Introduzione all'econometria - Master MEGAT[1]

Simple Logistic Regression

1 The model

Suppose that we perform N experiments. Form distances of 1 to 20 feet from the hoop a player shot the ball and recorded hit or miss. Setting $X =$ distance form the hoop, in the i-th experiment we observe $Y_i = 0$ or $Y_i = 1$. The the pdf of Y_i is given by:

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{1}$$

where: $y_i = 0, 1$ and π_i depend on x_i.

To model this dependency a popular choice is

$$\pi_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}. \tag{2}$$

that can be equivalently rewritten as

$$\log(\frac{\pi_i}{1-\pi_i}) = \alpha + \beta x_i. \tag{3}$$

We may have several Bernoulli trials for each selected value x_i, of the regressor variable. For instance, the player may shoot the ball M_i times from distance x_i. Let Y_i be the number of success out of the M_i trials conducted at X_i Assuming the independence of the individual Bernoulli trials, Y_i has a binomial distribution with M_i trials and success probability π_i.

2 Estimation

The simple logistic regression model consists of $N \geq 2$ independent random variables $Y_1, Y_2, ..., Y_n$ such that:

[1]Send comments/questions on this note to Carlo Ciccarelli: *carlo.ciccarelli@uniroma2.it* The tex file can be downloaded from `http://web.tiscali.it/cciccarelli`

1. $Y_i \sim$ binomial(M_i, π_i) for $i = 1, 2, ..., N$, where the $M_i \geq 1$ are fixed integers, and

2. $\text{logit}(\pi_i) = \log\frac{\pi_i}{1-\pi_i} = \alpha + \beta x_i$, for $i = 1, 2, ..., N$.

In this example all M_i are equal to 1.

In this model the conditional mean of Y, given $X = x$ is:

$$E[Y_i|x_i] = M_i\pi_i = M_i\frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}. \tag{4}$$

We are now going to find the maximum likelihood estimators of α and β. The log likelihood function is:

$$l(\alpha, \beta) = c + \sum_{i=1}^{n}[y_i \log \pi_i + (M_i - y_i)\log(1 - \pi_i)], \tag{5}$$

where c is a constant that does not depend on the unknown parameters.

Rearranging terms we have:

$$l(\alpha, \beta) = c + \sum_{i=1}^{n}[y_i(\alpha + \beta x_i) + (M_i)\log(1 - \pi_i)], \tag{6}$$

Therefore the elements of the score function are:

$$S_\alpha = \frac{\partial l(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^{n}(y_i - M_i\pi_i) \tag{7}$$

and

$$S_\beta = \frac{\partial l(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^{n} x_i(y_i - M_i\pi_i) \tag{8}$$

To find the maximum likelihood estimates we need to solve the equations $S_\alpha = 0$ and $S_\beta = 0$. For numerical maximization we may use the information matrix $I(\alpha, \beta)$.

$$I(\alpha, \beta) = \begin{bmatrix} \sum_{i=1}^{n} M_i\pi_i(1 - \pi_i) & \sum_{i=1}^{n} x_i M_i\pi_i(1 - \pi_i) \\ \sum_{i=1}^{n} x_i M_i\pi_i(1 - \pi_i) & \sum_{i=1}^{n} x_i^2 M_i\pi_i(1 - \pi_i) \end{bmatrix}$$

This matrix is positive definite provided that there are at least two different values x_i and provided that the value of π_i is not identically equal to 0 or 1 for $i = 1, 2, \ldots, n$

In this example:

1. $N = 20$

2. $M_i = 1$ for $i = 1, 2, ..., 20$

3. $x_i = i$ for $i = 1, 2, ..., 20$,

4. $y = (1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0)$.

The log-likelihood function take a unique maximum at $\hat{\alpha} = 2.271$ and $\hat{\beta} = -0.276$. Starting with $\alpha(0) = \beta(0) = 0$, the procedure converge in 5 iterations. The algorithm converges quite rapidly, due to the fact that the log-likelihood function is well approximated by a quadratic function in the neighborhood of the maximum.