

Introduction to Logistic Regression

Moo K. Chung

February 13, 2007

Update history: June 30, 2006, Jan 17, 2007.

Motivation: consider a mild cognition impairment (MCI) imaging study. The main research problem is if it is possible to predict the state of MCI purely based on imaging biomarkers such as cortical thickness. We can address this particular problem using a simple logistic regression analysis.

1 Logistic regression

Suppose p regressors $X_1, \dots, X_p$ are given. These are both imaging and non-imaging biomarkers such as local area, cortical thickness, gender, age, education level and memory test score. Let $x_{i1}, \dots, x_{ip}$ denote the measurements for the i -th subject. Let the response variable Y_i be the clinical state of the i -th subject modeled as

$$Y_i \sim \text{Bernoulli}(\pi_i).$$

$Y_i = 0, 1$ indicates the subject belongs to the elderly normal control (group 0) or mild cognition impairment group (group 1) respectively. π_i is the likelihood (probability) of a subject belong to the group 1, i.e. $\pi_i = P(Y_i = 1)$.

Now consider linear model

$$Y_i = \mathbf{x}_i' \beta + \epsilon_i,$$

where $\mathbf{x}_i' = (1, x_{i1}, \dots, x_{ip})$ and $\beta' = (\beta_0, \dots, \beta_p)$. We may assume $\mathbb{E}\epsilon_i = 0, \mathbb{V}\epsilon_j = \sigma^2$. In this case, the above linear model is no longer appropriate since

$$\mathbb{E}Y_j = \pi_i = \mathbf{x}_i' \beta$$

but $\mathbf{x}'_i\beta$ may not be in the range $[0, 1]$. This inconsistency is caused by trying to match continuous variables x_{ij} to categorical variable Y_i directly. To address this problem, we introduce the *logistic regression function* g :

$$\pi_i = g(x_i) = \frac{\exp(\mathbf{x}'_i\beta_i)}{1 + \exp(\mathbf{x}'_i\beta_i)}.$$

Then using the *logit function*, we can write this as

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} = \mathbf{x}'_i\beta_i.$$

The unknown parameters β are estimated via the maximum likelihood estimation (MLE). The likelihood function is

$$L(\beta|y_1, \dots, y_n) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \prod_{i=1}^n \left[\frac{\exp(\mathbf{x}'_i\beta_i)}{1 + \exp(\mathbf{x}'_i\beta_i)} \right]^{y_i} \prod_{i=1}^n \left[\frac{1}{1 + \exp(\mathbf{x}'_i\beta)} \right]^{1-y_i}.$$

The loglikelihood function is given by

$$\log L(\beta) = \text{const.} + \sum_{i=1}^n y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \quad (1)$$

$$= \text{const.} + \sum_{i=1}^n y_i \mathbf{x}'_i\beta + \log(1 - \pi_i) \quad (2)$$

and its maximum is obtained when

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^n \mathbf{x}_i (y_i - \pi_i) = 0.$$

In simplifying the expression, we used the following identities

$$\frac{\partial \pi_i}{\partial \beta_0} = \pi_i(1 - \pi_i)$$

and

$$\frac{\partial \pi_i}{\partial \beta_1} = x_i \pi_i(1 - \pi_i).$$

Since the logistic regression function π is in complicated form, the maximum is obtained numerically. Define the *information matrix* to be

$$I(\beta) = -\frac{\partial^2 \log L(\beta)}{\partial \beta' \partial \beta} - \sum_{i=1}^n \pi_i(1 - \pi_i) \mathbf{x}_i \mathbf{x}_i'.$$

Then the Newton-Raphson algorithm is used to find the MLE in an iterative fashion. Starting with an arbitrary initial vector β^0 , we estimate iteratively

$$\beta^{j+1} = \beta^j + I(\beta^j)^{-1} \frac{\partial \log L(\beta)}{\partial \beta}(\beta^j).$$

2 Statistical inference

Although we do not have the explicit formulas for the MLE, using the asymptotic normality of the MLE, the distributions of the estimators can be approximately determined. For large sample size n , the distribution of $\hat{\beta}$ is approximately multivariate normal with means β with the covariance matrix $I(\hat{\beta})^{-1}$.

Consider following full model:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Let $\beta^{(1)} = (\beta_0, \dots, \beta_q)'$ and $\beta^{(2)} = (\beta_{q+1}, \dots, \beta_p)'$. $\beta^{(1)}$, which corresponds to the parameters of the *reduced model*. Then we are interested in testing

$$H_0 : \beta^{(2)} = 0.$$

Define the *deviance* D of a model as $D = -2 \log L(\hat{\pi})$ which is distributed asymptotically as χ_{n-p-1}^2 . Let $\hat{\pi}^{(p)}$ and $\hat{\pi}^{(q)}$ be the estimated success probabilities for the full and reduced models, and let D_p and D_q be the associated deviances. Then the log-likelihood ratio statistic for testing $\beta^{(2)} = 0$ is

$$2[\log L(\hat{\pi}^{(p)}) - \log L(\hat{\pi}^{(q)})] = D_q - D_p \sim \chi_{p-q}^2.$$

Most statistical data analysis packages such as R and MATLAB has built-in routines for doing logistic regression.

3 Logistic discrimination

Discriminant analysis resulting from the estimated logistic model is called *logistic discrimination*. We classify the i -th subject according to a *classification rule*. The simplest rule is to assign the i -th subject as group 1:

$$P(Y_i = 1) > P(Y_i = 0).$$

This statement is equivalent to $\pi_i > 1/2$. Depending on the bias and the error of the estimation, the value $1/2$ can be adjusted. For the fitted logistic model, we classify the i -th subject as group 1 if $\mathbf{x}'_i\beta_i > 0$ and as 0 if $\mathbf{x}'_i\beta_i < 0$. The plane $\mathbf{x}'_i\beta = 0$ is the *classification boundary* that separates two groups. The performance of classification technique is measured by *error rate* γ , the overall probability of misclassification.

Cross-validation is used to estimate the error rate. This is done by randomly partitioning the data into the training set and the testing set. In the *leave-one-out* scheme, the training set consists of $n - 1$ subjects while the testing set consists of one subject. Suppose the i -th subject is taken as the test set. Then using the training set, we determine the logistic model. Using the predicted model, we test if the i -th subject is correctly classified. It is classified correctly if y_i

The error rate obtained in this fashion is denoted as e_{-i} . Note that $e_{-i} = 0$ if the subject is classified correctly while $e_{-i} = 1$ if the subject is misclassified. The *leave-one-out error rate* is then given by

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^n e_{-i}.$$

For reference on logistic regression, just open any textbook on multivariate analysis.