Simfit

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

Dose-response curves and related bioassay techniques are frequently used to estimate percentiles such as LD50, the median concentration causing 50% mortality. When the experiment consists of estimating proportions in disjoint groups as a function of some effector, the proportions are estimates of binomial probabilities so it may be reasonable to fit a generalized linear model (GLM) that assumes binomial distribution of error.

Example 1

From the main SimFIT menu choose [A/Z] then proceed as follows.

- Open program **gcfit**
- Select the option to perform bioassay (percentiles/EC50/LD50)
- Read in the default test file `ld50.tf1` using the [Demo] button on the SimFIT file selection control

Test file `ld50.tf1` contains the following data.

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

The results were for the number of animals dying in ten separate groups after dosing with toxin as follows.

1. **Column 1**
   The number of animals dying (y_i) within a fixed time at a dose x_i
2. **Column 2**
   The number of animals in the corresponding group (N_i)
3. **Column 3**
   The concentration of toxin (x_i)

Here, for $i = 1, 2, \ldots, 10$ there is a binomial distribution with probabilities p_i and estimates $\hat{p}_i$ as follows

$$\hat{p}_i = \frac{y_i}{N_i}.$$

Further, as the groups are independent, we can calculate exact non-symmetrical confidence limits for such estimates. However, recognizing the distribution of the proportions has no further value unless it is possible to construct a model for the dependence of such proportions on the effector substance at concentration x_i.

In the event that a deterministic model cannot be constructed in the usual way it is possible to explore generalized linear models to see if they can give an adequate fit and yield meaningful estimates of the percentiles with confidence limits. For instance, here are the results from analyzing the test data using the logistic, probit, and complementary log-log models in order to estimate the 50% point. From the results it is clear that, in this case, all three models give comparable estimates for the 50% point.

Method: GLM with binomial errors, Link: Logistic
Number of samples = 10, Deviance = 4.2461

Parameter	Value	std. error	Lower95%cl	Upper95%cl	*p*
Constant	-2.0986	0.47329	-3.1900	-1.0072	0.0022
Slope	0.4507	0.08725	0.2495	0.6519	0.0009
50% point	4.6564	0.44415	3.6322	5.6806	0.0000

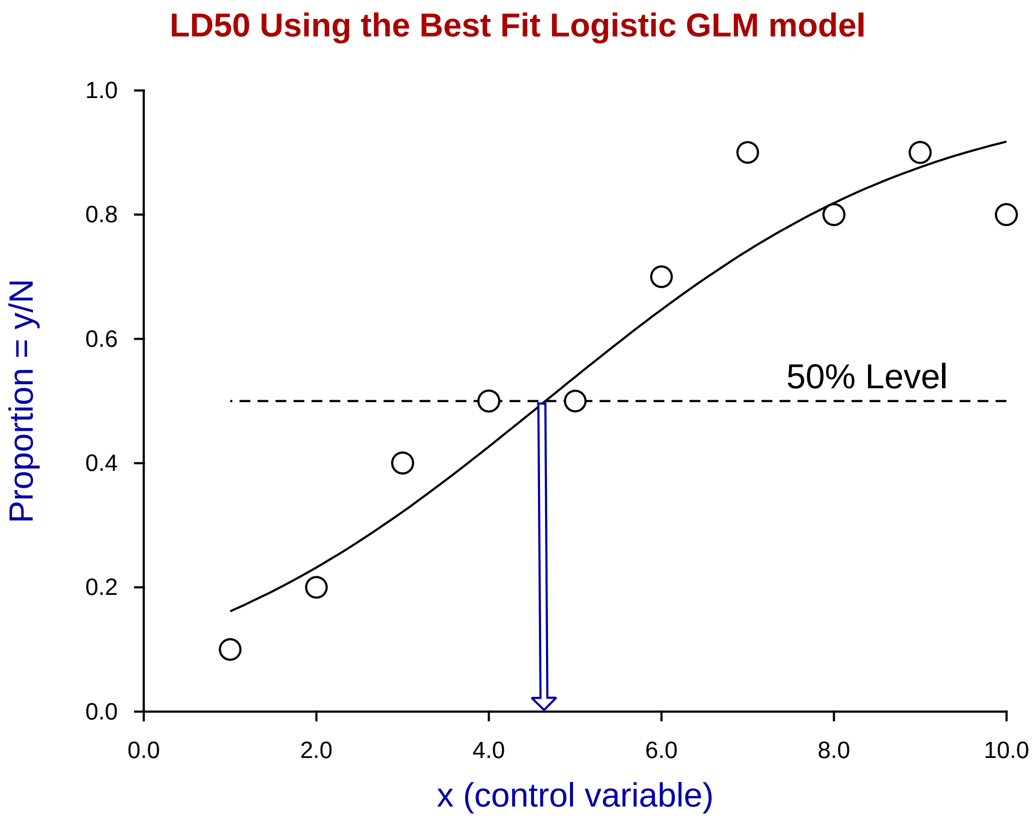
Method: GLM with binomial errors, Link: Probit
Number of samples = 10, Deviance = 4.5642

Parameter	Value	std. error	Lower95%cl	Upper95%cl	*p*
Constant	-1.2513	0.27078	-1.8757	-0.6269	0.0017
Slope	0.2668	0.04855	0.1548	0.3787	0.0006
50% point	4.6902	0.44633	3.6610	5.7194	0.0000

Method: GLM with binomial errors, Link: Complementary log-log
Number of samples = 10, Deviance = 6.6004

Parameter	Value	std. error	Lower95%cl	Upper95%cl	*p*
Constant	-1.6696	0.32951	-2.4294	-0.9097	0.0010
Slope	0.2664	0.05079	0.1492	0.3835	0.0008
50% point	4.8922	0.51820	3.6972	6.0872	0.0000

The next graph shows the best-fit curve obtained with the logistic model and indicates how the intersection of $\hat{p} = 0.5$ with the curve leads to the estimation of the 50% point. Actually other percentiles can also be estimated if required.



Example 2

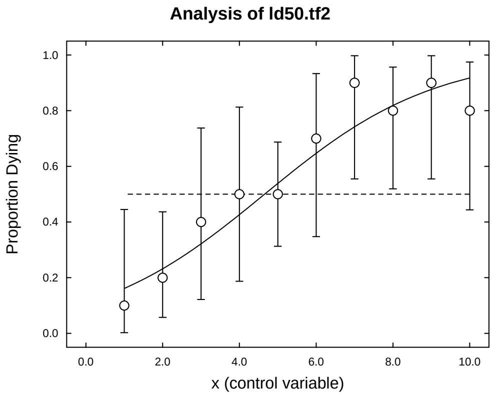
As GLM techniques are to be used it should be mentioned that the analysis just described can also be carried out with data in the standard GLM format. For instance, the column order for binomial GLM fitting is covariates first, successes, trials, then finally weighting factors. Generally, as in this case, the weighting factors would be 1, but they must be included so that the number of covariates is counted correctly.

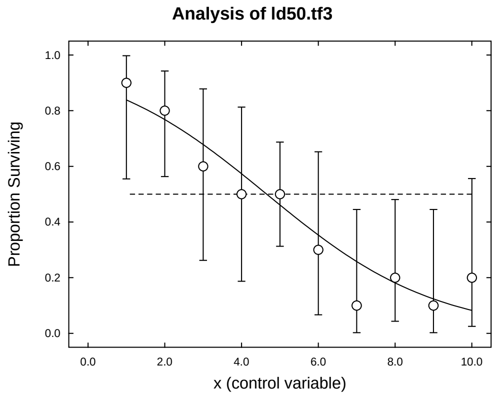
Another point is that the same estimates for percentiles can also be carried out with the data in complementary format, where the percent surviving is estimated and not the percent dying, as long as it is remembered to reflect values other than the 50% point. One major advantage of the GLM data format is that it can be extended to include other covariates such as time, or allowing for heterogeneity in the groups.

Here, for example, are the data contained in test files `ld50.tf2` and `ld50.tf3`.

ld50.tf2				ld50.tf3			
x	y	N	s	x	N-y	N	s
1	1	10	1	1	9	10	1
2	4	20	1	2	16	20	1
3	4	10	1	3	6	10	1
4	5	10	1	4	5	10	1
5	15	30	1	5	15	30	1
6	7	10	1	6	3	10	1
7	9	10	1	7	1	10	1
8	12	15	1	8	3	15	1
9	9	10	1	9	1	10	1
10	8	10	1	10	2	10	1

The next graphs illustrate the difference between analyzing data as proportion surviving instead of proportion dying, and from these it will be clear that both data sets give the same 50% point but that any other percentiles would have to be treated carefully. For instance, 10% dying would be equivalent to 90% surviving.





Now the standard SIMFIT analysis of proportions routines only involve an indicator variable x to identify samples or to plot the variation in estimates with confidence limits as functions of x. In the estimation of LD50 we have to make further assumptions as to how the binomial parameter depends on x, not as an indicator variable but as an independent variable or, as it is usually referred to in generalized linear models, a covariate. Such models are not based on biochemical theories as to how a toxin acts, but are empirical models that are only useful to the extent that they fit the dose-response curve adequately.

Theory

Given N trials at fixed x with probability p for success in each trial then the probability of y successes is

$$P(y) = \binom{N}{y} p^y (1-p)^{N-y}, \text{ for } y = 0, 1, \ldots, N$$

and the best estimate for p is

$$\hat{p} = \frac{y}{N}$$

where an unsymmetrical $100(1-\alpha)\%$ confidence range (p_l, p_u) may be obtained by solving the nonlinear equations

$$\sum_{t=y}^{N} \binom{N}{t} p_l^t (1-p_l)^{N-t} = \alpha/2$$
$$\sum_{t=0}^{y} \binom{N}{t} p_u^t (1-p_u)^{N-t} = \alpha/2$$

for p_l and p_u.

The generalized linear model for binomial errors supposes that the expectation of Y is to be estimated, i.e.,

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

given the distribution

$$f_Y = \binom{N}{y} p^y (1-p)^{N-y}$$

and the assumption 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, it is assumed that a link function $g(\mu)$ exists between the expected value of Y and the linear predictor. The choices for

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

with the binomial distribution, where y successes have been observed in N trials, are the logistic, probit, or complementary log-log link functions

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