ELSEVIER Computational Statistics & Data Analysis 43 (2003) 315–332

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Robustness against separation and outliers in logistic regression

Peter J. Rousseeuw[a], Andreas Christmann[b,*]

[a] *Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (UIA), Universiteitsplein 1, B-2610 Wilrijk, Belgium*

[b] *University of Dortmund, HRZ, Abteilung A1 University of Dortmund, D-44221 Dortmund, Germany*

Received 30 September 2002; received in revised form 30 September 2002

Abstract

The logistic regression model is commonly used to describe the effect of one or several explanatory variables on a binary response variable. It suffers from the problem that its parameters are not identifiable when there is separation in the space of the explanatory variables. In that case, existing fitting techniques fail to converge or give the wrong answer. To remedy this, a slightly more general model is proposed under which the observed response is strongly related but not equal to the unobservable true response. This model will be called the hidden logistic regression model because the unobservable true responses are comparable to a hidden layer in a feedforward neural net. The maximum estimated likelihood estimator is proposed in this model. It is robust against separation, always exists, and is easy to compute. Outlier-robust estimation is also studied in this setting, yielding the weighted maximum estimated likelihood estimator.

Keywords: Binary regression; Hidden layer; Neural net; Overlap; Robustness

1. Introduction

The logistic regression model assumes independent Bernoulli distributed response variables with success probabilities $\Lambda(x_i'\theta)$ where Λ is the logistic distribution function, $x_i \in \mathbb{R}^p$ are vectors of explanatory variables, $1 \leqslant i \leqslant n$, and $\theta \in \mathbb{R}^p$ is unknown. Under these assumptions, the classical maximum likelihood (ML) estimator has certain asymptotic optimality properties. However, even if the logistic regression assumptions

* Corresponding author. Tel.: +49-231-755-2763; fax: +49-231-755-2731.
E-mail address: a.christmann@hrz.uni-dortmund.de (A. Christmann).

0167-9473/03/$ - see front matter
PII: S0167-9473(02)00304-3

are satisfied there are data sets for which the ML estimate does not exist. This occurs for exactly those data sets in which there is no overlap between successes and failures, cf. Albert and Anderson (1984) and Santner and Duffy (1986). This identification problem is not limited to the ML estimator but is shared by all estimators for logistic regression, such as that of Künsch et al. (1989).

One way to approach this problem is to measure the amount of overlap. This can be done by exploiting a connection between the notion of overlap and the notion of regression depth proposed by Rousseeuw and Hubert (1999a), leading to the algorithm of Christmann and Rousseeuw (2001). A comparison between this approach and the support vector machine is given in Christmann et al. (2002).

Of course, finding that there is no overlap in the data set does not imply that the underlying population distributions have no overlap, and the practitioner often needs to obtain regression estimates and odds ratios anyway, e.g. in a comparative study.

In Section 2 we adopt a different approach, based on a slight extension of the logistic regression model. This model assumes that due to an additional stochastic mechanism the true response of a logistic regression model is unobservable, but that there exists an observable variable which is strongly related to the true response. E.g., in a medical context there is often no perfect laboratory test procedure to detect whether a specific illness is present or not (i.e., misclassification errors may sometimes occur). In that case, the true response (whether the disease is present) is not observable, but the result of the laboratory test is.

It can be argued that the true unobservable responses are comparable to a hidden layer in a feedforward neural network model, which is why we call this the hidden logistic regression (HLR) model. In Section 3 we propose the maximum estimated likelihood (MEL) technique in this model, and show that it is immune to the identification problem described above. In Section 4 we consider outlier–robust estimation in this setting. The MEL estimator and its robustification are studied by simulations (Section 5) and on real data sets (Section 6). Section 7 provides a discussion and an outlook to further research.

2. The hidden logistic regression model

The classical logistic regression model assumes n observable independent responses Y_i with Bernoulli distributions $\mathrm{Bi}(1, \Lambda(x_i'\theta))$, where $i=1, \ldots, n$ and $\theta \in \mathbb{R}^p$. Throughout this paper we assume that there is an intercept, so we put $x_{i,1} = 1$ for all i, and thus $p \geqslant 2$.

The new model assumes that the true responses are unobservable (latent) due to an additional stochastic mechanism. In medical diagnosis there is typically no test procedure (e.g. a blood test) which is completely free of misclassification errors. Another possible cause of misclassifications is the occurrence of clerical errors, which could be made when registering the response variable or (perhaps more often) one of the explanatory variables.

To clarify the model, let us first consider a medical application with only $n=1$ patient. His/her true status (e.g. presence or absence of the disease) has two possible values,