Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 3199–3221
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

# Cressie and Read power-divergences as influence measures for logistic regression models

J. Muñoz-García[a], J.M. Muñoz-Pichardo[a,*], L. Pardo[b]

[a] *Departamento de Estadística e I.O. Universidad de Sevilla, Spain*
[b] *Departamento de Estadística e I.O. Universidad Complutense de Madrid, Spain*

Received 24 June 2004; received in revised form 10 June 2005; accepted 10 June 2005
Available online 8 August 2005

**Abstract**

A sample version of the power-divergence measures of Cressie and Read is proposed for the influence analysis in the logistic regression model. Influence measures are obtained by quantifying the deviation between the sample distribution of an estimate obtained with all the observations and the sample distribution of the same estimate obtained without any observation. In particular, this approach is applied to three estimates of the model: the MLE of regression coefficients vector, the probabilities vector and the linear predictor of a future case. Some examples are considered to clarify the usefulness of the introduced diagnostics.


*Keywords:* Power-divergence measure; Logistic regression; Influence analysis; Residuals; Leverage

## 1. Introduction

Logistic regression is a very useful tool in the study of a binary data set obtained under experimental conditions as well as in observational studies. In both situations, especially the second one, the data set can contain observations that are not well explained by the model and/or observations that exert an undue influence on some aspect of the model. In this sense, it is interesting to remember the following sentence pointed out in the paper written by Pregibon (1982, p. 210): "after fitting a logistic regression model and prior to

* Corresponding author. Tel.: +34 954 55 769 39; fax: +34 954 62 28 00.
*E-mail address:* juanm@us.es (J.M. Muñoz-Pichardo).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.06.018

drawing inferences from it, the natural succeeding step is that of critically assessing the fit". In particular, the techniques of identifying of outlying and influential observations are included in this critical study.

Some of the methods used for assessing influence and finding outliers in logistic regression are similar to those used for linear regression, see for instance Cook and Weisberg (1982). Pregibon (1982) carried out an interesting study about this topic and presented diagnostics to identify observations which are influential relative to the estimation of the regression coefficient vector. Johnson (1985) gave diagnostics for detecting influential observations relative to the determination of probabilities and the classification of future observations. Later, these measures were applied to generalized linear models in some interesting papers: Williams (1987), Lee (1988), Thomas and Cook (1989, 1990) and Lee and Zhao (1997).

The standard methods of determining influence are based on considering a suitable scheme for perturbing the model as well as a procedure to compare the considered model and the perturbed model. The two most important schemes of perturbing are: considering the full binary data set except case $i$, case-deletion approach, and small perturbations of any elements of data, local influence approach (Cook, 1986). In general, for determining influence on an estimate of a parameter, all the procedures are focused on measuring the distance between this estimate and the estimate under the perturbed model. Recently, Jiménez-Gamero et al. (2002) proposed to consider the Rao's distance between the distribution of the estimated parameter of interest and the distribution of the estimate under the perturbed model. This approach can be described as follows: Let $D$ be a data set which is assumed to follow a model $M$, let $R = R(D, M)$ be a statistic based on the data, and let $F_R$ be its distribution function. By perturbing the model formulation, the model $M^*$ is obtained, with $R^*$ and $F_{R^*}$ being the statistic and its distribution under the perturbed model, respectively. The changes due to the perturbation on the statistics can be evaluated by the distance between $F_R$ and $F_{R^*}$.

Through this approach, it is possible to obtain influence diagnostics which are more complete than the diagnostics obtained by direct comparison of the estimate values. The term "more complete" is used because they not only compare the obtained values for a specific sample but the probability distributions of the estimators. The metrics used previously are Rao's distance (Jiménez-Gamero et al., 2002) and Frèchet's distance (Muñoz-Pichardo et al., 2004). Also Johnson (1985) used the Kullback–Leibler divergence (Kullback, 1968, Chapter 1) to quantify the influence on the estimate probabilities in the logistic regression model.

In this work the power-divergence family of divergence measurements introduced by Cressie and Read (1984) is considered to measure the deletion-case effect under the sample distribution of regression coefficients in logistic regression and on the estimated probability vector characterizing the model. This second procedure will be a generalization of the procedure given in Johnson (1985) based on the Kullback–Leibler divergence measure.

In Section 2, the most important results in relation to the Cressie–Read power divergence measures (Cressie and Read, 1984) are presented. These results will be necessary to use in the rest of the paper. Section 3 is devoted to giving an overview of the logistic regression model and to introduce some results and notation that will be necessary in the rest of the paper. Some new measures to quantify the influence on three aspects of the logistic