Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 16.4 Non-normal errors – Continuous Data.

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

Today: Non-normal errors with continuous data.

Wrap-up.

Non-normal errors can arise in many ways.

Non-normal errors arise from count data, which is bounded at zero.

Binomial counts arise when each statistical unit is scored as yes/no, present/absent, etc.

Poisson counts arise from an unknown number of trials within a statistical unit. Poisson counts result from rare and random events. The variance will be approximately equal to the mean count per unit.

Overdispersed counts (variance > mean) arise in several ways. The arise from an unknown number of trials per unit. They arise from a small number of heterogeneous rates.

Continous data skewed by a low frequency of large values arise from a small number of multiplicative processes. Transforming skewed data produces biased estimates of parameters.

Data bounded at 0 and 1 are not infinite, unlike the normal distribution. These data arise when a partial coverage of a unit is measured on a ratio scale. An example is % area covered by water within a larger area.

Each error structure has a canonical link between the response variable and structural model. The canonical link can be replaced by other links.

Choosing an error structure.
The distribution of the response variable cannot be relied upon when choosing an error structure. A normally distributed response variable is no guarantee of normal and homogeneous residuals. A non-normally distributed response variable is no guarantee that residuals are non-normal. We can, however, diagnose the response variable for traits that lead to an informed choice. The traits we consider are counts versus continuous data, whether the data are bounded at zero, whether the data are bounded by a maximum, and whether the data occur near or at a zero or an upper boundary.

Ratio scale (continuous) data
For these data the first choice is a normal error. However, residuals may well prove to be heterogeneous where the response variable is skewed to large values, bounded at zero, or bounded at zero and one.

Proportions bounded at zero.
Here are two examples
Williams, C.B. (1964) Patterns in the Balance of Nature. Academic Press, London

Diversity indices
Diversity indices summarize, as a single number, the information obtained by sorting a collection into n species, each with a count of N_i organisms, from which we obtain $p_i = N_i/\Sigma N_i$ the proportion of organisms in each species. Here are three common indices.

n	Species richness
$H = \sum^{n} p_i \ln p_i$	The Shannon-Weaver index
$D = \sum \left(p_i\right)^2$	The Simpson index

For any one species the proportion p_i can be treated as a binomial variable. However, the aggregate measures H and D are effectively continuous and bounded at zero. Both measures have the potential for non-homogeneous errors around means and regression lines. This heterogeneity may not matter. If the residuals in an analysis of a diversity index are homogeneous and normal, a GLM (normal error) will suffice. If residuals are heterogeneous or non-normal, our next choice is a gamma error for the indices. This distribution is bounded at zero and has a parameter that takes into account skewness toward large values.

Proportions bounded at zero and one.
Here are two examples.

Percent cover of any surface

Percent time in an activity.

Prior to the introduction of PROC GENMOD by the SAS system in 1993 the recommended procedure for analysis of data bounded between zero and one was the arcsin transform. The arcsin addressed the problem in principle, if not in practice, by a kluge—transforming to the degrees of a circle. A kluge is a fix-up for want of something better. It is a kluge with no guarantee or even a record of success in producing normal residuals. The arcsin transform has persisted into the present century by what might be called academic inertia, the tendency of supervisors and reviewers to use techniques learned as grad students.
The implementation of software for analysis of 0-1bounded continuous data has lagged considerably behind software for binomial proportions. Beta regression (beta error model) was introduced by SAS in PROC GLIMMIX circa 2010. It is available in R but not SPSS (2018). The topic is rarely treated in recent texts on generalized linear models. Examples are primarily from the social science literature, despite its applicability in the natural sciences.

Choosing the link
The choice of an error structure brings with it a ***canonical link*** between the response variable and structural model. Here is a listing of the canonical links for commonly used error structures.

Other links are possible, for any choice of error structure. For example, the canonical link for a gamma error structure is the inverse link. This is appropriate for a hyperbolic relation of response to explanatory variables, such as a Michaelis-Menten dynamics. The inverse link, is not readily interpretable in most applications. An alternative is a gammae error with identity link, which is readily interpretable. Estimate with an identity link can, however, fail. The next choice is a log link, resulting in an interpretable relation of multiplicative effects.