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

Logistic regression is widely used to model experiments with only one of two outcomes such as success or failure which, unlike the simple method for analysis of binomial proportions, depend on the values of k covariates $x_1, x_2, \ldots, x_k$, where $k \geq 1$.

Example 1: Alcohol and congenital abnormalities

From the main SIMFIT menu choose [Statistics], [Generalized linear models], then [Logistic regression], and examine the default test file `logistic.tf3` containing the following data.

x	y	N	s
0.0	48	17066	1
0.5	38	14464	1
1.5	5	788	1
4.0	1	126	1
7.0	1	37	1

These data were taken from a study of the effects of consuming alcoholic drinks on congenital abnormalities noted in infants after birth.

1. Column 1: x, alcoholic drinks consumed per day by mother
2. Column 2: y, infants born with abnormalities
3. Column 3: N, sample size
4. Column 4: s, weighting factors ($s = 1$ indicates unweighted analysis)

Logistic regression by was used to fit the GLM model

$$\log[y/(N-y)] \approx \beta_0 + \beta_1 x$$

which yielded the following parameter estimates, residuals, then observed and estimated frequencies.

Number of parameters = 2, Rank = 2, Number of points = 5, Degrees of freedom = 3

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p	$exp(\beta_1)$
Constant	-5.95840	-6.32583	-5.59097	0.115454	0.0000	
β_1	0.31927	-0.08038	0.71889	0.125574	0.0845 *	1.37611

Deviance = 1.96760

Number	Y-value	Theory	Dev-resid.	Leverage
1	48	43.9856	0.597220	0.584800
2	38	43.7119	-0.885127	0.476721
3	5	3.27338	0.886968	0.097194
4	1	1.15684	-0.149976	0.246568
5	1	0.87238	0.135174	0.594717

Observed	Estimated
0.0028	0.0026
0.0026	0.0030
0.0064	0.0041
0.0079	0.0092
0.0270	0.0236

Example 2: The symmetrical case with one variable

Logistic regression is frequently used to model the variation in binomial probability p as a simple linear function of a variable x often without realizing that, because of the necessary symmetry of the logistic function, this will usually lead to a biased fit.

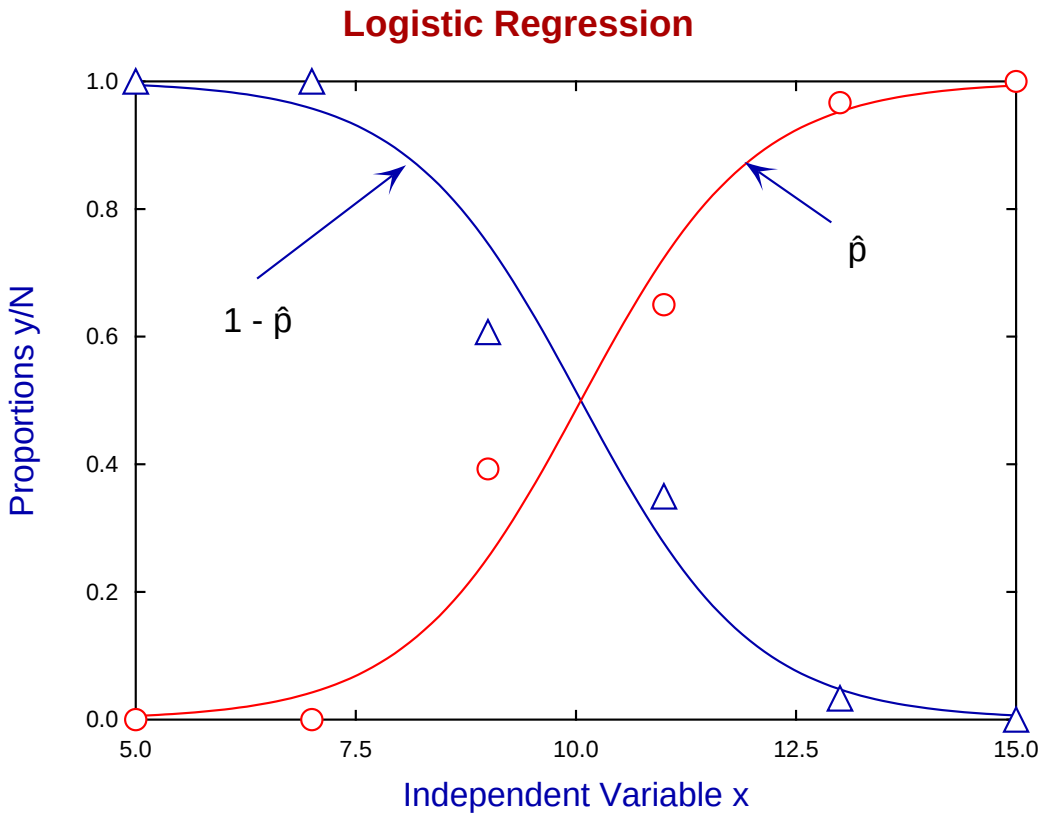
For instance, consider the data file `logistic.tf4` below

5	0	39	1
7	0	30	1
9	11	28	1
11	26	40	1
13	29	30	1
15	20	20	1

and its complement `logistic.tf5`

5	39	39	1
7	30	30	1
9	17	28	1
11	14	40	1
13	1	30	1
15	0	20	1

simulated by SimFIT using a random choice from an integer uniform distribution for N $(20 \leq N \leq 40)$ followed by a random choice from a binomial distribution for y given N and $p(x)$ which were then fitted as indicated in the next graph.



Exact parameters were $\beta_0 = -10, \beta_1 = 1$ for $p(x)$ and $\beta_0 = 10, \beta_1 = -1$ for $1 - p(x)$ and the best fit parameters are in the next tables.

Number of parameters = 2, Rank = 2, Number of points = 6, Degrees of freedom = 4

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p	$exp(\beta_1)$
Constant	-10.2573	-14.5246	-5.98991	1.53699	0.0026	
β_1	1.01996	0.60551	1.43440	0.14927	0.0024	2.77307

Deviance = 7.07433

Number of parameters = 2, Rank = 2, Number of points = 6, Degrees of freedom = 4

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p	$exp(\beta_1)$
Constant	10.2573	5.98391	14.5307	1.53915	0.0026	
β_1	-1.01996	-1.43498	-0.60494	0.14948	0.0024	0.360610

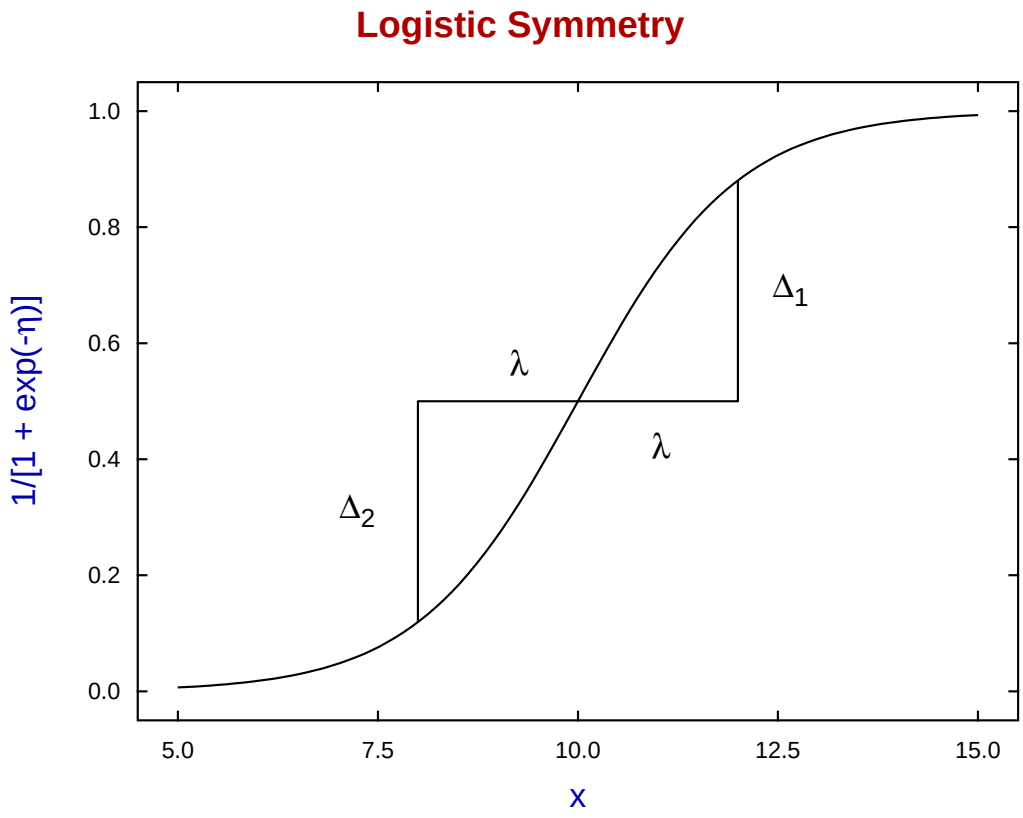
Deviance = 7.07433

Due to the symmetry of the logistic curves for p and $1-p$ about their midpoints $x_{1/2}$

$$p = \frac{1}{1+\exp(-\eta)} = 1 - p = \frac{1}{1+\exp(\eta)} = \frac{1}{2}$$

so that the mid point requires $\eta = 0$, that is $x_{1/2} = -\beta_0/\beta_1$.

Moving a horizontal distance λ to either side of the midpoint generates two vertical distances Δ_1 and Δ_2 which are equal as shown in this next graph.



This is because

$$\Delta_1 = \frac{1}{1+\exp(-\beta_1\lambda)} - \frac{1}{2} = \Delta_2 = \frac{1}{2} - \frac{1}{1+\exp(\beta_1\lambda)}.$$

Theory

In a situation where the probability of success is a fixed constant p and Y is the number of successes in N trials ($N \geq 1$), then the probability the Y equals any specific value y where $0 \leq y \leq N$ is

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

and p can be estimated as $\hat{p}$ where the estimate, expectation and variance are

$$\begin{aligned} \hat{p} &= y/N \\ E(y) &= Np \\ V(y) &= Np(1-p). \end{aligned}$$

When the binomial parameter is a function of some variables $x_1, x_2, \ldots, x_k$ then a functional relationship must be proposed to model $p(x)$ and this must be fitted to estimate parameters accounting for the variation in p. It is usual to do this by fitting a generalized linear model (GLM) with assumed binomial error and logistic link, but the reason for this model is not because it is the correct model but because of the following fact. In the simple case of one variable x the log odds ratio from fitting a GLM model with y_1 at x and y_2 at $x+1$ can then be expressed as

$$\begin{aligned} \log\left(\frac{y_1}{N-y_1}\right) &\approx \beta_0 + \beta_1 x \\ \log\left(\frac{y_2}{N-y_2}\right) &\approx \beta_0 + \beta_1 (x+1) \end{aligned}$$

and hence the odds ratio can be estimated as

$$\frac{\hat{p}_2/(1-\hat{p}_2)}{\hat{p}_1/(1-\hat{p}_1)} \approx \exp(\beta_1).$$

From this we could conclude that an increase of one alcoholic drink per day can be estimated to change the odds ratio for congenital malformation by about 1.376.

It is this seemingly easy method for interpreting the parameter estimates from logistic regression that has been responsible for the widespread and often uncritical adoption of this technique and its extension into areas such as

- Including multiple variables
- Analyzing cases with categorical variables.

So, in order to confirm goodness of fit, SimFIT outputs the deviance and deviance residuals defined as follows

$$\begin{aligned} \text{For binomial errors: } d_i &= 2\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (N_i - y_i)\log\left(\frac{N_i - y_i}{N_i - \hat{\mu}_i}\right)\right\} \\ \text{Deviance residuals: } r_i &= \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i} \\ \text{Deviance } &= \sum_{i=1}^{n} d_i. \end{aligned}$$

where there are n observations and $\hat{\mu}_i = N_i \hat{p}_i$, and these should always be considered before accepting a fit.