Analysis of 1:M Conditional Logistic Regression Modelling Method

Changmin Zhang, Zhihong Man, Thong Nguyen

Department of Electrical Engineering & Computer Science
University of Tasmania, GPO Box 252-65 Hobart 7001, Australia
Email: czhang@postoffice.sandybay.utas.edu.au

Abstract Logistic regression model is a nonlinear regression system. It has rapidly become a widely used multivariate statistical method. In this paper, we analyze the 1:M conditional logistic regression modelling equation, and use Newton-Raphson iteration algorithm to obtain the optimized value of the parameter estimates. Also, a flow-chart of the program is provided.

1. Introduction

In 1838, Verhulet developed a simple logistic equation for growing population. Later, Volterra , Pearl and Reed developed the life processing model for predicting analysis of biological growth, (from the womb to the tomb for the population). Recently personal computers have become cheaper and more compact. PC software packages (e.g. SAS, SPSS/PC+ etc.) have been developed for statistical modellings. During the 1970's~1990's, logistic regression modelling techniques not only have been successfully applied to the biological forecasting, chronic disease etiological research and setting diagnosis criterions, but also have becoming widely used for various fields, such as weather & forecasting, artificial intelligent, ecological, economic, and socioeconomic systems. Today, its theory and applications are still being further investigated. In this paper, we analyze the 1:M conditional logistic regression modelling equation and use Newton-Raphson iteration algorithm to obtain optimized values for the parameter estimates.

2. 1:M Conditional logistic Regression Modelling

Suppose that the i^{th} element of I matched sets contains M_i controls in addition to the case. Let $X_{i0} = (x_{i01},\dots,x_{i0k})$ be the K-vector of exposure for the case in this set and $X_{ij} = (x_{ij1},\cdots,x_{ijk})$ be the exposure vector for the j^{th} control $(j=1,\cdots,M_i)$, where x_{ijk} represents the value of the K^{th} exposure variable for the case (j=0) or j^{th} control in the i^{th} matched set. We may then write the conditional likelihood in the form (Liddell, McDonald & Thomas, 1977; Breslow et al., 1978):

$$\prod_{i=1}^{I}\frac{\exp(\sum_{k=1}^{K}\beta_k x_{i0k})}{\sum_{j=0}^{M_i}\exp(\sum_{k=1}^{K}\beta_k x_{ijk})}$$

$$=\prod_{i=1}^{I}\frac{1}{1+\sum_{j=1}^{M_i}\exp(\sum_{k=1}^{K}\beta_k(x_{ijk}-x_{i0k}))} \quad (1)$$

Based on conditional probability, exposure risk factors probability for case may be defined as

$$P(x_{i0}\mid y=1)=P(y=1\mid x_{i0})\cdot P(x_{i0})/P(y=1)$$

Exposure probability of risk factor for control, may be written as

$$P(x_{ij}\mid y=0)=P(y=0\mid x_{ij})\cdot P(x_{ij})/P(y=0)$$

Logistic function is defined as

$$P(y=1\mid x_{i0})=\frac{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})}{1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})} \qquad (2)$$

$$P(y=0\mid x_{ij})=1-P(y=1\mid x_{ij})=\frac{1}{1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})} \qquad (3)$$

where α represents the log odds of disease risk for a person with a standard **(x=0)** set of regression variables, while $\exp(\beta_k)$ is the fraction by which this is increased (or decreased) for every unit change in x_{ijk}.

Using the Bayesian formula, we have

$$P(x_1,\cdots,x_n)=\frac{\frac{P(y=1\mid x_{i0})\cdot P(x_{i0})}{P(y=1)}\cdot\prod_{j=1}^{T_i}\frac{P(y=0\mid x_{ij})\cdot P(x_{ij})}{P(y=0)}}{\sum_{j=0}^{T_i}\frac{P(y=1\mid x_{ij})\cdot P(x_{ij})}{P(y=1)}\cdot\prod_{j'\neq j=0}^{T_i}\frac{P(y=0\mid x_{ij'})\cdot P(x_{ij'})}{P(y=0)}}$$

$$=\frac{P(y=1\mid x_{i0})\cdot\prod_{j=1}^{T_i}P(y=0\mid x_{ij})\cdot\frac{\prod_{j=0}^{T_i}P(x_{ij})}{P(y=1)\cdot\{P(y=0)\}^{T_i}}}{\{\sum_{j=0}^{T_i}P(y=1\mid x_{ij})\cdot\prod_{j'\neq j=0}^{T_i}P(y=0\mid x_{ij'})\}\cdot\frac{\prod_{j=0}^{T_i}P(x_{ij})}{P(y=1)\cdot\{P(y=0)\}^{T_i}}} \qquad (4)$$

Using expressions (2) and (3), expression (4) can be written as

$$=\frac{\frac{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})}{1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})}\cdot\prod_{j=1}^{T_i}\frac{1}{1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})}}{\sum_{j=0}^{T_i}\frac{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})}{1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})}\cdot\prod_{j'\neq j=0}^{T_i}\frac{1}{1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ij'k})}}$$

$$=\frac{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})\cdot\frac{1}{[1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})]\cdot\prod_{j=1}^{T_i}[1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})]}}{\sum_{j=0}^{T_i}\{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})\cdot\frac{1}{[1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})]\cdot\prod_{j'\neq j=0}^{T_i}[1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ij'k})]}\}}$$

$$=\frac{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})\cdot\frac{1}{\prod_{j=0}^{T_i}[1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})]}}{\sum_{j=0}^{T_i}\{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})\cdot\frac{1}{\prod_{j=0}^{T_i}[1+\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})]}\}}$$

$$=\frac{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})}{\sum_{j=0}^{T_i}[\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})]} \qquad (5\text{-a})$$

$$=\frac{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})}{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})+\sum_{j=1}^{T_i}\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})} \qquad (5\text{-b})$$

If the numerator and denominator in expression (5-b) are simultaneously divided by,

$$\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})$$

then, expression (5-b) has the following form

$$\frac{1}{1+\sum_{j=1}^{T_i}\exp[(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})-(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})]}$$

$$=\frac{1}{1+\sum_{j=1}^{T_i}\exp[\sum_{k=1}^{p}\beta_k(x_{ijk}-x_{i0k})]} \qquad (6)$$

In fitting observation data of fields, we need to observe case-control of N using equations (5-a), (6). The value of conditional Likelihood function can then be written as

$$L^*=\prod_{i=1}^{N}\frac{\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{i0k})}{\sum_{j=0}^{T_i}[\exp(\alpha+\sum_{k=1}^{p}\beta_k x_{ijk})]} \qquad (7)$$

$$L^*=\prod_{i=1}^{N}\frac{1}{1+\sum_{j=1}^{T_i}\exp[\sum_{k=1}^{p}\beta_k(x_{ijk}-x_{i0k})]} \qquad (8)$$

The first and second order partial derivatives of expression (7) can be expressed as

$$\frac{\partial \ln L^*}{\partial \beta_k} = \sum_{k=1}^{p} x_{i0k} - \sum_{i=1}^{N} \frac{\sum_{j=0}^{T_i} [\exp(\sum_{k=1}^{p} \beta_k x_{ijk})] x_{ijk}}{\sum_{j=0}^{T_i} \exp(\sum_{k=1}^{p} \beta_k x_{ijk})} \quad (9)$$

$$\frac{\partial^2 \ln L^*}{\partial \beta_k \partial \beta_{k'}} = -\sum_{i=1}^{N} \{ \frac{[\sum_{j=0}^{T_i} \exp(\sum_{k=1}^{p} \beta_k x_{ijk})][\sum_{j=0}^{T_i} [\exp(\sum_{k=1}^{p} \beta_k x_{ijk})] x_{ijk} \cdot x_{ijk'}]}{[\sum_{j=0}^{T_i} \exp(\sum_{k=1}^{p} \beta_k x_{ijk})]^2}$$

$$\frac{[\sum_{j=0}^{T_i} [\exp(\sum_{k=1}^{p} \beta_k x_{ijk})] x_{ijk}] \cdot [\sum_{j=0}^{T_i} [\exp(\sum_{k=1}^{p} \beta_k x_{ijk})] x_{ijk'}]}{[\sum_{j=0}^{T_i} \exp(\sum_{k=1}^{p} \beta_k x_{ijk})]^2} \} \quad (10)$$

3. Newton-Raphson (NR) Method

The Newton-Raphson method is a powerful method of solving non-linear algebraic equations. It works faster and is sure to converge in most casts as compared to the Gauss-Seidel (GS) method. The convergence can be speeded up considerably by performing the first iteration through the GS method and using the values obtained for starting the NR iterations.

Before explaining how the NR method is applied to solve the optimum value of parameter estimate problem, we briefly review the Newton-Raphson method.

Consider a set of p non-linear algebraic Equations

$L_i(\beta_1, \beta_2, \cdots, \beta_p) = 0;$

i=1, 2, ..., p (11)

Assume initial values $\beta_1^0, \beta_2^0, \cdots, \beta_p^0$ are unknown. Let $\Delta\beta_1^0, \Delta\beta_2^0, \cdots, \Delta\beta_p^0$ be the corrections, which added to the initial guess, give the actual solution as follow

$L_i(\beta_1^0 + \Delta\beta_1^0, \beta_2^0 + \Delta\beta_2^0, \cdots, \beta_p^0 + \Delta\beta_p^0) = 0;$

i=1, 2, ..., p (12)

Expanding the above equations using Taylor series around the initial guess, we have

$$L_i(\beta_1^0, \beta_2^0, ..., \beta_p^0) + [(\frac{\partial \ln L_i^*}{\partial \beta_1})^0 \Delta\beta_1^0 + (\frac{\partial \ln L_i^*}{\partial \beta_2})^0 \Delta\beta_2^0 + \cdots + (\frac{\partial \ln L_i^*}{\partial \beta_p}) \Delta\beta_p^0] +$$

higher order terms =0 (13)

the Higher order terms in expressions (13) can write in the following matrix form

$$\begin{pmatrix} L_1^0 \\ L_2^0 \\ \vdots \\ L_p^0 \end{pmatrix} + \begin{pmatrix} (\frac{\partial^2 \ln L^*}{\partial \beta_1^2})^0 & (\frac{\partial^2 \ln L^*}{\partial \beta_1 \partial \beta_2})^0 & \cdots & (\frac{\partial^2 \ln L^*}{\partial \beta_1 \partial \beta_p})^0 \\ (\frac{\partial^2 \ln L^*}{\partial \beta_2 \partial \beta_1})^0 & (\frac{\partial^2 \ln L^*}{\partial \beta_2^2})^0 & \cdots & (\frac{\partial^2 \ln L^*}{\partial \beta_2 \partial \beta_p})^0 \\ \vdots & \vdots & \vdots & \vdots \\ (\frac{\partial^2 \ln L^*}{\partial \beta_p \partial \beta_1})^0 & (\frac{\partial^2 \ln L^*}{\partial \beta_p \partial \beta_2})^0 & \cdots & (\frac{\partial^2 \ln L^*}{\partial \beta_p^2})^0 \end{pmatrix} \begin{pmatrix} \Delta\beta_1^0 \\ \Delta\beta_2^0 \\ \vdots \\ \Delta\beta_p^0 \end{pmatrix},$$

$$\cong \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (14)$$

Or in vector matrix form

$$L^0 + J^0 \Delta\beta^0 \cong 0 \quad (15)$$

where J^0 is known as the Jacobian matrix (obtained by differentiating the function vector L^* with respect to β and evaluating it at β^0). Equation (15) can be written as

$$L^0 \approx [-j^0] \Delta\beta^0 \quad (16)$$

Approximated values of corrections $\Delta\beta^0$ can be obtained from expression (16). These being a set of linear algebraic equations can be solved efficiently by triangularization and back substitution .

Updated values of β are then

$\beta^1 = \beta^0 + \Delta\beta^0$

or, in general , for the (r+1)th iteration

$$\beta^{(r+1)} = \beta^{(r)} + \Delta\beta^{(r)} \quad (17)$$

The iterations are continued till equation (11) is satisfied to an arbitrarily desired accuracy, i.e

$\left|L_i(\beta^{(r)})\right| < \xi$ (a specified value);

i=1,2,..., p (18)

4. General Definition of the Relative Risk (RR)

So far the 1:M conditional logistic model has been used solely as a means of relating disease probabilities to one or more categorical risk factors whose levels are represented by indicator variables. More generally the model relates a dichotomous outcome variable y which, in this paper, denotes whether (y=1) or not (y=0) the individual develops the disease during the study period, to a series of K regression variables via the equation (2), (3).

This formulation implies that the Relative Risk (RR) for individuals having two different sets X^* and X of risk variables is

$$RR=\frac{P(X^*)\{1-P(X)\}}{P(X)\{1-P(X^*)\}}=\exp\{\sum_{k=1}^{K}\beta_k(X_k^*-X_k)\} \qquad (19)$$

Here the ratio of incidence rates for individuals with exposures X * and X is given exactly by the equation (19). This approach has the conceptual advantage of eliminating the Odds Ratio (OR) approximation altogether, and thus obviates the rare disease assumption.

5. Criteria for Assessing Model Fit

Suppose the model contains S explanatory variables. Let y_i be the response value of the j^{th} observation. The estimate $\hat{p}_j$ of $p_j = P(Y_j = y_j)$ is obtained by replacing the regression coefficients by the maximum likelihood estimates (MLEs). The three criteria used by the SAS logistic procedure are calculated as follows:

(i) -2Log Likelihood

$$-2\ \mathrm{Log}\ L= -2\sum_j W_j \log(\hat{p}_j) \qquad (20)$$

where W_j is the weight of the j^{th} observation.

(ii) Akaike Information Criteria (AIC)

$$\mathrm{AIC}= -2\mathrm{LogL}+2(K+S) \qquad (21)$$

where K is the number of ordered values for the response and S is the number of explanatory variables.

(iii) Schwartz Criterion

$$\mathrm{SC}= -2\mathrm{LogL}+(K+S)\log(N) \qquad (22)$$

where K and S are defined as above, and N is the total number of observations (for the actual model syntax) or the total number of trials (for the events/trials model syntax).

The –2Log Likelihood statistic has a Chi-square distribution under the null hypothesis (that all the explanatory variables in the model are zero), and the SAS, SPSS/PC+ etc software package procedure prints a p-value for this statistic.

The AIC and SC statistic give two different ways of adjusting the –2Log Likelihood statistic for the number of terms in the model and the number of observations used. These statistics should be used when comparing different models for the same data, for example, when reader use the SELECTION=STEPWISE option in the MODEL Statement of SAS package; Lower values of the statistic indicate a more desirable model.

6. Program Flow-Chart

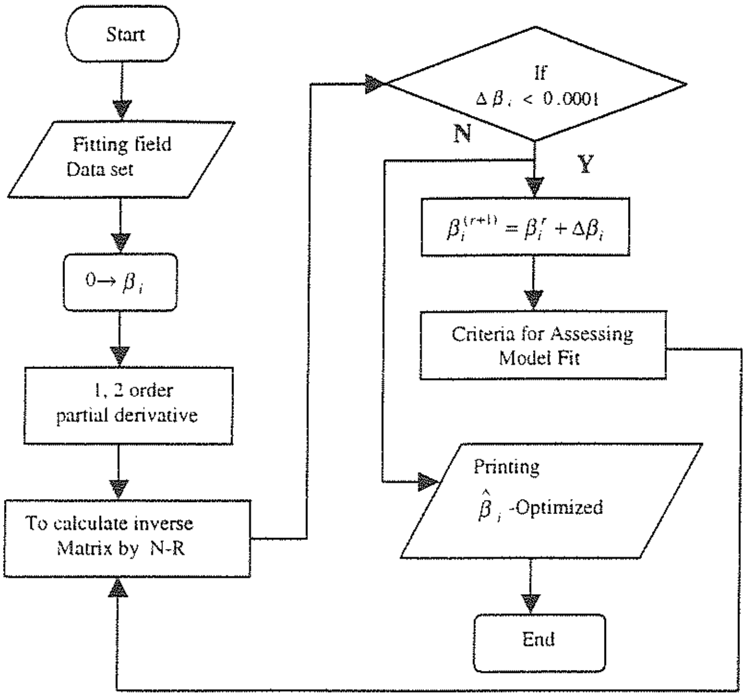
Start
Fitting field Data set
$0 \rightarrow \beta_i$
1, 2 order partial derivative
To calculate inverse Matrix by N-R
If $\Delta\beta_i < 0.0001$
N
Y
$\beta_i^{(r+1)} = \beta_i^r + \Delta\beta_i$
Criteria for Assessing Model Fit
Printing $\hat{\beta}_i$ -Optimized
End

Figure 1. 1:M Conditional Logistic Regression Model Program Flow-Chart

7. Conclusion

In this paper, we have analyzed the 1:M conditional logistic regression modelling method and used Newton-Raphson iteration algorithm to obtain the estimate values (β_i) of parameter-optimized. Although equations (7) and (8) are equivalent, we have compared them and recorded their iterations and convergence time, respectively, we found that equation (7) has faster convergence than (8) when studied for fitting field data. Another, the bias of risk variables confounding can not only be well overcome by the 1:M conditional logistic model, but also it has widely adapting range of matching field data, such as binary response variables (for example, success, failure) and ordinal response variables (for example, none, mild, severe) and continuous data. Meanwhile, we still introduce the Relative Risk (RR) (see equation (19)) . A consequence of equation (1) is that the RR associated with the risk factors under study are constant over strata. By including such interaction terms among the X's, one may model changes in the relative risk which accompany changes in the stratification variables. The fact that the parameters of the logistic model are so easily interpretable in terms of RR. We have introduced a new methodology progress in applied statistics which is criteria for assessing model fit, as mentioned above is one of the main reasons for using the model.

References

Chen, Q.G., A Simple Method of logistic Curve Calculating and Matching, *J. Railway Medicine. P.R.China.*, 10(3), 160-165, 1989.

Billet, J. Delisi M. Smith B G. Gates C., Use

Of Regression Techniques to Predict Hail Size and the Probability of Large Hail Size, ***Weather & Forecasting,*** 12(1), 154-164, 1997.

Tokunaga, R. matsumoto T., A Nonlinear Prediction Technique for Parametrized Families of Chotic Dynamics, ***International J. of Intelligent System***, 12(4), 291-309,1997.

Scott, A.J. Wild C.J., Fitting Regression Models to Case-Control data by Maximum Likelihood, ***Biometrika,*** 84(1), 57-71, 1997.

Breslow, N.E. & N.E. Day., Statistical Methods in Cancer Research, ***IARC Scientific Publication, No.32, Lyon France, Sixth Reimpression***, 248-334, 1994.

Nagrath, I.J. D.P.Kothari., Modern Power System Analysis, ***Tata McGraw-Hill Publishing Company Limited, New Delhi, Second Edition***, 183-191, 1989.

Rollin, B., Digesting Logistic Regression Results, ***The American Statistician***, 50(2), 117-119, 1996.

Ioannis, K. Argyros., Concerning the Convergence of Inexact Newton Method, ***J. of computational and Applied Mathematics***, 79, 235-247, 1997.

Saratchandran, PP. Nandakumaran, VM. Ambika G., Dynamics of the Logistic Map Under Discrete Parametric Perturbation, ***Pramana-Journal of Physics***, 47(5), 339-345, 1996.

Loesche, WJ. Taylor G. Giordano J. Hutchinson R. Rau CF. Chen YM. Schork MA., A Logistic Regression Model for the Decision to Perform Access Surgery, ***J. of Clinical Periodontology***, 24(3), 171-179, 1997.

Hsu, JSJ. Leonard T., Hierarchical Bayesian Semiparametric Procedures for Logistic Regression, ***Biometrika,*** 84(1), 85-93, 1997.

Jefferson, MF. Pendleton N. Lucas SB. Horan MA., Comparison of A Genetic Algorithm Neural Network with Logistic Regression for Predicting Outcome After Surgery for Patients with Nonsmall Cell Lung Carcinoma, ***Cancer,*** 79(7), 1338-1342, 1997.

SAS/STAT, User's Guide, Volume 2, ***Version 6, fourth Edition***, SAS Institute Inc, Cray, NC 27513, 1088-1089, 1990.

Hosmer, D.W, Jr. and Lemeshow, S., Applied Logistic Regression, New York, ***John Wiley & Sons, Inc.,*** 1989.