The problem of separation in logistic regression

Gabriela Cintra Raquel
Mário de Castro Andrade Filho*

Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação,
Departamento de Matemática Aplicada e Estatística,
Avenida Trabalhador São-carlense, 400, 13566-590, São Carlos, SP, Brasil

April 21, 2015

*Orientador. Tel.: (16) 3373 6631; fax: (16) 3373 9650. E-mail: mcastro@icmc.usp.br.
1 Introduction

In many application areas, when dealing with regression models for a
dichotomous response variable, logistic regression is one of the most adopted
modelling approaches. The literature on the subject is huge and growing
fast; see, for example, Hosmer et al. (2013) for a recent account.

The so-called problem of separation or monotonic likelihood, studied by
Albert & Anderson (1984), occurs in the fitting process of a logistic model
when the likelihood function converges while at least one regression coefficient
estimate diverges to $-\infty$ or $+\infty$, so that the maximum likelihood estimator
does not exist. Such situation takes place if the subsets of the data with
response variable equal to 1 and 0 can be perfectly separated by a single
covariate or by an intricate linear combination of covariates. For example,
if a covariate represents a risk factor with two levels (absence and presence,
say) and for all subjects with the risk factor absent, the response is equal to
1, the problem happens.

2 Goals

This project will be developed along the lines sketched below:

a) To study the separation problem and its consequences to the inferential
   results.

b) To study the sensitivity to the choice of the penalty function (see Sec-
   tion 3). The penalty function in (2) comprises the whole sample of size
   $n$. It would be of interest to assess some properties of a penalty function
   that uses $n_0 < n$ observations.
c) To carry out a simulation study in R language (R Core Team, 2015) to assess some properties of the penalized solution.

d) To study the problem of separation in the context of probit regression (Collett, 2003).

e) To apply the methods to the analysis of real data sets suffering from the problem of separation.

3 Methods

The inferential methods in this project are built on a modification of the likelihood function. Let $\beta$ denote the $p \times 1$ vector of regression coefficients. The maximum likelihood estimate of $\beta$ is a solution of the score equations $U(\beta) = 0$, where $U(\beta) = \partial \log L(\beta) / \partial \beta$ and $L(\beta)$ denotes the likelihood function of $\beta$. Based on a procedure by Firth (1993), Heinze & Schemper (2002) proposed a solution to the separation problem by means of penalized maximum likelihood estimation. The estimate $\hat{\beta}^*$ is obtained by solving the equations $U^*(\beta) = 0$, where

\[
U^*(\beta) = \frac{\partial}{\partial \beta} \log L(\beta) + \frac{1}{2} \text{trace} \left\{ I(\beta)^{-1} \frac{\partial I(\beta)}{\partial \beta} \right\}, \tag{1}
\]

with $I(\beta)$ denoting the Fisher information matrix. This modification guarantees finite coefficient estimates. The modified score function in (1) arises from the penalized likelihood function

\[
L^*(\beta) = L(\beta) |I(\beta)|^{1/2}, \tag{2}
\]

where $| \cdot |$ denotes the determinant. The influence of the penalty function $|I(\beta)|^{1/2}$ is asymptotically negligible.
Hypotheses on the regression coefficients are tested using the likelihood ratio test statistic derived from $\log L^*(\beta)$. This statistic is inverted in order to compute confidence intervals for the coefficients.

A computational framework for making inference using the penalized likelihood approach is available in an R package developed in Heinze et al. (2013).

References


