Simfit

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

It is often of interest to fit a model to data in order to estimate parameters such as the 50% point from dose-response curves, and SIMFIT provides several dedicated programs for this purpose such as the following.

- **exfit**: fits one or sums of exponentials and calculates the area under the curve (AUC).
- **mmfit**: fits one or sums of Michaelis-Menten models and calculates the apparent K_m.
- **hlfit**: fits one or sums of binding models and calculates the apparent K_a.
- **sffit**: fits cooperative binding models and calculates half saturation points.
- **gcfit**: fits nonlinear growth models and calculates maximal growth rates.
- **inrate**: fits several models and calculates initial rates.
- **polnom**: fits polynomials and calculates y given x.
- **calcurve**: fits cubic splines and calculates y given x.
- **qnfit**: fits user defined models and calculates y given x.

These programs all assume uncorrelated normally distributed errors, but there are many procedures, such as bioassay, dose-response curves, determination of LD50, or EC50 etc. where binomially distributed errors would be more appropriate, and so it would be better to fit general linear models (GLM)

This would be a situation such as the following dose-response data set contained in the default test file `ld50.tf1` which can be inspected after opening the main SIMFIT menus, followed by selecting [Statistics], [Analysis of proportions], then [Bioassay, dose response curves and LD50].

y	N	x
1	10	1
4	20	2
4	10	3
5	10	4
15	30	5
7	10	6
9	10	7
12	15	8
9	10	9
8	10	10

Data for determination of LD50 by GLM requires the above format as follows for k groups and $i = 1, 2, \ldots, k$.

- Column 1: $y_i \geq 0$, the number of animals dying in group i
- Column 2: $N_i \geq y_i$, the number of animals tested in group i
- Column 3: $x_i \geq 0$, the amount of poison being tested on group i

If the k groups are all independent and each group is homogeneous, i.e., each animal in the group has exactly the same probability p of dying given the same time of exposure to poison at amount x for the same period of time, then y is binomially distributed and p_i can be estimated as $\hat{p}_i = y_i/N_i$, together with exact confidence limits.

It is usual to investigate a data set to choose a model with the lowest deviance and the next table shows the results from analysis of the data in the default test file using the three GLM link functions indicated.

Method: GLM with binomial errors, Link: Logistic
Number of groups = 10, Deviance = 4.246

Parameter	Value	Standard error	Lower 95%cl	Upper 95%cl	*p*
Constant	-2.0986	0.4733	-3.190	-1.007	0.0022
Slope	0.45070	0.08725	0.2495	0.6519	0.0009
50% point	4.6564	0.4441	3.632	5.681	0.0000

Method: GLM with binomial errors, Link: Probit
Number of groups = 10, Deviance = 4.564

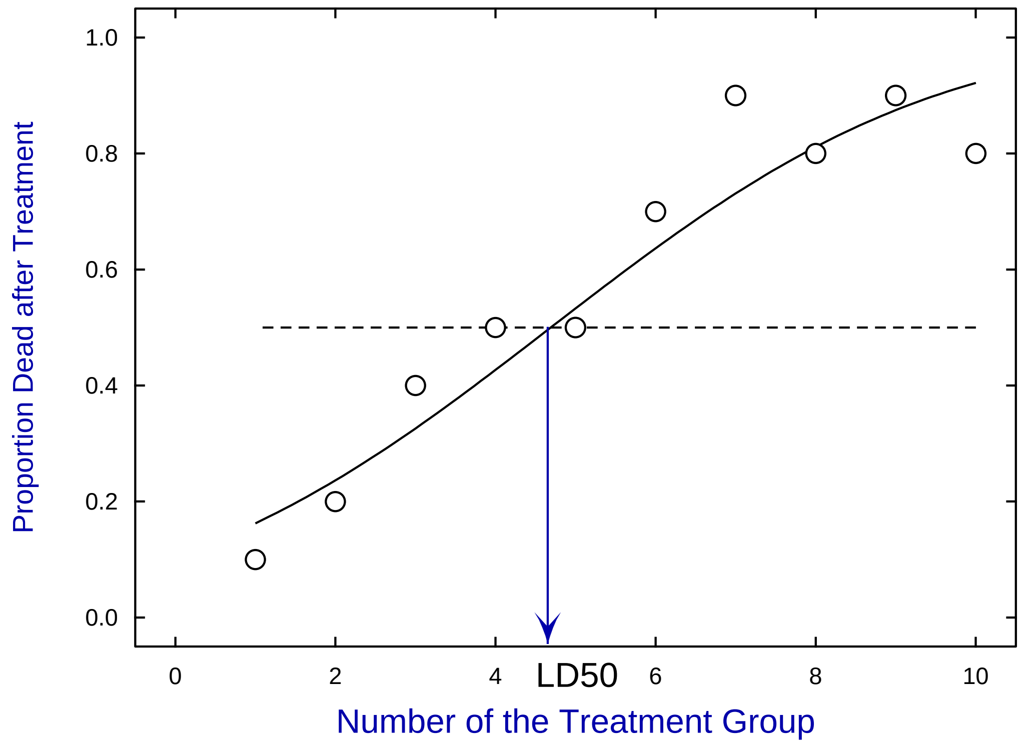
Parameter	Value	Standard error	Lower 95%cl	Upper 95%cl	*p*
Constant	-1.2513	0.2708	-1.876	-0.6269	0.0017
Slope	0.26678	0.04855	0.1548	0.3787	0.0006
50% point	4.6902	0.4463	3.661	5.719	0.0000

Method: GLM with binomial errors, Link: Complementary log-log
Number of groups = 10, Deviance = 6.600

Parameter	Value	Standard error	Lower 95%cl	Upper 95%cl	*p*
Constant	-1.6696	0.3295	-2.429	-0.9097	0.0010
Slope	0.26635	0.05079	0.1492	0.3835	0.0008
50% point	4.89220	0.5182	3.697	6.087	0.0000

In this case the logistic and probit models give a similar fit, which is somewhat better than the complementary log-log, and so the standard probit graph is shown next.

Data, best-fit Probit, and 50% point (LD50).

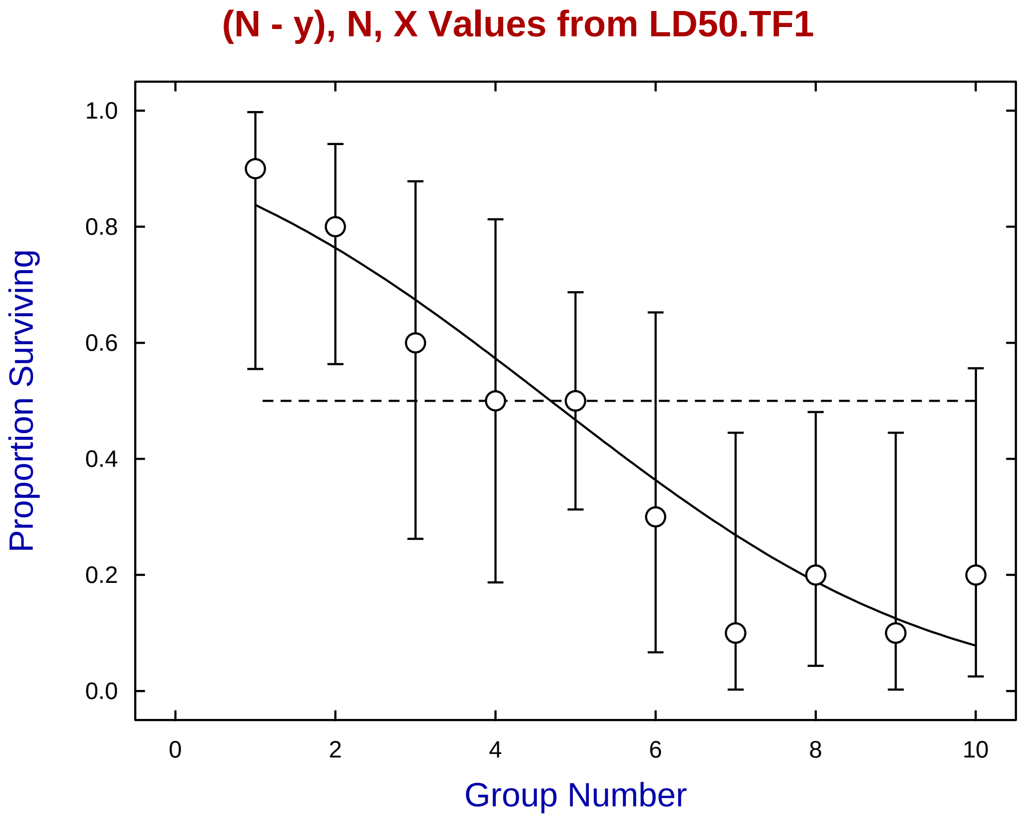


Various options are available for testing the goodness of fit by plotting residuals or inspecting tables of residuals as shown next.

Number	Y-value	Theory	Deviance	Leverages
1	1	1.624	-0.5692	0.2308
2	4	4.729	-0.3912	0.3731
3	4	3.260	0.4907	0.1391
4	5	4.270	0.4645	0.1030
5	15	15.99	-0.3612	0.2652
6	7	6.366	0.4227	0.09853
7	9	7.311	1.328	0.1282
8	12	12.17	-0.1118	0.2492
9	9	8.749	0.2476	0.1985
10	8	9.217	-1.219	0.2143

Note that exact 95% confidence limits can also be plotted but, as these can be large and very distracting with small samples, they can be switched off.

A further point can be made about this GLM procedure. Suppose that, instead of a file with y, N, x for the proportion failing, we input a file with $N - y, N, x$. This would then be the proportion surviving as plotted below.



Note that this change from proportion failing to the complement, that is the proportion surviving, leads to exactly the same estimate for LD50.

Another variant of the this technique is that a parameter can be changed in order to estimate other percentiles than the 50% point, e.g., LD25, LD75, EC25, ID25, ED75, etc.

Theory 1: GLM

It is important to understand that fitting dose-response curves in the manner just described does not correspond to the usually understood technique of adjusting the parameters of a deterministic mathematical model by optimization to obtain a best-fit curve that minimizes the sum of squared residuals. For this reason a brief overview of generalized linear modeling (GLM) is now presented.

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. 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 normal, binomial, Poisson, or gamma distributions, 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.

Theory 2: 95% confidence regions in inverse prediction

The calculation of confidence limits for derived values, such as LD50 in the present case, that are obtained from the parameter estimates from fitting along with the estimated parameter covariance matrix should be noted.

polnom estimates non-symmetrical confidence limits assuming that the N values of y for inverse prediction and weights supplied for weighting are exact, and that the model fitted has n parameters that are justified statistically. **calcurve** uses the weights supplied, or the estimated coefficient of variation, to fit confidence envelope splines either side of the best fit spline, by employing an empirical technique developed by simulation studies. Root finding is employed to locate the intersection of the y_i supplied with the envelopes. The AUC, LD50, half-saturation, asymptote and other inverse predictions in SIMFIT use a t distribution with $N-n$ degrees of freedom, and the variance-covariance matrix estimated from the regression. That is, assuming a prediction parameter defined by $p = f(\theta_1, \theta_2, \ldots, \theta_n)$, a central 95% confidence region is constructed using the prediction parameter variance estimated by the propagation of errors formula

$$\hat{V}(p) = \sum_{i=1}^{n}\left(\frac{\partial f}{\partial \theta_i}\right)^2 \hat{V}(\theta_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{\partial f}{\partial \theta_i}\frac{\partial f}{\partial \theta_j}\hat{CV}(\theta_i, \theta_j).$$

Note that this formula for the propagation of errors can be used to calculate parameter standard errors for parameters that are calculated as functions of parameters that have been estimated by fitting, such as apparent maximal velocity when fitting sums of Michaelis-Menten functions. However, such estimated standard errors will only be very approximate.