Model Based Statistics in Biology.
Part V. The Generalized Linear Model.
Chapter 16.5 Notation and choice of probability model

on chalk board

ReCap (Ch 16) We extend the model based approach we have learned to non-normal errors.
GLM (normal errors) is a special case of GzLM
Count data are analyzed with discrete probability models
Continuous data bounded at zero are analyzed with a Gamma error
Continuous data bounded at 0 - 1 are analyzed with a Betabinomial error

Today: GzLM notation and choice of error structures and link functions.

Wrap-up.

The GzLM consists of a structural model, an error model, and a function that links the two.

The notation for the GzLM differs from the GLM. It is shown in 3 parts: the probability model, the link function, and the structural model η.

The distributional assumptions do not apply to the response variable. They apply to the residuals from the model.

Statistical tests of distributional assumptions reliably produce the wrong answer.

Each probability model has a canonical link function. We can choose links other than the canonical.

Normal errors can be used with count and zero bounded data if zeros and values close to zero are absent.

Notation

The generalized linear model is specified in three parts: a probability model, a structural model, and a function that links the two. Here is a list of the most commonly used models, in 3 part notation.

GLM – Normal error, identity link.

Mass is used as an example of a variable that is often distributed normally.

Distribution $Mass \sim Normal(\mu, \sigma)$

μ refers to the distribution of residuals around the fitted model

$\mu = 0$ for unbiased estimates.

σ refers to the standard deviation of the population.

$\hat{\sigma}$ is the estimate of the standard deviation,

s is a common symbol for the estimate of σ

Link $Mass = \mu$ This is the identity link

Structural Model $\eta = \sum \beta_i X_i$

X_i = Explanatory variable, $i = 0$ to n

β_i consists of one or more contrasts (categorical) or slopes (ratio scale regression variable).

Normal error, log link for exponential rates

$[C]$ = concentration of substance C.

$[C]$ is used as an example of a variable that often remains distributed normally when changing at an exponential rate.

Distribution $[C] \sim Normal(\mu, \sigma)$

$[C]$ = concentration of substance C.

Link $[C] = e^{\mu}$ This is the log link

The log link is used in preference to $\ln(C)$ when estimating an exponential rate, such as a clearance rate.

Structural Model $\eta = \sum \beta_i X_i$

Binomial response variable

Live is used as an example of a variable that is scored in two categories, *Live* or *Dead* in *N* organisms.

Distribution *Live* ~ *Binomial(N, π)*

$p = Live/N \quad 1\text{-}p = Dead/N$

$Odds = p / (1\text{-}p) = Live/Dead$

Link $Odds = e^{\eta}$ This is the logit link

The logit link is used to analyze multiplicative effects expressed as odds ratios.

Structural Model $\eta = \sum \beta_i X_i$

Poisson response variable

Count is used as an example of a variable that is scored as a count in a in a defined unit such as a quadrat.

Distribution *Count* ~ *Poisson(λ)*

$\lambda = Mean(Count) = \sum Counts / \sum Units$

$Variance(Count) = \lambda$

$Variance(Count) / Mean(Count) = 1$

Count data rarely meet this restriction.

Link $Count = e^{\eta}$ This is the log link

The log link is used to analyze multiplicative effects expressed as proportions.

Structural Model $\eta = \sum \beta_i X_i$

Quasipoisson response variable

Count is used as an example of a variable that is scored as a count in a in a defined unit such as a quadrat.

Distribution *Count* ~ *Poisson(λ,CD)*

$\lambda = Mean(Count) = \sum Counts / \sum Units$

$CD = Variance(Count) / Mean(Count)$

CD = Coefficient of Dispersion.

CD is estimated from the data.

Link $Count = e^{\eta}$ This is the log link

The log link is used to analyze multiplicative effects expressed as proportions.

Structural Model $\eta = \sum \beta_i X_i$

Negative binomial response variable

Infected is used as an example of a variable that is scored as a count resulting from a binomial process that varies according to the failure rate (1-*p*)

Distribution *Infected* ~ *NB(r, π)*

π = binomial proportion

r = shape parameter

variance(*Infected*) = $r \cdot \pi / (1-\pi)^2$

Link *Infected* = e^{η} This is the log link

Structural Model $\eta = \sum \beta_i X_i$

Gamma response variable

TreeAge is used as an example of a continuous variable bounded at zero with a right skewed distribution, such as a very low occurrence of very old trees.

Distribution *TreeAge* ~ *Gamma(k, θ)*

k = shape parameter

θ = scale parameter

mean(TreeAge) = $k\theta$

variance(*TreeAge*) = $k\theta^2$

Link *TreeAge* = $e^{1/\eta}$ This is the inverse link

Structural Model $\eta = \sum \beta_i X_i$

Beta binomial response variable

%DOC is used as an example of a continuous variable that is scaled from 0 to 1 against a fixed maximum.

DOC = Dissolved Organic Carbon.

DOC is scaled to the total carbon in a sample.

Distribution *%DOC* ~ *BetaBinom(μ, φ)*

μ = mean proportion

φ = dispersion

Link *%DOC* = e^{η} This is the log link

Structural Model $\eta = \sum \beta_i X_i$

Commentary.

Binomial Count Data. Intrinsic Hypotheses.

We usually do not have an extrinsic hypothesis, derived from a model external to the data at hand. In the absence of an extrinsic hypothesis we use an intrinsic hypothesis. Examples of intrinsic hypotheses are that the odds are the same across groups or that the odds do not change in relation to a regression variable.
In other words, the extrinsic hypothesis is that the odds ratio is OR = 1.

Poisson Count Data

For Poisson counts we use the Poisson error structure. We use the log link because we are usually interested in proportional changes.
For example, if we were analyzing counts of two species of fruitfly in relation to altitude, the model has the same structural model as
as in the analysis of heterozygosity in relation to altitude. However, the error structure is appropriate to the assumption that the response variable is a Poisson count.
We use the log link because we are usually interested in proportional changes: does the proportion change in altitude?

The generalized linear model allows us to use the same suite of structural models as the general linear model. Here, for example is the model for proportional change in counts of two species of fruitfly in two habitats.
As we will see in a later chapter, the interaction term test is equivalent to a two-way contingency test. It tests whether the count (as a proportion) depends on species.

If we are interested in absolute changes in counts, not proportions, we can use the identity link with Poisson errors

Negative Binomial Count Data.

Counts of organisms are often overdispersed: the variance is greater than the mean, and hence the Poisson error structure (variance = mean) is inappropriate. When this occurs the appropriate error structure is the negative binomial.

In general we use the log link, thus comparing counts as proportions. However, we can use the identity link if we are interested in absolute differences in counts.

Lognormal and Gamma Errors.

Count data is not the only source of heterogeneous errors. Often the variance increases with the mean, leading to heterogeneous errors, for ratio scale data that are not counts. For example, we may be interested in dispersal distances of fruitflies. This response variable may prove to lognormally distributed, with 'typical' value but with an occasional very large distance. In this circumstance the errors will often not

be normally distributed. The gamma error structure will often be appropriate, as it takes into account the increase in variance with increase in mean or expected values.

If we are interested in absolute rather than relative comparison across habitat and species we can use the identity link. The inverse link is often recommended on mathematical grounds (McCullagh and Nelder 1989).