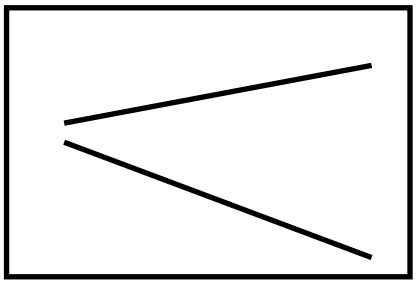
Göran Arnqvist

GENERALIZED LINEAR MODELS

- Most statistical evaluation is based on **data** and some form of a statistical **model**.
- By specifying a model you typically try to fit a certain relationship between variables – such as in a two way anova/ancova with an interaction:



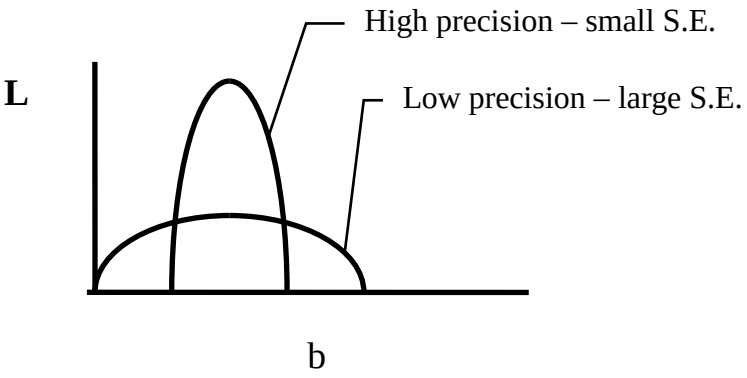
- An algorith is then used to "parameterize" the structure you impose; to find the values of the components of the model that best descibes/fit your data.

As mentioned before this can be done in several ways. In conventional **general linear models** (anova, ancova, conventional regression) this is done analytically by a method known as "**least-squares**". Minimizes the error (sums of squares of residuals). However, least-squares methods are sometimes problematic.

- Fairly **rigid**: makes relatively many assumptions, several of which are often not upheld:
  1) The response variable continuous
  2) The residuals (errors) normally distributed
  3) Homogeneity of variance
- Can yield **biassed estimates**, unless assumptions are completely upheld.

- **Analytically constrained**: inference difficult or impossible with some type data (e.g. unbalanced designs or unbalanced data; empty cells in certain anova/manova/ancova models).

An alternative method is **maximum likelihood (ML)** estimation. Estimates the empirical support for a certain model (the likelihood ≈ the probability of attaining the observed data under the model), then modifyes the model and estimates again; until the "best" possible model is found ---> an *iterative* procedure (can actually involve least-squares weighting). The model that maximizes the likelihood is deemed the "best" model. Precision of estimates, for example for the slope (*b*) in a regression, is determined from the "peakedness" of the likelihood function:



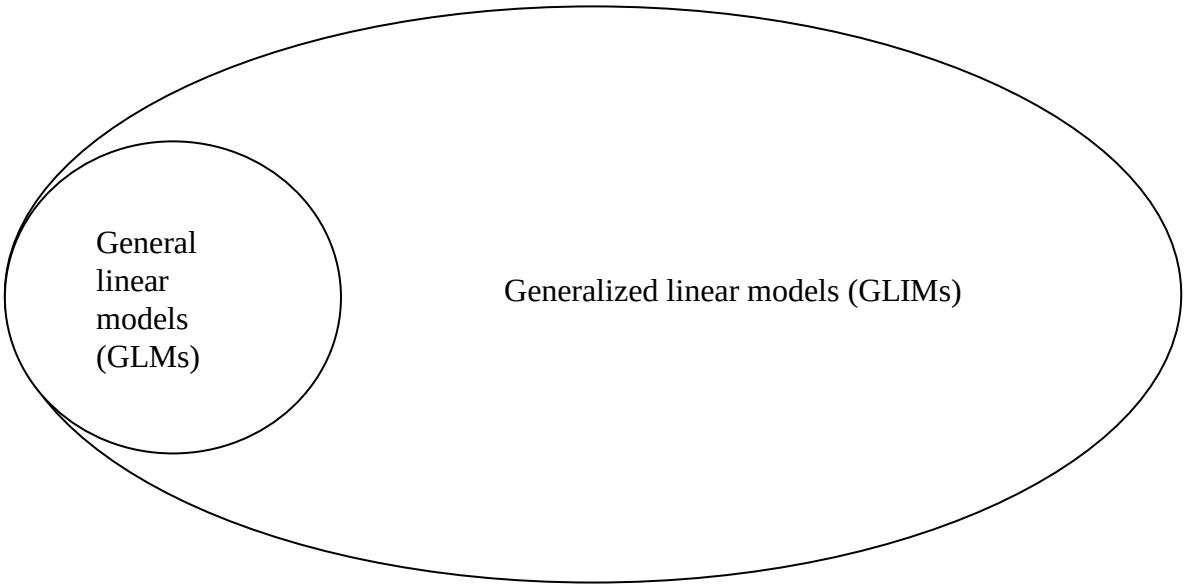
Maximum likelihood estimation has several important advantages:

- **Flexible**: a more general theorem that involves much fewer inherent assumptions - can be used to estimate almost any type of models and any type of data, by specifying the structure of the model.
- Can **handle "difficult"** data in a better way (e.g. unbalanced data, empty cells).
- Yields **unbiassed estimates** under a wider range of circumstances, including conventional general linear models as special cases, of course.

For mixed models, a third model fitting strategy is often used; **restricted (or residual) maximum likelihood (REML)**. REML produces relatively unbiassed estimates of the random effects factors when data is unbalanced - often used for variance and covariance component estimation.

A fourth main method of fitting models to data is **Bayesian inference, using Markov chain Monte Carlo simulations** (a form of iterative resampling of data) to fit models. Gaining in popularity, largely because of it's utility (e.g., efficient in estimating many parameters simultaneously), but is principally criticized by many statisticians – fear that it may sometimes yield biassed estimates.

Generalized linear models (GLIMs) refer to a class of very general and flexible statistical models that normally uses maximum likelihood to fit linear combinations of independent variables (and/or factors) to a response, or dependent, variable. **Any models can be estimated**; including multifactorial designs, nested models, random models, mixed models, repeated measures designs, etc etc. General linear models can be seen as a subset of generalized linear models:



Three basic concepts in generalized linear modelling:

1. **Error structure**
2. **The linear predictor**
3. **The link function**

1. Error structure

So far in this course, you have dealt with cases where the errors (residuals) are

1) Normally/Gaussian distributed (and Y is continuous).
2) Independent of the explanatory variables / factor levels (homogeneity of variance).

Errors are **frequently non-normal** (visualized by frequency plots of residuals), the response variable may not be continuous and errors are **sometimes dependent of the explanatory variables** (visualized by plots of residuals versus e.g. independent variables or predicted values). Generalized models handle this elegantly, by simply specifying error structure in the the model!

- Normal/Gaussian errors - for continuous response variables with normal errors.
- Poisson errors - for count data and multi-category data.
- Binomial errors - for proportions and dichotomous (binary) category data.
- Gamma errors - for data where variance in Y changes with X (heteroscedasticity - dependent errors).
- Many others for special cases; e.g. inverse Gaussian errors, negative binomial errors and quasi-likelihood models where variance functions are explicitly specifyed.
-

2. The linear predictor

The predicted value of the dependent variable from a linearly structured GLIM model is called the linear predictor (η - "eta") of Y, such that the model structure is:

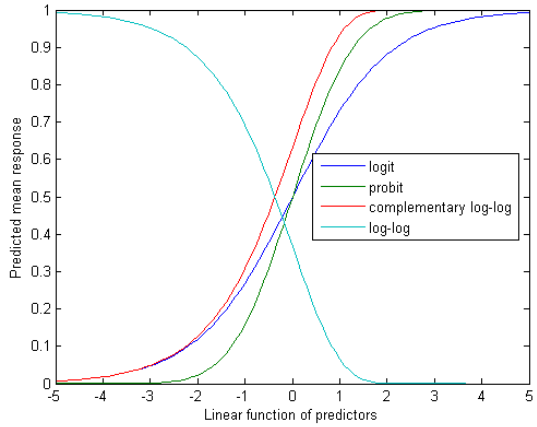
$$\eta = \beta_1 x_1 + \beta_2 x_2 + ... + \varepsilon$$

Note that we do not model Y but a linear predictor of Y!

3. The link function

The link function is a simple but brilliant trick and a very **central concept** in generalized linear models. It represents a function of the dependent variable that yields a linear function of the independent variables. In other words, it relates the mean value of Y (μ) to its linear predictor eta (η). Put very crudely, one can say that the link function forms the **mathematical link** between a "non-linearized" and a "linearized" function and it hence enables a "linear" modelling approach.

Some common link functions are:



- Identity link; η = μ (if combined with Gaussian error, it makes a generalized linear model a general linear model....)
- Log link; η = log μ
- Logit link; η = log μ/*n*-μ
- Reciprocal; η = 1/μ
- Others are e.g. the probit link, the complimentary log-log link, square root and exponent.

Families

The class of generalized linear models handled by facilities supplied in *S-PLUS* includes *gaussian*, *binomial*, *poisson*, *inverse gaussian* and *gamma* response distributions and also *quasi-likelihood* models where the response distribution is not explicitly specified. In the latter case the *variance function* must be specified as a function of the mean, but in other cases this function is implied by the response distribution.

Each response distribution admits a variety of link functions to connect the mean with the linear predictor. Those automatically available are as in Table 3

	Family Name					
Link Function	binomial	gaussian	Gamma	inverse.gaussian	poisson	quasi
logit	⊛					*
probit	*					*
cloglog	*					*
identity		⊛	*		*	*
inverse/recip.			⊛			*
log			*		⊛	*
1/mu^2				⊛		*
sqrt					*	*

Table 3: Families and the link functions available to them ⊛ = Canonical, or "default", links.

The combination of a response distribution, a link function and various other pieces of information that are needed to carry out the modelling exercise is called the *family* of the generalized linear model.

Table 10.1 The link functions used by GLIM. The canonical link function for normal errors is the identity link, for Poisson errors the log link, for binomial errors the logit link and for gamma errors the reciprocal link. These canonical link functions are defined by default when the error structure is declared

Symbol	Link function	Formula	Use
I	Identity	$\eta = \mu$	Regression or ANOVA with normal errors
L	Log	$\eta = \log \mu$	Count data with Poisson errors
G	Logit	$\eta = \log\left(\frac{\mu}{n-\mu}\right)$	Proportion data with binomial errors
R	Reciprocal/Inverse	$\eta = \frac{1}{\mu}$	Continuous data with gamma errors
P	Probit	$\eta = \Phi^{-1}(\mu/n)$	Proportion data in bioassays
C	Complementary log-log	$\eta = \log[-\log(1-\mu/n)]$	Proportion data in dilution assay
S	Square root	$\eta = \sqrt{\mu}$	Count data
E	Exponent	$\eta = \mu^{**}number$	Power functions
O	User defined own link function.		

One important difference between generalized linear modelling and conventional modelling (anova, regression etc) is that the "regression"/"anova" is not actually carried out on the response variable, Y, but on a linearized version of the link function as applied to Y. By going **"backwards"** through the link function, one can attain everything we are used to "getting" from a model (i.e., predicted values of Y, s.e. of Y, residuals, regression coefficients expressed in terms of Y, etc).

Modelling

In order to chose the correct error structure and link function (which is very important), you will need to:

- think about and examine your data, in particular your **response variable!**
- consult a reference on generalized linear modelling (e.g., see books below – or search the web for information on generalized linear models)
- compare model fit between models using different link functions (see below)
- assess, and remedy, potential problems with overdispersion (see below)
- assess model fit with graphical inspection of residuals
- perhaps discuss things with someone with some experience of generalized linear modelling.

All of this will require more than just a "click on a button", but you will learn from the experience, and you can be certain to end up with a more appropriate analysis!

Some common basic model types

- *Count data*

  When response data is **count data** – use Poisson errors and a log (possibly identity or square root) link function. Also called Poisson regression models.

- *Proportion data*

  When response data is expressed as a **proportion** (y out of n) – use binomial errors and a logit (possibly probit or complimentary log-log) link function.

- *Dichotomous outcome*

  When response data is expressed as a **binary outcome** (e.g., yes or no, 1 or 0, etc.) – use binomial errors and a logit (possibly probit or complimentary log-log) link function with 1 as the binomial denominator (i.e., 0 or 1 "out of" 1). Also termed logistic regression models.

Deviance

The discrepancy between the estimated model and the actual data is called the **deviance**. It is, in many ways, precisely analogous to **SSE** in conventional LS modelling.

ANODEV - testing your model and effects

Generalized linear models are typically assessed with **analysis of deviance**, closely analogous to analysis of variance in general linear models. The overall test statistic **G, sometimes (e.g., by Quinn and Keough) called G^2,** is also called the log-likelihood ratio (**LLR**) and is a measure of **goodness-of-fit** of model to data. The subsequent tests are called log-likelihood ratio tests or G-tests. *They are actually very simple*!

- LLR is calculated simply as the difference in deviance between the null model and the model under assessment: **G = D1 - D2**
- G is χ^2 distributed with df = difference in the total number of fitted parameters in the two models (1 for each continuous explanatory variable and k-1 for each factorial variable [k = number of levels of factor]).

1. Fit a null model, stating that Y is invariant of X_{ij} (only the constant/mean in the model).
2. Fit the full model with X_{ij} as independent variables / factors.
3. Test the model by taking the difference in deviance and comparing it to χ^2, df = difference in number of fitted parameters in two models.
4. Hierarchical models can also be compared in the same way, by taking the difference in deviance and comparing it with the chi-square value for df = difference in numbers of parameters in reduced and full models. This will tell you if the model has "improved" significantly by the addition of new factors/variables - this is precisely analogous to partial F-tests in regression/ANOVA (with type II sums-of-squares) – gives tests for effects of factors in your ANODEV.

Tests of single factors/variables

Generalized linear models generates estimates of coefficients of "slopes" (directly analogous to β in general linear models), and standard error estimates based on the likelihood function, **for covariates and for continuous variables in regression type models**. These can also be used to test effects of single variables in uni- or multivariate models by regular t-tests (as in regression/multiple regression).

- $t = \beta / SE_\beta$
- df = n - k - 2, where n is number of observations and k is the number of parameters/variables in the model.

Factors are typically tested by improvement in model fit by adding the factor (see above) – by χ^2 based tests of G or with F approximations based on ANODEV.

Overdispersion of errors

A word of caution: a common “problem with” binomial and poisson errors is **overdispersed errors**. This occurs either when you have failed to measure some important explanatory variables or when the distribution of data is not quite what you want it to be, and it means that the probability you are modelling behaves like a random variable. You can spot overdispersion by a “too large” residual deviance. ***Whenever the deviance of your full model is substantially larger than your residual degrees of freedom (that is, your sample size (n)) you should not trust your P-values*.** If the deviance is more than, say, 1.5 times as large as the degrees of freedom (**ideally; residual deviance = sample size**) you should start worrying. There are several easy and robust "fixes" to problems with overdispersion, which either involves some form of “scaling” deviances prior to making tests (consult e.g. Crawly 1993) or the use of quasi-likelihood estimation. Occasionally, **underdispersion** will also occur (need to be dealt with).

How do you report your analysis?

A good idea is to consult other papers where people have used generalized linear models, to see how results and analyses have been reported.

A) State your **model type** clearly in the M&M section (which error structure, and which link function). Also briefly motivate your choice and assessment of models. Inspect the residuals by visual inspection of residual plots, and state clearly that this was done in the M&M section.
B) Give deviances of models in the Results section, and give a clear account of your **deviance analysis**: which models are compared, which is the LLR and the associated df. This is often best done in a table, where the complete deviance analysis is presented.

C) If you are interested in effects of single variables in regression-type models, give a table with the coefficient estimates, their s.e., and the associated t and p values.

Computer programs

- Many (most?) computer programs can perfrom some form of GLIMs - typically, however, restricted to only logistic regression and/or probit regression...
- Genstat - a great, very powerful (but pretty expensive!) and reasonably "user friendly" program for GLIM (as well as almost everything else!). You can email them to try a free trial version download at:
  http://www.vsni.co.uk/downloads/genstat/
  Genstat is much used in genetics and agricultural research.
- R
- SAS
- S-plus
- Lisp-stat
- Stata
- MatLab (e.g. GLMFIT)
- GLIM - an older but excellent DOS program for GLIMs (free "inofficial" version from me).
- XploRe - an older but useful program for GLIM, freely available at http://fedc.wiwi.hu-berlin.de/xplore.php

Some good books to consult

Crawly, M. J. 1993. GLIM for Ecologists. Blackwell, London. *(A pretty basic but **very** good book on GLIMs using ecological data examples; also serves as a very pedagogical hands-on manual for the program GLIM! Highly recommended!)*

Crawly, M. J. 2007. The R book. Wiley, Chichester. *(A very good book on R and how to run all sorts of models, including GLIMs)*

Dobson, A. J. 1990. An Introduction to Generalized Linear Models. Chapman & Hall. *(A good general intro to GLIMs)*

McCullagh, P. & Nelder, J. A. 1989. Generalized Linear Models. 2nd edition. Chapman and Hall, London. *(**The** standard reference (bible....) for GLIMs; often hard to grasp, though, since it is fairly technical)*

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. 1989. Statistical modelling in GLIM. Clarendon Press, Oxford. *(A detailed and comprehensive "manual" to the program GLIM)*

Hardin, J. and Hilbe, J. 2001. Generalized Linear Models and Extensions. College Station, Texas: Stata Press. *(A pretty applied but good book covering the fundamentals, including worked out analyses using the software Stata).*