LOGISTIC REGRESSION IN DEPRESSION CLASSIFICATION

J. Kukal[1], *Q. V. Tran*[1], *M. Bareš*[2]

[1]KSE, FJFI, CVUT v Praze

[2]PCP, CNS, 3LF UK v Praze

Abstract

Well known logistic regression and the other binary response models can be used in the area of statistics, biomedical data analysis and artificial intelligence. Beginning with basic properties of proposed model, we use likelihood ratio testing as a tool for the pruning of complete model to produce the best sub-model. We are focused to optimum and automated pruning which is an alternative to traditional stepwise logistic regression. The general methodology of logistic regression with optimum pruning was developed together with a library of functions in the Matlab environment. The library was used to the analysis of neuropsychological and biomedical data related to the systematic research of resistant depressions. The aim of the paper is in the selection of optimum sub-model (set of patient properties) which will help with the decision whether the patient has resistant or non-resistant form of the depression.

1 Introduction

There are many statistical methods which seem to be also powerful tools in the area of artificial intelligence (cluster analysis, linear and nonlinear regression, kernel methods etc.). One of them is logistic regression [4] and its generalization to binary response index model [3]. The nonlinearities of these models are very similar to models of artificial neurons. So, the logistic regression and the testing of sub-models can be useful in designing of multilayer perceptron (MLP) networks. The model and sub-model difference can be useful for the decision, whether given input or a group of inputs plays significant role in hierarchical decision process inside ANN. Thus, statistical testing based on likelihood ratio (LR) can help to eliminate redundant weights (ANN pruning) or to generate hidden layer of MLP.

2 Binary Response Index Model

We suppose a model [3] with m real inputs $\boldsymbol{x}$ and single binary output y in the form

$$y = \mathrm{h}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + e) \tag{1}$$

where

$$\mathrm{h}(z) = \begin{cases} 1 & \text{for} \quad z > 0 \\ 0 & \text{for} \quad z \le 0 \end{cases} \tag{2}$$

is Heaviside's unit step function, $\boldsymbol{x}, \boldsymbol{\beta} \in \mathbf{R}^{m+1}$, $x_0 = 1$, e is a continuous random variable with positive and symmetric probability density function g(z) around zero, of course. Its cumulative distribution function is

$$\mathrm{G}(z) = \int_{-\infty}^{z} \mathrm{g}(u)\,\mathrm{d}u \tag{3}$$

So, the output y is also of stochastic nature and it can be described via probability

$$\mathrm{p}(\boldsymbol{x}) = \mathrm{prob}(y = 1 \mid \boldsymbol{x}) = \mathrm{prob}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + e > 0) = \mathrm{prob}(e > -\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) = 1 - \mathrm{G}(-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathrm{G}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) \tag{4}$$

which is well known formula [3, 4] for *binary model* related to *logistic regression*. There is necessary to satisfy $0 < G(z) < 1$ for all real arguments z. Fortunately, it implies from $g(z) > 0$ everywhere, which was declared above.

3 Special Cases of Binary Model

Binary model was first published by Bliss [1] in 1934 as *probit model* with nonlinearity

$$G(z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \tag{5}$$

and corresponding density function of standard normal distribution as

$$g(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \tag{6}$$

Lately in 1944, Berkson [2] published *logit model* (logistic regression) with nonlinearity

$$G(z) = \frac{1}{1+\exp(-z)} \tag{7}$$

and corresponding density function of logistic distribution as

$$g(z) = \frac{\exp(-z)}{(1+\exp(-z))^2} \tag{8}$$

This model is frequently used in many applications. The logit model approaches its asymptotes less rapidly then probit one. The other models are also possible to use. The density of Cauchy (t_1) distribution

$$g(z) = \frac{1}{\pi(1+z^2)} \tag{9}$$

generates the nonlinearity

$$G(z) = \frac{1}{2} + \frac{1}{\pi} \arctan z \tag{10}$$

with several amazing properties corresponding to heavy tail effect of Cauchy distribution. The variety of nonlinear characteristics can help us to choose a binary model and its parameters with the best possible quality of fitting.

4 Parameter Estimation

The estimation of model parameters is frequently performed via maximization of likelihood function or its logarithm respectively. Resulting point estimate (if exists) has a very good asymptotic properties [3, 4]. Let N be number of observations and $(\boldsymbol{x}_k, y_k)$ be individual observation for $k = 1, \ldots, N$. Thus, the density of y for individual $\boldsymbol{x}_k$ is

$$f(y \mid \boldsymbol{x}_k, \boldsymbol{\beta}) = \left[G(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{\beta})\right]^{y} \left[1 - G(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{\beta})\right]^{1-y} \tag{11}$$

The *logarithmic likelihood function* over all observations is defined as

$$L(\boldsymbol{\beta}) = \sum_{k=1}^{N} \left(y_k \log G(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{\beta}) + (1-y_k) \log(1 - G(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{\beta})) \right) \tag{12}$$

The point estimate $\boldsymbol{b}$ of parameter vector is obtainable via maximization of objective function L on closed convex domain **D** as

$$\boldsymbol{b} = \hat{\boldsymbol{\beta}} \in \arg\max_{\boldsymbol{\beta} \in \mathbf{D}} \mathrm{L}(\boldsymbol{\beta}) \tag{13}$$

The existence but not uniqueness of the estimate $\boldsymbol{b}$ is guaranteed in this case. Replacing **D** by opened set $\mathbf{R}^n$ will cause problems when the observations are separable, which is ideal case for classifier tuning but not for parameter estimation. The analysis of asymptotic variance begins with matrix

$$\boldsymbol{U} = \sum_{k=1}^{N} \frac{\mathrm{g}(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{b})\boldsymbol{x}_k\boldsymbol{x}_k^{\mathrm{T}}}{\mathrm{G}(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{b})(1-\mathrm{G}(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{b}))} \tag{14}$$

When the matrix $\boldsymbol{U}$ is regular, it is also positive definite and the asymptotic variance of estimate $\boldsymbol{b}$ is

$$\mathrm{Avar}(\boldsymbol{b}) = \boldsymbol{V} = \boldsymbol{U}^{-1} \tag{15}$$

The asymptotic standard error of estimate $\boldsymbol{b}$ is

$$\mathrm{Astd}(\boldsymbol{b}) = \boldsymbol{s} = \mathrm{diag}(\boldsymbol{V})^{1/2} \tag{16}$$

and corresponding approximate 95% CI (confidence interval) is

$$\boldsymbol{\beta} \in \langle \boldsymbol{b} - 1.96\boldsymbol{s}, \boldsymbol{b} + 1.96\boldsymbol{s} \rangle \tag{17}$$

The confidence interval is rather useful for final report then for significance testing due to its sensitivity to singularity of matrix $\boldsymbol{U}$.

5 Hypothesis Testing

There are many approaches to binary model testing. One of them is likelihood ratio (LR) test. It is based on the comparison of given model and its sub-models. Let

$$\boldsymbol{r} = (1, r_1, ..., r_m)^{\mathrm{T}} \in \{0,1\}^{m+1} \tag{18}$$

be *selection vector*, which describes, whether adequate components of vector $\boldsymbol{x}$ will be used in the model. The vector $\boldsymbol{r}$ decomposes the vector $\boldsymbol{x}$ to two parts: active input vector $\boldsymbol{u} \in \mathbf{R}^{K+1}$ and eliminated input vector $\boldsymbol{v} \in \mathbf{R}^{Q}$. Here, K is the number of real inputs (excluding x_0), Q is the number of eliminated inputs, $K, Q \in \mathbf{N}_0$, $K + Q = m$. Analogical decomposition of parameter vector $\boldsymbol{\beta}$ comes to vectors $\boldsymbol{\mu} \in \mathbf{R}^{K+1}$, $\boldsymbol{\eta} \in \mathbf{R}^{Q}$.

When $Q = 0$, then $K = m$ and we obtain *complete model* as

$$\mathrm{p}(\boldsymbol{x}) = \mathrm{G}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) = \mathrm{G}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\eta}) \tag{19}$$

with optimum log-likelihood value L_{compl}.

Anyway, $Q > 0$, then $K < m$ and we obtain *restricted model* as sub-model in the form

$$\mathrm{p}(\boldsymbol{u}) = \mathrm{G}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\mu}) \tag{20}$$

with optimum log-likelihood value L_{restr}.

The traditional *likelihood ratio test* (LR-test) is about hypothesis $\mathrm{H}_0 : \boldsymbol{\eta} = \mathbf{0}$ against alternative $\mathrm{H}_\mathrm{A} : \boldsymbol{\eta} \neq \mathbf{0}$.

The testing criterion

$$LR = 2(L_{\mathrm{compl}} - L_{\mathrm{restr}}) \tag{21}$$

has limiting distribution χ^2_Q with Q degrees of freedom. The adequate p_{value} is directly obtained as

$$p_{\text{value}} = 1 - \mathrm{F}_Q(LR) \tag{22}$$

where F_Q is cumulative distribution function of chi-squared distribution.

The comparison of the model and its sub-model via LR-test brings a view on the significance of the model and its parameters. There are two useful particular cases.

When $K = 0$, then $Q = m$ and we obtain *constant model* as

$$\mathrm{p}(\boldsymbol{u}) = \mathrm{G}(\mu_0) = p_{\mathrm{c}} \tag{23}$$

Maximum likelihood estimation procedure (13) comes to trivial estimate

$$p_{\mathrm{c}} = \frac{1}{N}\sum_{k=1}^{N} y_k \tag{24}$$

Log-likelihood value of this constant model is

$$L_{\text{const}} = N\big(p_{\mathrm{c}} \log p_{\mathrm{c}} + (1 - p_{\mathrm{c}}) \log(1 - p_{\mathrm{c}})\big) \tag{25}$$

Comparing complete model with constant one, we can measure the model significance (its quality) via probability

$$p_0 = 1 - \mathrm{F}_m(2(L_{\text{compl}} - L_{\text{const}})) \tag{26}$$

Lower value of p_0 indicates the higher significance of given complete model and various models can be ordered according to p_0.

In the second case, we compare the complete model and a sub-model with eliminated k^{th} input, where $k > 0$. So, we have $K = m - 1$, $Q = 1$ and adequate log-likelihood value can be denoted as L_k. We can easily measure the significance of k^{th} input, in the meaning of hypothesis $\mathrm{H}_0 : \beta_k = 0$ against alternative $\mathrm{H}_\mathrm{A} : \beta_k \neq 0$, via probability

$$p_k = 1 - \mathrm{F}_1(2(L_{\text{compl}} - L_k)) \tag{27}$$

Stepwise strategy of parameter elimination is frequently used to obtain the model, which every parameter is significant in the meaning of $p_k < \alpha$, together with $p_0 < \alpha$. If the model is not unique, we prefer the model with minimum value of p_0.

6 Biomedical Application

The theory of optimum logistics sub-model came to the realization of library in the Matlab environment. Then we used the library to the analysis of biomedical data about resistant depressions. The data consists of 114 patients with 93 properties. Three properties are output ones (positive response to treatment of two weeks, four weeks and total response). Seven properties are based on EEG cordance measurement which is very efficient indicator for responder-resistant decision. The rest of 83 properties are basic, biochemical, psychological and psychiatric characteristics of patients. The set of 114 patients consists of 89 total responders and 25 resistants and it was analyzed via logistic regression with optimum pruning technique of the best sub-model finding. The best sub-model is mainly oriented to the cordance based properties and will be used for patient classification.

7 Conclusion

Logistic regression was employed to form the best sub-models in the case of logistic regression. The general methodology of logistic regression with optimum pruning was developed together with a library of functions in the Matlab environment. The library was used to the analysis of neuropsychological and biomedical data related to the systematic research of resistant depressions. The research will help with the decision whether the patient has resistant or non-resistant form of the depression.

Acknowledgement

The paper was created under the support of grant OHK4 - 165/11 CTU in Prague.

References

[1] Bliss C.I., The method of probits, *Science*, Vol.79, No.2037, 1934, pp. 38–39.

[2] Berkson J., Application of the logistic function to bio-assay, *J Am Stat Assoc*, Vol.39, 1944, pp. 357–365.

[3] Wooldridge J. M., *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press, 2002

[4] Hilbe J. M., *Logistic Regression Models*, Chapman & Hall/CRC Press, 2009

Author contacts:

jaromir.kukal@fjfi.cvut.cz

tran@vse.cz

bares@pcp.lf3.cuni.cz