Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Generalized linear modeling (GLM) is used to fit statistical models to observations that have a known error distribution, and where covariates can be assumed to contribute in a linear manner via a specified intermediate link function.

Introduction

The GLM technique is intermediate between linear regression, which is trivial and gives uniquely determined parameter estimates but is rarely appropriate, and nonlinear regression, which is very hard and does not usually give unique parameter estimates, but is justified with normal errors and a known model.

To understand the motivation for this technique, it is usual to refer to a typical doubling dilution experiment in which diluted solutions from a stock containing infected organisms are plated onto agar in order to count infected plates, and hence estimate the number of organisms in the stock. Suppose that before dilution the stock had N organisms per unit volume, then the number per unit volume after $x = 0, 1, \ldots, m$ dilutions will follow a Poisson dilution with $\mu_x = N/2^x$. Now the chance of a plate receiving no organisms at dilution x is the first term in the Poisson distribution, that is $\exp(-\mu_x)$, so if p_x is the probability of a plate becoming infected at dilution x, then

$$p_x = 1 - \exp(-\mu_x), \;\; x = 1, 2, \ldots, m.$$

Evidently, where the p_x have been estimated as proportions from y_x infected plates out of n_x plated at dilution x, then N can be estimated using

$$\log[-\log(1 - p_x)] = \log N - x \log 2$$

considered as a maximum likelihood fitting problem of the type

$$\log[-\log(1 - p_x)] = \beta_0 + \beta_1 x$$

where the errors in estimated proportions $p_x = y_x/n_x$ are binomially distributed.

The SimFIT generalized models interface can be used from **gcfit**, **linfit** or **simstat** as it finds many applications, ranging from bioassay to survival analysis.

Basic theory

So, to fit a generalized linear model, you must have independent evidence to support your choice for an assumed error distribution for the dependent variable Y from the following possibilities:

- normal
- binomial
- Poisson
- gamma

in which it is supposed that the expectation of Y is to be estimated, i.e.,

$$E(Y) = \mu.$$

The associated *pdfs* are parameterized as follows.

$$\begin{aligned}
\text{normal}: f_Y &= \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)\\
\text{binomial}: f_Y &= \binom{N}{y}\pi^y(1-\pi)^{N-y}\\
\text{Poisson}: f_Y &= \frac{\mu^y\exp(-\mu)}{y!}\\
\text{gamma}: f_Y &= \frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu}\exp\left(-\frac{\nu y}{\mu}\right)\frac{1}{y}
\end{aligned}$$

It is a mistake to make the usual unwarranted assumption that measurements imply a normal distribution, while proportions imply a binomial distribution, and counting processes imply a Poisson distribution, unless the error distribution assumed has been verified for your data. Another very questionable assumption that has to made is that a predictor function η exists, which is a linear function of the m covariates, i.e., independent explanatory variables, as in

$$\eta = \sum_{j=1}^{m}\beta_j x_j.$$

Finally, yet another dubious assumption must be made, that a link function $g(\mu)$ exists between the expected value of Y and the linear predictor. The choice for

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

depends on the assumed distribution as follows. For the binomial distribution, where y successes have been observed in N trials, the link options are the logistic, probit or complementary log-log

$$\begin{aligned}
\text{logistic}: \eta &= \log\left(\frac{\mu}{N-\mu}\right)\\
\text{probit}: \eta &= \Phi^{-1}\left(\frac{\mu}{N}\right)\\
\text{complementary log-log}: \eta &= \log\left(-\log\left(1-\frac{\mu}{N}\right)\right).
\end{aligned}$$

Where observed values can have only one of two values, as with binary or quantal data, it may be wished to perform binary logistic regression. This is just the binomial situation where y takes values of 0 or 1, N is always set equal to 1, and the logistic link is selected. However, for the normal, Poisson and gamma distributions the link options are

$$\begin{aligned}
\text{exponent}: \eta &= \mu^a\\
\text{identity}: \eta &= \mu\\
\text{log}: \eta &= \log(\mu)\\
\text{square root}: \eta &= \sqrt{\mu}\\
\text{reciprocal}: \eta &= \frac{1}{\mu}.
\end{aligned}$$

In addition to these possibilities, you can supply weights and install an offset vector along with the data set, the regression can include a constant term if requested, the constant exponent a in the exponent link can be altered, and variables can be selected for inclusion or suppression in an interactive manner. However, note that the same strictures apply as for all regressions: you will be warned if the SVD has to be used due to rank deficiency and you should redesign the experiment until all parameters are estimable and the covariance matrix has full rank, rather than carry on with parameters and standard errors of limited value.

The simplified GLM interface

Although generalized linear models have widespread use, specialized knowledge is sometimes required to prepare the necessary data files, weights, offsets, etc.

For this reason, there is a simplified SIMFIT interface to facilitate the use of GLM techniques in such fields as the following.

- Bioassay, assuming a binomial distribution and using logistic, probit, or log-log models to estimate percentiles, such as the LD50.
- Logistic regression and binary logistic regression.
- Logistic polynomial regression, generating new variables interactively as powers of an original covariate.
- Contingency table analysis, assuming Poisson errors and using log-linear analysis to quantify row and column effects.
- Survival analysis, using the exponential, Weibull, extreme value, and Cox (i.e., proportional hazard) models.

Of course, by choosing the advanced interface, users can always take complete control of the GLM analysis, but for many purposes the simplified interface will prove much easier to use for many routine applications.

Warning

The GLM procedure involves an iterative technique to estimate parameters from starting estimates and, in this respect, it is similar to nonlinear regression in that it will only succeed if the following conditions are satisfied.

1. The error type and link function (i.e. the model) must be chosen sensibly.
2. The data must be formatted in a specific manner depending on the error type and link function selected, as will be explained in subsequent worked examples.
3. The data must be sufficiently accurate and cover a wide enough range to allow the parameters to be estimated.
4. Error messages about failure to fit or poor parameter estimates must be interpreted sensibly, and then appropriate action taken.

Only when all these conditions are satisfied will SIMFIT be able to fit a GLM model.