the variance is greater than what should be expected from these equations (overdispersion)
    - For example, in the Poisson, a simple index of overdispersion is the variance/mean ratio, which should be ~1
- Overdispersion should be gauged after you fit your model—typically you compare the residual deviance to the residual degrees of freedom for Poisson models, and there are analogs for binomial response data
    - See McCullagh and Nelder 1989 or Zuur et al. 2009
- If over or underdispersion is not accounted for, cannot trust the model (point estimates and SEs)