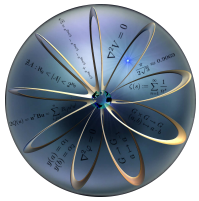

planetmath.org

Math for the people, by the people.

logistic regression

Canonical name	LogisticRegression
Date of creation	2013-03-22 14:47:51
Last modified on	2013-03-22 14:47:51
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	12
Author	CWoo (3771)
Entry type	Definition
Classification	msc 62J12
Classification	msc 62J02
Defines	logit
Defines	probit
Defines	complementary-log-log

Given a binary respose variable Y with probability of success p, the *logistic regression* is a non-linear regression model with the following model equation:

$$\mathrm{E}[Y] = \frac{\exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})},$$

where $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$ is the product of the transpose of the column matrix $\boldsymbol{X}$ of explanatory variables and the unknown column matrix $\boldsymbol{\beta}$ of regression coefficients. Rewriting this so that the right hand side is $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$, we arrive at a new equation

$$\ln\Big(\frac{\mathrm{E}[Y]}{1-\mathrm{E}[Y]}\Big) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}.$$

The left hand side of this new equation is known as the logit function, defined on the open unit interval $(0,1)$ with range the entire real line $\mathbb{R}$:

$$\mathrm{logit}(p) := \ln(\frac{p}{1-p}) \text{ where } p \in (0,1).$$

Note that the logit of p is the same as the natural log of the odds of success (over failures) with the probability of success $= p$. Since Y is a binary response variable, so it has a binomial distribution with parameter (probability of success) $p = \mathrm{E}[Y]$, the logistic regression model equation can be rewritten as

$$\mathrm{logit}\,\big(\mathrm{E}[Y]\big) = \mathrm{logit}(p) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}. \tag{1}$$

Logistic regression is a particular type of generalized linear model. In addition, the associated logit function is the most appropriate and natural choice for a link function. By natural we mean that $\mathrm{logit}(p)$ is equal to the natural parameter θ appearing in the distribution function for the GLM (generalized linear model). To see this, first note that the distribution function for a binomial random variable Y is

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

where n is the number of trials and $Y = y$ is the event that there are y success in these n trials. p, the parameter, is the probability of success. Let there be N iid binomial random variables $Y_1, Y_2, \ldots, Y_N$ each corresponding to n_i trials with p_i probability of success. Then the joint probability distribution of these N random variables is simply the product of the individual binomial

distributions. Equating this to the distribution for the GLM, which belongs to the exponential family of distributions, we have:

$$\prod_{i=1}^{N} \binom{n_i}{y_i} {p_i}^{y_i} (1-p_i)^{(n_i-y_i)} = \prod_{i=1}^{N} \exp\big[y_i\theta_i - b(\theta_i) + c(y_i)\big].$$

Taking the natural log on both sides, we have the equality of log-likelihood function in two different forms:

$$\sum_{i=1}^{N} \big[\ln \binom{n_i}{y_i} + y_i \ln p_i + (n_i - y_i)\ln(1-p_i)\big] = \sum_{i=1}^{N} \big[y_i\theta_i - b(\theta_i) + c(y_i)\big].$$

Rearranging the left hand side and comparing term i, we have

$$y_i \ln(\frac{p_i}{1-p_i}) + n_i \ln(1-p_i) + \ln \binom{n_i}{y_i} = y_i\theta_i - b(\theta_i) + c(y_i),$$

so that $\theta_i = \ln\big(p_i/(1-p_i)\big) = \text{logit}(p_i)$.

Next, setting the natural link function logit of the expected value of Y_i, which is p_i, to the linear portion of the GLM, we have

$$\text{logit}(p_i) = {\boldsymbol{X}_i}^{\mathrm{T}}\boldsymbol{\beta},$$

giving us the model formula for the logistic regression.

Remarks.

- Comparing model equation for the logistic regression to that of the normal or Gaussian linear regression model, we see that the difference is in the choice of link function. In normal liner model, the regression equation looks like

  $$\mathrm{E}[Y] = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}. \tag{2}$$

  The link function in this case is the identity function. The model equation is consistent because the linear terms on the right hand side allow $\mathrm{E}[Y]$ on the left hand side to vary over the reals. However, for a binary response variable, Equation (2) would not be appropriate as the left hand side is restricted to only within the unit interval, whereas the right hand side has the possibility of going outside of $(0, 1)$. Therefore, Equation (1) is more appropriate when we are dealing with a binary response data variable.

- The logit function is not the only choice of link function for the logistic regression. Other, "non-natural" link functions are available. Two such examples are the probit function, or the inverse cumulative normal distribution function $\Phi^{-1}(p)$ and the complimentary-log-log function $\ln(-\ln(1-p))$. Both of these functions map the open unit interval to $\mathbb{R}$.