Linear programming for detecting separation in polytomous logistic regression

Inácio Andruski-Guimarães
UTFPR - Universidade Tecnológica Federal do Paraná, Curitiba, Brasil - andruski@utfpr.edu.br

Abstract

The parameter estimation in logistic regression model is known to be dependent on the data configuration. While the logistic model work well for many situations, may not work when the groups of observations are completely separated. Separation is a common problem in the logistic regression. Mathematical Programming approaches have been used for detecting separated data in logistic regression, but most of these researches have focused on the two group problem. In this paper we propose a linear programming formulation to detect separation in polytomous logistic regression. The proposed approach classifies data as completely separated, quasi-separated or overlapped, and can be used as part of the parameter estimation. Comparison with other methods, using different data sets taken from the literature, shows that our formulation may suggest an efficient alternative to mathematical programming approaches for the multiple-group problem.

Keywords: Linear Programming; Polytomous Logistic Regression.

1. Introduction

The logistic regression model is known as a powerful method widely applied to model the relationship between a categorical - or ordinal - dependent variable and a set of explanatory variables, or covariates, that may be either continuous or discrete. The accuracy of the Logistic Regression Model has been reported in many studies involving bankruptcy prediction, marketing applications and cancer classification, among others applications. However, the parameter estimation in logistic regression model is known to be dependent on the data configuration. While the logistic regression model work well for many situations, may not work when the data set has no overlapping. In order to way out the problem that arises when the data set has no overlapping, Rousseeuw and Christmann (2003) proposes the Hidden Logistic Regression Model (HLR). Another approaches to deal with separation can be found in Heinze and Schemper (2002), for binary response, and Andruski-Guimarães and Chaves-Neto (2009), for polytomous response, to name just a few.

Mathematical Programming approaches have been used for detecting separated data in discriminant analysis, but almost all researches have focused on the two group problem. An algebraic approach, suggested by Albert and Anderson (1984), uses ideas of linear programming and specifies the necessary constraints, but not an objective function. A mixed integer linear program, presented by Santner and Duffy (1986), determines whether data is separated or overlapped. Silvapulle and Burridge (1986) uses linear programming to check the necessary conditions for the existence of a finite maximum likelihood estimate for the logistic model. A single linear programming formulation, proposed by Bennett and Mangasarian (1992), generates a plan that minimizes an average sum of misclassified points belonging of two disjoint point sets in n-dimensional real space. An algorithm proposed by Konis (2007) is based on a linear program with a nonnegative objective function that has a positive optimal value when separation is detected.

In this paper we propose a linear programming formulation to detect separation in polytomous logistic regression. The proposed approach classifies data as completely separated, quasi-separated or overlapped, and can be used as part of the parameter estimation. A comparative analysis with other conventional LP methods shows that our approach may suggest an efficient alternative to traditional statistical methods and LP formulations for the multiple-group classification problem.

This paper is organized as follows. Section 2 consists of a brief review of the Polytomous Logistic Regression Model. Section 3 presents an overview on linear programming for detecting separation. In section 4 we propose a linear programming formulation to detecting separated data. In Section 5 we apply the proposed model on data sets taken from the literature and compare their performance with those that were obtained from another methods. Section 6 gives a brief conclusion and makes suggestions for future studies.

2. Polytomous Logistic Regression Model

Let us consider a sample of n independent observations, available from the groups $G_1, ..., G_s$, and a vector of explanatory variables, $\underline{\mathbf{x}}^T = (x_0, x_1, ..., x_p)$, where $x_0 \equiv 1$, for convenience. Let Y denote the polytomous dependent variable with s possible outcomes. We will summarize the n observations in a matrix form given by:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & ... & x_{p1} \\ 1 & x_{12} & ... & x_{p2} \\ ... & ... & ... & ... \\ 1 & x_{1n} & ... & x_{pn} \end{bmatrix}$$

The Classical Logistic Regression (CLR) Model assumes that the posterior probabilities have the form:

$$P\left(G_k \mid \underline{\mathbf{x}}\right) = \frac{exp\left(\underline{\mathbf{B}}_k\right)}{\sum\limits_{i=1}^{s} exp\left(\underline{\mathbf{B}}_i\right)} \ ,$$

where $\underline{\mathbf{B}}_k = \beta_{k0} + \sum\limits_{j=1}^{p} \beta_{kj} x_j$, $k = 1\,,\, 2\,,\, ...\,,\, s-1$ and $\underline{\mathbf{B}}_s = \mathbf{0}$. In this paper the group s is called reference group. The model involves $(s-1)\,(p+1)$ unknown parameters and the conditional likelihood function is:

$$L\left(\underline{\mathbf{B}} \mid \mathbf{Y}, \underline{\mathbf{x}}\right) = \prod_{i=1}^{n} \prod_{k=1}^{s} \left[P\left(G_k \mid \underline{\mathbf{x}}_i\right)\right]^{Y_{ki}} \ ,$$

where $\mathbf{Y} = \left(\mathbf{Y}_1\,,\, ...\,,\, \mathbf{Y}_n\right)^T$ and $\mathbf{Y}_i = (Y_{1i}, ..., Y_{si})$, with $Y_{ki} = 1$ if $Y = k$, and $Y_{ki} = 0$ otherwise. Taking the logarithm, the log-likelihood function is given by:

$$\ell\left(\underline{\mathbf{B}} \mid \mathbf{Y}, \underline{\mathbf{x}}\right) = \sum_{i=1}^{n} \sum_{k=1}^{s} Y_{ki} ln\left[P\left(G_k \mid \underline{\mathbf{x}}_i\right)\right] \ .$$

Thus:

$$\frac{\partial}{\partial \beta_{kj}} \ell\left(\underline{\mathbf{B}} \mid \mathbf{Y}, \underline{\mathbf{x}}\right) = \sum_{i=1}^{n} x_{ij} \left(Y_{ki} - P\left(G_k \mid \underline{\mathbf{x}}_i\right)\right) \ .$$

The Maximum Likelihood Estimator (MLE) $\hat{\underline{\mathbf{B}}}$ is obtained by setting the derivatives to zero and solving for $\underline{\mathbf{B}}$. The solution is found using an iterative procedure, such as Newton-Raphson method.

In practice, the estimation of unknown parameters should be considering the possible configurations of the sample points. Albert and Anderson (1984) suggested a sample classification into three categories: Overlapped, completely separated and quasi-completely separated, when perfect prediction occurs only for a group of observations. They also proved that the MLE do not exists if, and only if, there is overlap of the data points. If there is complete, or quasi-complete separation, existing iterative methods fail to converge, or give a wrong answer.

3. Linear Programming for Detecting Separation

The use of linear programming for detecting separation among the sample points was proposed by Albert and Anderson (1984). We say that two groups are completely separated if there exists a vector $\underline{\mathbf{B}} \in R^m$, where $m = (s-1)(p+1)$, such that for all $i \in G_j$, and $j, t = 1, ..., s\ (j \neq t)$:

$$(\underline{\mathbf{B}}_j - \underline{\mathbf{B}}_t)^T \underline{\mathbf{x}}_i > 0\,.$$

We say there is quasi-complete separation if, for all $i \in G_j$, and $j, t = 1, ..., s\,(j \neq t)$:

$$(\underline{\mathbf{B}}_j - \underline{\mathbf{B}}_t)^T \underline{\mathbf{x}}_i \geq 0\,,$$

with equality for at least one (i, j, t) triplet. The points for which the equality holds are said to be quasi-separated.
In binary logistic regression, if there is complete separation, the MLE do not exist. But, in polytomous logistic regression, complete separation does not make the same sense, although the parameter estimation is not necessarily affected. We say that two groups G_i and G_j are linearly separable if there exists a vector $\underline{\mathbf{\Omega}} = (\omega_1, ..., \omega_p)$ and a real number δ such that $\underline{\mathbf{\Omega}}\,\underline{\mathbf{x}}_k > \delta$ if $\underline{\mathbf{x}}_k \in G_i$ and $\underline{\mathbf{\Omega}}\,\underline{\mathbf{x}}_k < \delta$ if $\underline{\mathbf{x}}_k \in G_j$, where $i, j = 1, ..., s$, $i \neq j$ and $k = 1, ..., n$. When there are more than two groups, the difference between linear separability and separation becomes more important. In this case linear separability means the existence of a set of vectors $\underline{\mathbf{\Omega}}_1, ..., \underline{\mathbf{\Omega}}_s$ satisfying $s(s-1)$ inequalities given by:

$$(\underline{\mathbf{\Omega}}_j - \underline{\mathbf{\Omega}}_t)^T \underline{\mathbf{x}}_i \geq \delta\,,$$

for all $i = 1, ..., n$, and $j, t = 1, ..., s\,(j \neq t)$.
For $s > 2$ groups, Bennett and Mangasarian (1992), defined the piecewise linear separability of data from s groups as: The data from s groups are piecewise-linear-separable if there exist a vector $({\beta_0}^k, {\beta_1}^k, ..., {\beta_p}^k) \in R^{p+1}$, $k = 1, ..., s$, such that, $\forall i \in G_h$:

$${\beta_0}^h + \sum_{j=1}^{p} x_{ij}{\beta_j}^h \geq {\beta_0}^k + \sum_{j=1}^{p} x_{ij}{\beta_j}^k + 1$$

where $k \neq h$.
Let $\mathbf{X}_j$ be the matrix with rows $\mathbf{x}_i^T$ such that $i \in G_j$. We can define the $(s-1) \times s$ matrix $\tilde{\mathbf{X}}_j$ to have blocks $\mathbf{X}_j$ in each element of column j, blocks $-\mathbf{X}_j$ in row k and column k, for $k < j$, and in row $k-1$ and column k, for $j < k$, and to be zero otherwise. For example, if the problem has four groups, the matrix $\tilde{\mathbf{X}}_3$, is given by:

$$\tilde{\mathbf{X}}_3 = \begin{bmatrix} -\mathbf{X}_3 & 0 & \mathbf{X}_3 & 0 \\ 0 & -\mathbf{X}_3 & \mathbf{X}_3 & 0 \\ 0 & 0 & \mathbf{X}_3 & -\mathbf{X}_3 \end{bmatrix}.$$

As stated in Konis (2007), if we let

$$\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{X}}_1 \\ \vdots \\ \tilde{\mathbf{X}}_s \end{bmatrix},$$

then $\tilde{\mathbf{X}}\underline{\mathbf{B}} \geq 0$ implies that $\underline{\mathbf{B}}$ satisfies the conditions for quasi-complete separation. If $\tilde{\mathbf{X}}_{\mathbf{j}}\underline{\mathbf{B}} \geq 0$, for a given j, then

$$(\underline{\mathbf{B}}_j - \underline{\mathbf{B}}_t)^T \underline{\mathbf{x}}_i \geq 0\,,$$

4. Linear Programming Model for Detecting Separated Data

Let us consider s groups $G_1, ..., G_s$, and a vector of explanatory variables, $\underline{\mathbf{x}}^T = (x_1, ..., x_p)$. Suppose there is complete separation among two groups, G_i and G_j, $i = 1, ..., s-1, j = 2, ..., s \, (i < j)$. Then there is a hyperplane H_{ij} such that all of the sample points in G_i lie on one side of H_{ij} and all of the sample points in G_j lie on the other side of the hyperplane. The distance of the $\underline{\mathbf{x}}_k$ point from the hyperplane is given by $d_{kij} = \underline{\mathbf{x}}_k^T \underline{\mathbf{U}}_{ij}$, where $\underline{\mathbf{U}}_{ij}$ is a unit vector normal to H_{ij}. If there is complete separation among G_i and G_j, then there is a vector $\underline{\mathbf{U}}_{ij}$ such that $d_{kij} < 0$, if $\underline{\mathbf{x}}_k \in G_i$, and $d_{kij} > 0$, if $\underline{\mathbf{x}}_k \in G_j$. If there is quasi-complete separation among the groups G_i and G_j, then $d_{kij} \leq 0$, if $\underline{\mathbf{x}}_k \in G_i$, and $d_{kij} \geq 0$, if $\underline{\mathbf{x}}_k \in G_j$, with equality for at least one value $k = 1, ..., n$. Let $S(\underline{\mathbf{U}}_{ij}) = \sum_{k=1}^{n} d_{kij}$ and a hyperplane P_{ij} with normal vector $\underline{\mathbf{U}}_{ij}^*$ such that $S(\underline{\mathbf{U}}_{ij}^*)$ is maximum. In this case, finding $\underline{\mathbf{U}}_{ij}^*$ can be posed as:

$$\begin{aligned} \text{Max } S(\underline{\mathbf{U}}_{ij}) &= \textstyle\sum_{k=1}^{n} d_{kij} \\ s.t \quad d_{kij} &= \underline{\mathbf{x}}_k^T \underline{\mathbf{U}}_{ij} \geq 0 \\ d_{kij} &= \underline{\mathbf{x}}_k^T \underline{\mathbf{U}}_{ij} \leq 0 \end{aligned}$$

If there is no vector $\underline{\mathbf{U}}_{ij}$ satisfying the constraints, then there is overlap among G_i and G_j. If the model above is feasible, its solution provides a vector $\underline{\mathbf{U}}_{ij}^*$ such that $S(\underline{\mathbf{U}}_{ij}^*)$ is maximum. An additional constraint, given by $\underline{\mathbf{U}}_{ij}^T \underline{\mathbf{U}}_{ij} = 1$ is not considered because our purpose is to determine if the sample point is completely separated, or quasi-completely separated, by $\underline{\mathbf{U}}_{ij}$. Hence the length of $\underline{\mathbf{U}}_{ij}$ is not relevant. Furthermore, the referred constraint gives a nonlinearly constrained optimization problem, which is not appropriated for our purpose. The proposed model can be solved using techniques of linear programming, such as Simplex Method or Interior Points Method.

5. Applications

In this section we consider two benchmark data sets, taken from the literature. Iris Data, taken from Fisher (1936), and Fatty Acid Composition Data, taken from Brodnjak − Vončina et al. (2005). We have applied the proposed model to both data sets. The results achieved are given in the sequence.

Example 1: Iris Data. There are three groups of Iris flowers: *Iris Setosa* (G_1), *Iris Versicolor* (G_2) and *Iris Virginica* (G_3). For each group there are 50 observations and four independent variables: Sepal Length (x_1), Sepal Width (x_2), Petal Length (x_3) and Petal Width (x_4). In this paper, the reference group is Iris Virginica.

Our results showed that two groups, G_2 (*Iris Versicolor*) and G_3 (*Iris Virginica*), overlap and form a cluster completely separated from G_1 (*Iris Setosa*).

Example 2: Fatty Acid Data. There are 120 observations, five groups and seven variables, representing the percentage levels of seven fatty acids, namely palmitic (x_1), stearic (x_2), oleic (x_3), linoleic (x_4), linolenic (x_5), eicosanoic (x_6) and eicosenoic (x_7) acids. In this paper we consider five groups: rapeseed (G_1), sunflower (G_2), peanut (G_3), corn (G_4) and pumpkin (G_5) oils. In this paper the reference group is G_5 (pumpkin oil). The original data set have eight groups, and the complete table of the original data can be found in Brodnjak − Vončina et al. (2005).

Our results showed that the group G_3 (pumpkin oil) is completely separated from G_1 (rapeseed oil). Furthermore, the group G_4 (corn oil) is completely separated from G_5) (pumpkin oil).

6. Conclusions

The purpose with this job is simply to develop and implement a simple and direct model based on Linear Programming which allows the detection of separation among sample points, in order to estimate the parameters for the Polytomous Logistic Regression Model, and to explore the performance of the model. This paper is not intended to give a detailed explanation about theoretical aspects which involves neither linear programming techniques nor the existence of the referred parameters. The results achieved suggest that the proposed approach is a promising alternative to detecting separation, even when a large number of dimensions have to be considered. The proposed approach does not need any particular ordering arrangement of data, which is an advantage for practical purposes. Furthermore, there are not computational difficulties to implement the referred approach.

References

Albert, A. and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, **71**, pp. 1-10.

Anderson, J. A. (1975). Quadratic logistic discrimination. Biometrika, **62**, pp. 149-154.

Andruski-Guimarães, I. and Chaves-Neto, A. (2009). Estimation in polytomous logistic model: comparison of methods. Journal of Industrial and Management Optimization, **5**, pp. 239-252.

Bennett, K. P. and Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. Optimizations Methods and Software, **1**, pp. 23-34.

Brodnjak − Vončina, D., Kodba, Z.C. and Novič, C. (2005). Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. Chemometrics and Intelligent Laboratory Systems **75**, pp. 31-43.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, **3**, pp. 179-188.

Gehrlein, W. V. (1986). General mathematical programming formulations for the statistical classification problem. Operations Research Letters, **5**, 6, pp. 299-304.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. Statistics in Medicine, **21**, pp. 2409-2419.

Konis, K. (2007) Linear programming algorithms for detecting separated data in binary logistic regression models. DPhill in Computational Statistics, Worcester College, University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, United Kingdom.

Rousseeuw, P. J. and Christmann, A. (2003). Robustness against separation and outliers in logistic regression. Computational Statistics & Data Analysis **43**, pp. 315-332.

Santner, T. J. and Duffy, D. E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum lihelihood estimates in logistic regression models. Biometrika **73**, 3, pp. 755-758.

Silvapulle, M. J. and Burridge, J. (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. Journal of Royal Statistical Society B **48**, 1, pp. 100-106.