Chapter 16

Outlier Detection in Logistic Regression

A. A. M. Nurunnabi
SLG, University of Rajshahi, Bangladesh

A. B. M. S. Ali
CQUniversity, Australia

A. H. M. Rahmatullah Imon
Ball State University, USA

Mohammed Nasser
University of Rajshahi, Bangladesh

ABSTRACT

The use of logistic regression, its modelling and decision making from the estimated model and subsequent analysis has been drawn a great deal of attention since its inception. The current use of logistic regression methods includes epidemiology, biomedical research, criminology, ecology, engineering, pattern recognition, machine learning, wildlife biology, linguistics, business and finance, et cetera. Logistic regression diagnostics have attracted both theoreticians and practitioners in recent years. Detection and handling of outliers is considered as an important task in the data modelling domain, because the presence of outliers often misleads the modelling performances. Traditionally logistic regression models were used to fit data obtained under experimental conditions. But in recent years, it is an important issue to measure the outliers scale before putting the data as a logistic model input. It requires a higher mathematical level than most of the other material that steps backward to its study and application in spite of its inevitability. This chapter presents several diagnostic aspects and methods in logistic regression. Like linear regression, estimates of the logistic regression are sensitive to the unusual observations: outliers, high leverage, and influential observations. Numerical examples and analysis are presented to demonstrate the most recent outlier diagnostic methods using data sets from medical domain.

DOI: 10.4018/978-1-4666-1830-5.ch016
INTRODUCTION

The logistic regression model has found wide range of usage in various fields. Over the last few years, use of logistic regression has been exploded. Naturally, its use spurs the need to have diagnostic tools to justify the appropriateness of the model. Hosmer and Lemeshow (2000) pointed, “In recent years diagnostics has become an essential part of logistic regression”. When analysing a data set some observations are often occurred that are somehow different from the majority, simply such observations are treated as the outliers. Sometimes the outlying observations are not incorrect rather they are made under exceptional circumstances, or they belong to other population(s). We often observe that outliers greatly affect the covariate pattern and consequently their presence can mislead our interpretation. So we need to identify such observations and study their impact on the model (Imon and Hadi, 2008). Although a rich pool of literature (Pregibon, 1981; Hosmer and Lemeshow, 2000; Ryan, 1997) exists for studying outlier diagnostics, the use of the logistic regression modelling especially for the detection of multiple outliers is still a major concern to the researchers and in need to the practitioners. It is now evident that most of the popular diagnostic methods based on single-case deletion approach can mislead the analysis in the presence of multiple outliers and /or influential cases because of the well-known masking and/or swamping problem (see Atkinson, 1986). As remedy to masking and swamping effects, the group-deletion approach is used in regression diagnostics. Group-deletion technique helps us to reduce the maximum disturbance by deleting the suspect group of influential cases at a time (see Hadi and Simonoff, 1993). It helps to make the data more homogeneous after the group-deletion (Nurunnabi et al., 2011). Since most of the existing diagnostics approaches are originated from the ideas in linear regression, this chapter states the basic ideas in linear regression as well as logistic regression where it is relevant, and for the benefit of the reader.

In this chapter, we introduce logistic regression, the nature and general ideas of outliers and how we can find out outliers, without getting into the deep of the subject from the mathematical and/or statistical point of view. We concentrate here to give the general understanding and to show the importance of outlier detection prior decision making and model building for the researchers in different fields. The rest of the chapter is structured as follows. Section 2 makes a basic discussion about logistic regression and different types of outliers. In Section 3, we introduce several recent and effective outlier detection methods followed by the numerical examples and analysis in section 4. Section 5 concludes the chapter with findings and some future issues of outlier detection in Logistic regression.

LOGISTIC REGRESSION
MODEL FORMULATION

Regression analysis deals how the values of the response (dependent variable) change with the change of one or more explanatory (independent) variables. It is appealing because it provides a conceptually simple method for investigating functional relationship among variables (Chatterjee and Hadi, 2006). In any regression problem the key quantity is the mean value of the outcome (dependent or response) variable, given the value of the explanatory (independent) variable(s), $E(Y|X)$. In linear regression, we assume that this mean is expressed as an equation linear in $X$ (or some transformations of $X$ or $Y$) such as

$$E(Y / X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p. \quad (1)$$

Hence $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon, \quad (2)$

$$= X \beta + \varepsilon = E(Y / X) + \varepsilon, \quad (3)$$