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

Binary 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: Fitting a binary logistic model

From the main SIMFIT menu choose [Statistics], [Generalized linear models], then [Binary logistic regression] (with no strata), then examine the default test file `logistic.tf1` containing the following data.

x_1	x_2	y	N	s
3.70	0.825	1	1	1
3.50	1.090	1	1	1
0.75	1.500	1	1	1
1.25	2.500	1	1	1
0.80	3.200	1	1	1
0.70	3.500	1	1	1
0.60	0.750	0	1	1
1.10	1.700	0	1	1
0.90	0.750	0	1	1
0.90	0.450	0	1	1
0.80	0.570	0	1	1
0.55	2.750	0	1	1
0.60	3.000	0	1	1
1.40	2.330	1	1	1
0.75	3.750	1	1	1
2.34	1.640	1	1	1
3.20	1.600	1	1	1
0.85	1.415	1	1	1
1.70	1.060	0	1	1
1.80	1.800	1	1	1
0.40	2.000	0	1	1
0.95	1.360	0	1	1
1.35	1.350	0	1	1
1.50	1.360	0	1	1
1.60	1.780	1	1	1
0.60	1.500	0	1	1
1.80	1.500	1	1	1
0.95	1.900	0	1	1
1.90	0.950	1	1	1
1.60	0.400	0	1	1
2.70	0.750	1	1	1
2.35	0.030	0	1	1
1.10	1.830	0	1	1
1.10	2.200	1	1	1
1.20	2.000	1	1	1
0.80	3.330	1	1	1
0.95	1.900	0	1	1
0.75	1.900	0	1	1
1.30	1.625	1	1	1

The format of this data file will now be explained. These are vasoconstriction data from Finney D. J. (1947) Biometrika, 34, 320-34 with the following meanings.

- Column 1: x_1 (volume of air inspired)
- Column 2: x_2 (rate of air inspiration)
- Column 3: $y = 1$ (vasoconstriction), or $y = 0$ (no vasoconstriction)
- Column 4: $N = 1$ (sample size)
- Column 5: $s = 1$ (unweighted)

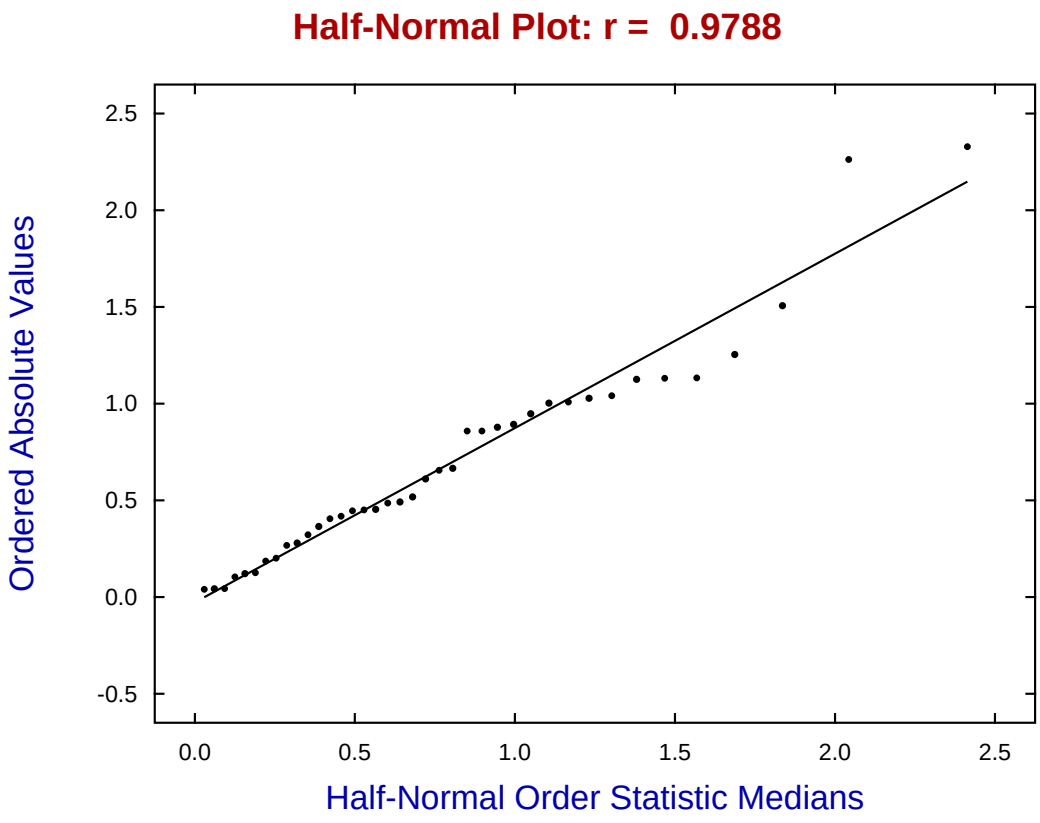
It is important to note that for binary logistic regression, y must be 1 (e.g. for success) or 0 (e.g. for failure), and N must be 1 because y is the outcome from a single Bernoulli trial, whereas for normal logistic regression, y must be in the range $0 \leq y \leq N$ with $N \geq 1$ for the number of trials resulting in y (e.g. the number of successful outcomes). The weighting factors would normally be $s = 1$ except for experienced users.

Fitting a generalized linear model with a mean, no offsets, binomial error and a logistic link leads to the following results table and half-normal residuals plot.

Number of parameters = 3, Rank = 3, Number of points = 39, Degrees of freedom = 36

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p	$\exp(\beta_i)$
Constant	-9.51999	-16.0587	-2.98131	3.22405	0.0055	
β_1	3.87719	0.986847	6.76753	1.42515	0.0100	48.2882
β_2	2.64683	0.797466	4.49619	0.91187	0.0063	14.1092

Deviance = 29.7656



Example 2: Predicting probabilities

Binary logistic regression seeks to find an approximation $\hat{p}(x)$ to a population binomial probability $p(x)$ that is not a constant probability but one that depends on covariates x as in the following model for the logodds

$$\log\left(\frac{p(x)}{1-p(x)}\right) \approx \eta(x)$$
$$\text{where } \eta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

and where the approximation results from fitting a generalized linear model (GLM) with binomial error, using a logistic link, when there are k covariates. The constant parameter β_0 in this polynomial simply estimates the logodds when all k covariates are zero, and can be included or omitted from the model.

Having estimated best-fit parameters and confirmed that the model is satisfactory it is often useful to predict what the probability would be given a set of covariates using the best-fit parameters in the next expressions

$$\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$
$$\hat{p} = \frac{\exp(\hat{\eta})}{1+\exp(\hat{\eta})}$$
$$= \frac{1}{1+\exp(-\hat{\eta})}.$$

So, after fitting has been completed, SIMFIT offers the possibility to do this either by inputting a set of covariates from the terminal or from a data file containing a matrix of covariates. Here, for instance, are the data contained in the test file `logistic.tf2`

x_1	x_2
0.4	1.0
0.6	1.0
0.8	1.0
1.0	1.0
1.0	0.8
1.0	0.6
1.0	0.4

which lead to these predictions after evaluation using the best-fit model from fitting `logistic.tf1`.

File: logistic.tf2
Data: Covariates to evaluate p after fitting logistic.tf1
$Y(x)$ evaluated for x_1 to x_2
Model includes a constant term
Binomial $N = 1$

y predicted	Probability estimated	Range of covariates
y_1 = 4.85785E-03	Binomial p = 0.004858	x = 0.4, ..., 1.0
y_2 = 1.04893E-02	Binomial p = 0.010489	x = 0.6, ..., 1.0
y_3 = 2.25015E-02	Binomial p = 0.022502	x = 0.8, ..., 1.0
y_4 = 4.76080E-02	Binomial p = 0.047608	x = 1.0, ..., 1.0
y_5 = 2.85997E-02	Binomial p = 0.028600	x = 1.0, ..., 0.8
y_6 = 1.70450E-02	Binomial p = 0.017045	x = 1.0, ..., 0.6
y_7 = 1.01099E-02	Binomial p = 0.010110	x = 1.0, ..., 0.4

As binary logistic regression is so widely used, often uncritically and without justification, to predict probabilities given covariates it is as well to consider the basic principles behind this method which will now be done.

Theory

The ideas behind binary logistic regression will be explained under several headings, namely

1. definitions;
2. one quantitative variable;
3. several quantitative variables;
4. categorical variables;
5. fitting technique; then
6. conclusion.

1. Definitions

A random integer variable Y can be formulated as taking a value depending on the result of an experiment with only one of two possible outcomes, such as heads/tails in coin tossing, death/survival following a serious illness, positive/negative of a value with respect to a baseline, etc. Arbitrarily calling one outcome a success and the other a failure we can sometimes define Y as taking two possible values, 1 or 0, depending on a probability p where $0 \leq p \leq 1$ that is

$$\begin{aligned} \text{Probability}(y = 1) &= p \\ \text{Probability}(y = 0) &= 1 - p. \end{aligned}$$

When the certain and impossible outcomes are excluded (i.e. $0 < p < 1$) the Odds can be defined as the ratio of success to failure as can the Log Odds, its natural logarithm, that is

$$\begin{aligned} \text{Odds} &= \frac{p}{1-p} \\ \text{Log Odds} &= \log\left(\frac{p}{1-p}\right). \end{aligned}$$

Given one trial with probability p_1 where $0 < p_1 < 1$ and one with probability p_2 where $0 < p_2 < 1$ the Odds Ratio and Log Odds Ratio are then given by

$$\begin{aligned} \text{Odds Ratio} &= \frac{p_2/(1-p_2)}{p_1/(1-p_1)} \\ \text{Log Odds Ratio} &= \log\left(\frac{p_2/(1-p_2)}{p_1/(1-p_1)}\right). \end{aligned}$$

At this point it should be emphasized that exponentials and logarithms to base e are used in theoretical developments and computational implementation, but many users prefer to present results using logarithms to base 10 in order to immediately clarify changes in orders of magnitude as powers of 10.

Of course such probabilities are never known exactly but must be estimated by sampling. In the case of N successive Bernoulli trials which are independent with identical probability it is usual to define a binomial variable W as

$$W = y_1 + y_2 + \cdots + y_N$$

so that the probability that $W = w$ where $0 \leq w \leq N$ is

$$P(W = w) = \binom{N}{w} p^w (1-p)^{N-w}$$

with expectation and variance given by

$$E(W) = Np$$
$$V(W) = Np(1-p).$$

From such a series of trials an estimate for the binomial parameter $\hat{p}$ is easily seen to be

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

and SimFIT provides dedicated analysis of proportions routines to calculate $\hat{p}$ with unsymmetrical confidence limits and plot these as a function of a parameter such as time, which only serves to order the observations for plotting and does not enter into the calculations.

Logistic regression using GLM is an extension of this subject to the case where additional variables x affect the probabilities under the assumption that each set of additional variables alters the binomial distribution distribution, i.e. $p = p(x)$, while binary logistic regression is just the special case where $N = 1$.

2. One quantitative variable

The GLM technique is used to adjust the two parameters β_0 and β_1 until the best-fit values are located to satisfy the approximation

$$\log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) \approx \hat{\beta}_0 + \hat{\beta}_1 x$$

according to the maximum likelihood criterion.

A special case is where the continuous variable is used at just two levels differing by one unit, say x and $x+1$, for then we have that, since

$$\begin{aligned}\text{Odds} &= \frac{p}{1-p} \\ &= \exp(\eta)\end{aligned}$$

then for the two levels x and $x+1$

$$\begin{aligned}\text{Odds Ratio} &= \frac{\text{Odds}(x+1)}{\text{Odds}(x)} \\ &= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1(x+1))}{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)} \\ &= \exp(\hat{\beta}_1)\end{aligned}$$

so that the Odds multiply by the factor $\exp(\hat{\beta}_1)$ for every one unit increase in the variable x or, alternatively,

$$\hat{\beta}_1 = \text{Log Odds Ratio}.$$

3. Several quantitative variables

Now, for k variables we have

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

which becomes difficult to fit as k increases, especially if the variables are not expressed in units so that the values are of similar size. Further, the arguments about interpreting the estimated parameters in terms of Log Odds Ratios, imply that only one variable is increased at a time with the others remaining fixed.

4. Categorical variables

Frequently the covariates are qualitative variables which can be included in the model by defining appropriate dummy indicator variables. For instance, suppose a factor has m levels, then we can define m dummy indicator variables $x_1, x_2, \ldots, x_m$ as in the next table.

Level	x_1	x_2	x_3	...	x_m
1	1	0	0	...	0
2	0	1	0	...	0
3	0	0	1	...	0
...	...	...	...	...	...
m	0	0	0	...	1

The data file would be set up as if to estimate all m parameters for the m factor levels but because only $m-1$ of the dummy indicator variables are independent, one of them would have to be suppressed if a constant were to be fitted, to avoid aliasing, i.e., the model would be overdetermined and the parameters could not be estimated uniquely. Suppose, for instance, that the model to be fitted was for a factor with three levels, i.e.,

$$\log\left\{\frac{p(x)}{1-p(x)}\right\} = a_0 + a_1x_1 + a_2x_2 + a_3x_3$$

but with any one of the x_i suppressed, x_1 for instance, since

$$x_{1i} + x_{2i} + x_{3i} = 1$$

for every i.

Then the estimated parameters could be interpreted as log odds ratios for the factor levels with respect to level 1, the suppressed reference level. This is because for probability estimates $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3$ we would have the odds estimates

$$\begin{aligned}
\frac{\hat{p}_1}{1-\hat{p}_1} &= \exp(\hat{a}_0) && (x_1 = 1, x_2 = 0, x_3 = 0)\\
\frac{\hat{p}_2}{1-\hat{p}_2} &= \exp(\hat{a}_0 + \hat{a}_2) && (x_1 = 0, x_2 = 1, x_3 = 0)\\
\frac{\hat{p}_3}{1-\hat{p}_3} &= \exp(\hat{a}_0 + \hat{a}_3) && (x_1 = 0, x_2 = 0, x_3 = 1)
\end{aligned}$$

and estimates for the corresponding log odds ratios involving only the corresponding estimated coefficients

$$\begin{aligned}
\log\left\{\frac{\hat{p}_2/(1-\hat{p}_2)}{\hat{p}_1/(1-\hat{p}_1)}\right\} &= \hat{a}_2\\
\log\left\{\frac{\hat{p}_3/(1-\hat{p}_3)}{\hat{p}_1/(1-\hat{p}_1)}\right\} &= \hat{a}_3.
\end{aligned}$$

5. Fitting technique

The first thing to note about binary logistic regression is that we cannot fit the model

$$\log\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx^k$$

directly as all the y values are 0 or 1. Instead starting estimates are generated and an iterative procedure is used to find the maximum likelihood solution point and estimate the deviance and deviance residuals. In the

event that the data matrix is not of full rank this will be reported and the singular value decomposition will be used, so parameter estimates will still be generated but will not then be unique.

Note that, for standard logistic regression where w_i is the i'th binomial variable for m samples of size N_i, and not either 0 or 1 with a sample size of 1 as with binary logistic regression, the deviance is

$$\sum_{i=1}^{m} \text{dev}(w_i, \hat{\mu}_i) = 2 \sum_{i=1}^{m} \left\{ w_i \log\left(\frac{w_i}{\hat{\mu}_i}\right) + (N_i - w_i) \log\left(\frac{N_i - w_i}{N_i - \hat{\mu}_i}\right) \right\}$$

and the deviance residuals r_i are

$$r_i = \text{sign}(w_i - \hat{\mu}_i)\sqrt{\text{dev}(w_i, \hat{\mu}_i)}.$$

Of course these expressions are corrected for the extreme cases $w_i = 0$ or $w_i = N_i$ which will happen from time to time with standard logistic regression, but will happen all the time with binary logistic regression.

6. Conclusions

Binary logistic regression is widely used to analyze large data sets, sometimes even containing mixtures of qualitative and quantitative variables, and often in order to estimate Log Odds Ratios. It is incumbent upon users that any conclusions drawn about predicting probabilities are justified by taking account of the following suggestions.

- Add a constant term to the regression unless it is clear that $p = 0.5$ when all the covariates are zero.
- Scale all the variables to similar orders of magnitude prior to regression.
- Take care about the need with categorical variables to suppress a variable if a constant is fitted.
- Check the deviance, deviance residuals, and leverages to make sure the model gives a sensible fit.
- Do not ignore warnings if the rank is less than full or iteration has not converged.
- Only using parameters to estimate log odds ratio if the number of variables is small and that a unit change in a variable makes sense.
- Be careful not to confuse exponentials and logarithms to base e with powers of ten and logarithms to base 10.