Logistic Regression

Menu:	QCExpert	Regression	Logistic

Logistic regression assumes one or more real independent variables and a binary response variable with values 0 or 1, usually representing logical value like false/true, good/bad, etc. Alternatively, the response may have a form of frequency ratio in the interval <0, 1> in case of repeated measurements. This ratio should be the number of positive results divided by number of trials n_1/n at a given value of the independent variable. Logistic regression is then used to model probability of some event in dependence on the independent variables x. It is supposed that the response is a random variable with alternative distribution with parameter π which denotes the probability of a positive outcome of a trial. Thus, the number of positive outcomes out of a fixed number n of trials have a binomial distribution Binom(n, π). This parameter depends on x monotonously and logistic regression model will be an estimate of this dependence. Applications of logistic models are wide and include diverse fields of science and technology. Typically logistic models are used to estimate risks or failures under given conditions, bank credit scoring of a client, probability of survival of an organism in given environment, in toxicology, pharmaceutical, medicine, ecology, reliability analysis, market research, etc.

The probability π is modeled with a logistic model

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

or after rearrangement

$$\log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x \ .$$

The expression log($\pi(x)$ / (1 − $\pi(x)$)) is called logit. Parameters α and β are regression coefficients and their estimates a, b are computed with an iterative least squares methods. Such values of α a β are maximum likelihoods estimates. If x is univariate, logistic model may be plotted as a sigmoid-shape curve $\pi(x)$ describing the dependence of probability of a positive outcome on x. This model may then be used for prediction of the probability at any new value of x. The independent variable may be multivariate, $\mathbf{x} = (x_1, ..., x_m)$. Corresponding model for multivariate logistic regression can be expressed by

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + +\beta_m x_m)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + +\beta_m x_m)} \ .$$

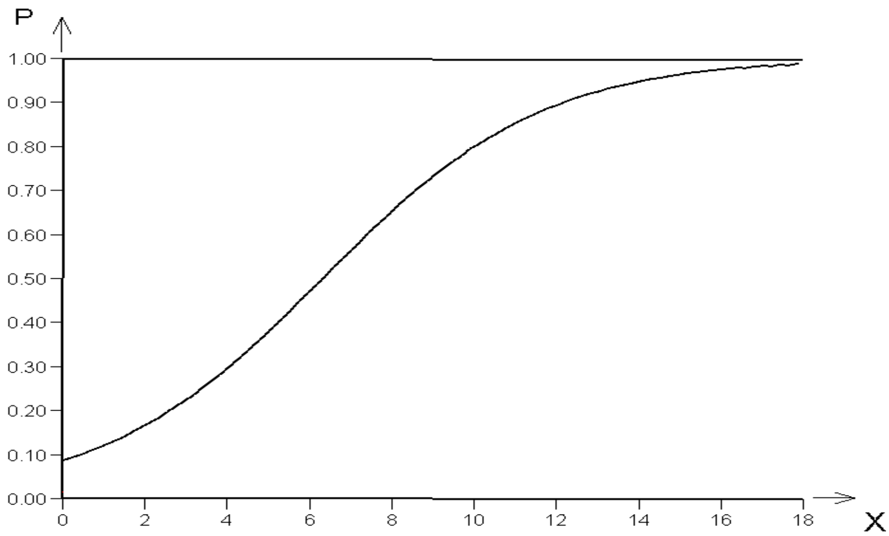


Fig. 1 Logistic regression model for one variable *x*

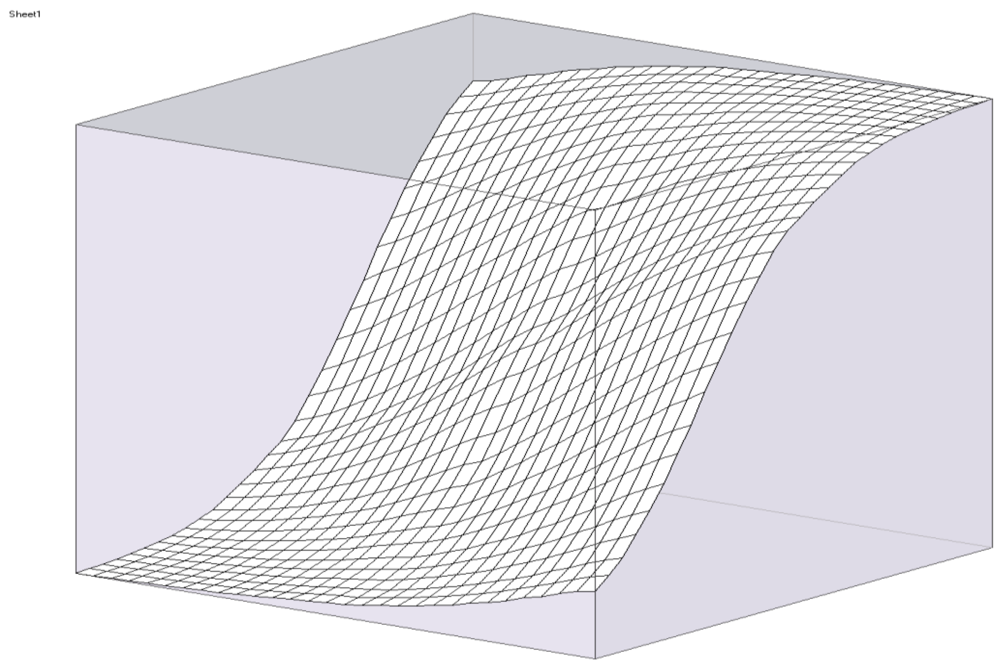


Fig. 2 Example of 3D logistic regression model for two variables x = (x_1, x_2)

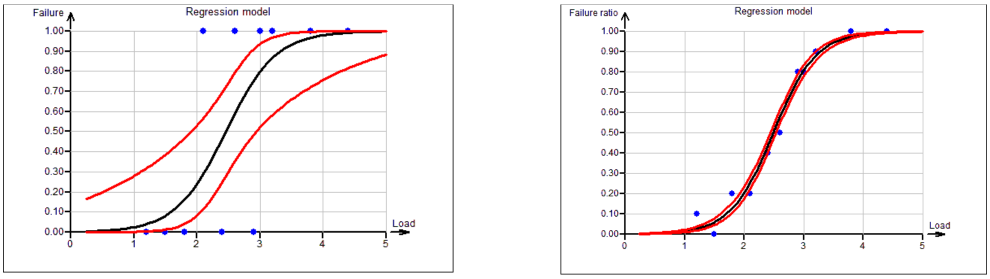
Computed model can be used to predict the probability of the positive outcome of the experiment of test based on any new user-defined values of the independent variable.

Data and parameters

Data for this module must contain at least two columns, one of which is the independent variable, second is the test outcome (0 or 1), or the fraction of positive outcomes from a set of trials performed independently on each other at the same value of *x*. Generally, it is recommended to use the original binary data instead of summary fractions. The choice of 0 or 1 to denote the outcome is arbitrary. The dependence of probability on *x* may be ascending as well as descending. Two example of data for logistic regression are given on Fig. 3.

Load	Failure
1.2	0
1.5	0
1.8	0
2.1	1
2.4	0
2.6	1
2.9	0
3	1
3.2	1
3.8	1
4.4	1

Load	Failure ratio
1.2	0.1
1.5	0
1.8	0.2
2.1	0.2
2.4	0.4
2.6	0.5
2.9	0.8
3	0.8
3.2	0.9
3.8	1
4.4	1



A. Single binary responses, 11 measurements B. Summary ratio data, 110 measurements

Fig. 3 Examples of data and resulting logistic curves

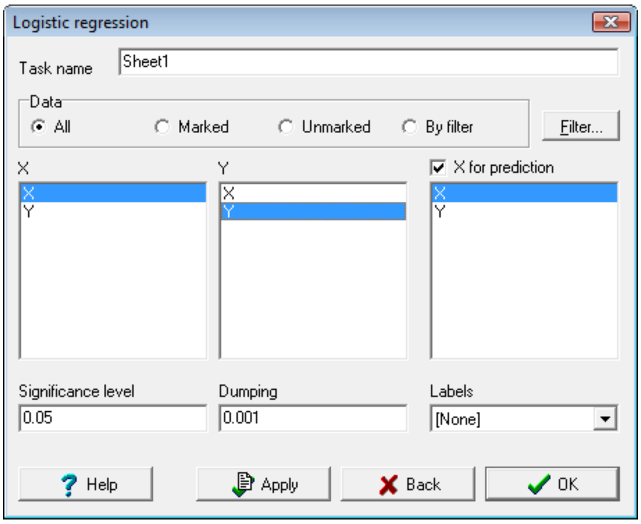


Fig. 4 Dialog box for logistic regression

In the dialog panel *Logistic regression* choose one or more dependent variables X and one independent variable Y. Values for prediction can be selected in the field *X for prediction*. These values can be identical with the independent variable column. Predicted probability including confidence intervals will be computed for each value for prediction.

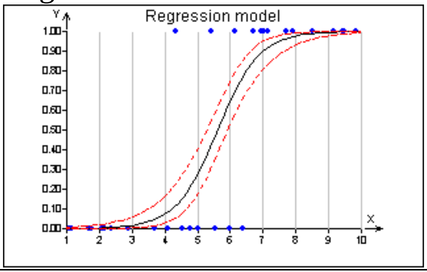
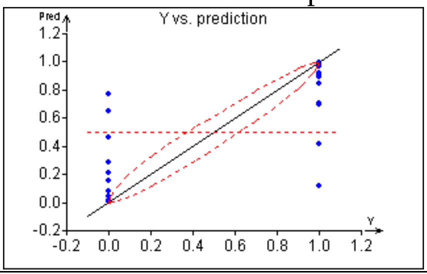
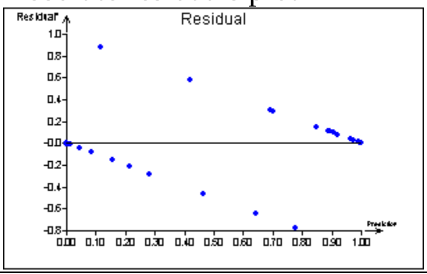
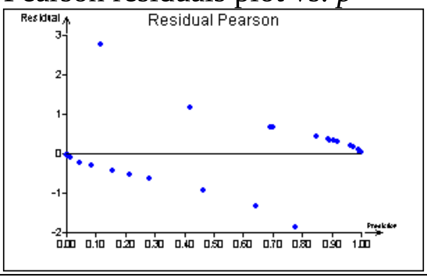
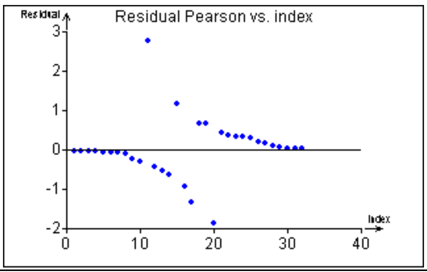
The number of variables and the variables independently for the prediction must be the same.

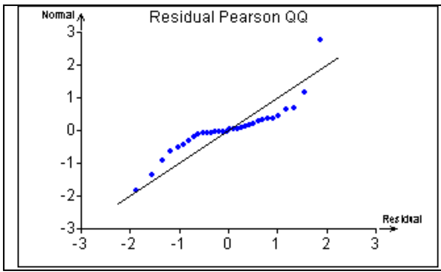
Protocol

No of cases	Number of rows used
Independent variables	List of independent variables
Dependent variables	Name of the dependent variable
No of iterations	The number of iterations used by the algorithm to calculate parameters by maximum likelihood method
Max likelihood	Logarithm of the maximal likelihood reached
Parameter estimates	Logistic model parameter values
Parameter	Parameter names taken from names of the independent variables, *Abs* stands for the absolute term
Estimate	Parameter estimates
Std Deviation	Parameter standard deviations
p-value	p-value gives the theoretical significance level at which the statistical significance of a parameter would just be rejected
Table of prediction	Table of predicted probability values for each value (or row) of the

	independent variable selected in the *X for prediction* field in the dialog box, see Fig. 4.
Variable name	Names of the independent variable (variables)
Prediction	Probability of an event as predicted by the logistic model
Lower limit	Lower confidence limit for the predicted probability
Upper limit	Upper confidence limit for the predicted probability

Graphs

Regression model 	Logistic regression curve with its confidence band and measured data. This plot is displayed only for a single independent variable *x*. The confidence band (the upper and lower red curve) defines a confidence interval at selected confidence level 1 – α (usually 0.95, i.e. 95%) of the predicted probability for each value of the independent variable. For higher precision it is advisable to zoom in the plot.
Predicted vs. measured plot 	Model versus data plot (predicted vs. observed). This plot is an analogy of the prediction plot in regression. Points far from the line y=x are not necessarily outliers.
Absolute residuals plot 	Graf of absolute residuals or a distance between data and the logistic curve. The residuals are of different nature than the residuals in a linear or non-linear regression. They do not have a normal distribution and their values are always between -1 and 1.
Pearson residuals plot vs. *p* 	Transformed Pearson's residuals are comparable with classic residuals such as in the linear regression. They have a standard normal distribution $N(\mu = 0, \sigma^2 = 1)$ and can be used for diagnosis of outlying measurement.
Pearson residuals vs. index 	The same residuals as in the previous plot. If the data are unordered, one can better observe the distribution of residuals than in the previous plots.
Pearson residual QQ-plot	QQ-graph of transformed residuals is an effective diagnostic tool for assessing the normality of residuals as though the data have

	alternative, or binomial distribution, the Poisson-transformed residuals should be normal. Points that are significantly apart from the line are suspected outlying points, and it is appropriate to verify their validity, or conclusions from the computed model should be drawn with with caution. Ideally, points are scattered around the designated line.