Logistic Regression

Logit Function and Logistic Function

To understand the concept of logistic regression it is useful to start by anchoring the discussion to the following basic linear regression model:

$$y = \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \theta_3 \cdot x_3 + \varepsilon \tag{1}$$

In this type of model, which is addressed in another document, the left-hand side is a scalar response quantity, y, and the right-hand side contains a collection of model parameters, $\boldsymbol{\theta}$, and regressors, $\mathbf{x}$, as well as the model error, ε. A nonlinear regression model is similar, except the model parameters, $\boldsymbol{\theta}$, appear in a nonlinear form on the right-hand side. Regardless of the formulation of the right-hand side of Eq. (1), logistic regression is unique in the formulation for the left-hand side. Instead of a response quantity, y, logistic regression aims at modelling a probability, p. For example, logistic regression can be employed to model the probability that a structural component is in a particular damage state. Because probability values must lie in the interval from zero to one, a special formulation is necessary in the left-hand side. Specifically, logistic regression employs the "logit function," or inversely the "logistic function," to ensure that $p \in [0,1]$. In mathematics, the logistic function

$$p = \frac{1}{1+e^{-y}} \tag{2}$$

varies between zero and unity as the argument, y, varies from minus to plus infinity. Solving Eq. (2) for y yields the logit function:

$$y = \ln\left(\frac{p}{1-p}\right) \tag{3}$$

Thus, a model of the form

$$\ln\left(\frac{p}{1-p}\right) = \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \theta_3 \cdot x_3 + \varepsilon \tag{4}$$

yields a model for the quantity $p \in [0,1]$. This is the essence of logistic regression; the right-hand side and the overall modelling procedure remains the same as in linear or nonlinear regression, while the quantity p is guaranteed to remain in the interval zero to unity. With reference to Eq. (4), the analyst would have observations for x_1, x_2, x_3, and p and carry out linear regression to assess θ_1, θ_2, θ_3, and ε, thus yielding a model for the probability p.

Data with Probabilities

(To be written)

Data with Categories

Suppose the observations are provided in the form shown in Table 1. In words, suppose the 0-1 indicator shows whether the event of interest occurred for the observed regressor values. For example, 1 indicates that the component was damaged, while 0 indicates that the component was undamaged. In that situation the likelihood function is:

$$L = p^{y} \cdot (1-p)^{1-y} \tag{5}$$

Notice that $L=p$ for observations $y=1$, and $L=1-p$ for observations $y=0$; this shows that the likelihood function is indeed proportional to the probability of making the observations.

Table 1: Observations

y	x_1	x_2	x_3	x_4
1	…	…	…	…
0	…	…	…	…
1	…	…	…	…
1	…	…	…	…
0	…	…	…	…
1	…	…	…	…