Computational Statistics and Data Analysis 78 (2014) 33–42

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

M-regression, false discovery rates and outlier detection with application to genetic association studies☆

V.M. Lourenço [a,*], A.M. Pires [b]

[a] Department of Mathematics and CMA, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal

[b] Department of Mathematics and CEMAT, Instituto Superior Técnico, Universidade Técnica de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal

ARTICLE INFO

Article history:
Received 17 April 2013
Received in revised form 28 March 2014
Accepted 28 March 2014
Available online 12 April 2014

Keywords:
Robust regression
Robust outlier test
False discovery rate
Genetic association studies
Single nucleotide polymorphism

ABSTRACT

Robust multiple linear regression methods are valuable tools when underlying classical assumptions are not completely fulfilled. In this setting, robust methods ensure that the analysis is not significantly disturbed by any outlying observation. However, knowledge of these observations may be important to assess the underlying mechanisms of the data. Therefore, a robust outlier test is discussed, together with an adequate false discovery rate correction measure, to be used in the context of multiple linear regression with categorical explanatory variables. The methodology focuses on genetic association studies of quantitative traits, though it has much broader applications. The method is also compared to a benchmark rule from the literature and its good performance is validated by a simulation study and a real data example from a candidate gene study.

1. Introduction

Outlier detection is a widely discussed issue in the literature in many research fields, including genetic association studies, where there is interest in cleaning the data before analysis (Paschou et al., 2008; Arshadi et al., 2009; Inouye et al., 2010). Moreover, the decision as to whether or not an observation should be flagged as an outlier is frequently based on a rule of thumb.

In the last fifty years or so there has been a major effort to develop statistical procedures resistant to small deviations from the assumptions of the model, which are robust to outliers and stable with respect to deviations from normality (Ronchetti, 2006; Dell'Aquila and Ronchetti, 2006). This interest is mainly stimulated by the fact that the identification and removal of outlying observations from the data, so that a classical analysis may be performed, does not come without concern: (i) true outliers are not always visible, due to *masking* and *swamping* effects, among other reasons; (ii) mild outliers cannot always be differentiated from regular data; (iii) removing outliers from the data reduces sample size, may affect the distribution theory and variances may be underestimated from the cleaned data.

Since outliers deriving from intrinsic characteristics of the individual being measured may provide relevant information for the analysis, their identification is essential to better understand the underlying mechanisms of the system described by

☆ Includes two pdf files with supplementary Tables and Figures, a pdf file with duplicated Figures from the paper (see Appendix A), another pdf file with a brief description of False Discovery Rates and one text file with the proposed outlier test and outlier diagnostic plot implemented in the R language.

* Correspondence to: Department of Mathematics, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal. Tel.: +351 212948388; fax: +351 212948391.

E-mail address: vmml@fct.unl.pt (V.M. Lourenço).

http://dx.doi.org/10.1016/j.csda.2014.03.019
0167-9473/

the data. For example, in the context of environmental sciences, an outlier could very well represent an unusual event that took place (Daszykowski et al., 2007) and whose comprehension could be helpful for the study. This also occurs in many other areas of research, e.g., genetic association studies (Tzeng et al., 2003).

Robust estimation and outlier detection are intimately related. McKean et al. (1993) showed that residual plots from M-regression, performed via a convex objective function, can be interpreted as least squares (LS) residual plots, as will be the case when we consider Huber's proposal (Huber, 1964). Given such robust estimates of the true values of an assumed null model, analyzing the residuals obtained from the robust fit plotted against the fitted values, can indicate which, if any, of the observations are regression outliers. This is not usually the case when classical residuals are analyzed, since outliers often tend to pull the regression line towards them, thus making their residuals look normal. As a result, outliers are difficult to distinguish from regular observations and sometimes regular observations may be incorrectly identified as outliers.

Rice and Spiegelhalter (2006) proposed a robust outlier test to be used in the context of a simple linear regression model that was shown to perform well not only in terms of power but also in terms of good agreement between sample and nominal false discovery rate (FDR) levels. However, they only considered the intercept in the robust fit in their simulations which does not tell us whether there is FDR control in the multiple regression setting (e.g., if we are simultaneously testing for association between a quantitative trait and several to many single-nucleotide polymorphisms—SNPs). Here, the choice of both an FDR correction procedure (Cerioli and Farcomeni, 2011; Cerioli et al., 2013) and a robust estimate of scale for the robust outlier test may be crucial to maintain FDR control and power and we have not seen this problem tackled in the literature. Therefore, in this work, the "extension" of the test to the multiple regression setting is studied, in particular within the framework of genetic association studies of quantitative traits, although it may easily be generalized to any studies where a continuous response variable and categorical explanatory variables are considered. We underline that M-regression was shown by Lourenço et al. (2011) to be suitable for these studies.

Hence, following a brief literature review (Section 2), we evaluate the performance of Rice and Spiegelhalter's robust outlier test in the context of a robust multiple regression fit via simulation, for two robust estimates of scale, several multiple testing corrections and various contamination scenarios, where a benchmark rule from the literature is also considered (Section 3). In Section 4 we analyze data from a candidate gene study where the adequacy of the "upgraded" robust outlier test is validated. Finally, in Section 5 we summarize the results.

2. Methodology

The general multiple linear regression model can be written as

$$Y = X\beta + \varepsilon, \tag{1}$$

under the usual assumptions of independence, homoscedasticity and normality of the errors, where $Y = (Y_1, \ldots, Y_m)^T$ is the $(m \times 1)$ vector of the response variable, X is the $(m \times p)$ design matrix coding for $p-1$ explanatory variables, the first column of X having 1's, $\beta = (\beta_0, \ldots, \beta_{p-1})^T$ are the unknown parameters and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)^T$ is a vector of non-observable errors with $E(\varepsilon) = \mathbf{0}$. Note that we consider $m > p$.

In M-regression, the estimates are the solutions to the following minimization problem,

$$\widehat{\beta}_R = \arg\min_{\beta} \sum_{i=1}^{m} \rho\Big(\frac{\varepsilon_i}{\widehat{\sigma}}\Big) = \arg\min_{\beta} \sum_{i=1}^{m} \rho\Big(\frac{r_i(\beta)}{\widehat{\sigma}}\Big), \tag{2}$$

where $\widehat{\sigma}$ is a robust scale estimate of σ and ρ is a function with specific properties.

There are several proposals in the literature for the robust scale (Rousseeuw and Croux, 1993; Daszykowski et al., 2007). For example the re-scaled MAD (Hampel, 1974)

$$s = 1.4826 \times \underset{1\le i\le m}{\text{median}}\, |r_i - \underset{1\le j\le m}{\text{median}}(r_j)|, \tag{3}$$

where the value of 1.4826 is to assure consistency with the Gaussian distribution. Although this estimator has a high breakdown point (50%), it lacks efficiency for the normal model (37%). Thus, Croux and Rousseeuw (1992) proposed the estimator

$$Q_n = d \times \{|r_i - r_j| : i < j\}_{(k)}, \tag{4}$$

where d is a constant and $k = \binom{h}{2}$ with $h = \lfloor m/2 \rfloor + 1$, which also has a high breakdown point (50%) but with 83% efficiency for the normal when we take $d = 2.2219$; here k refers to the kth order statistic of the $\binom{h}{2}$ interpoint distances.

As to the objective function ρ, the convex function proposed by Huber (1964)

$$\rho(x) = \begin{cases} x^2/2, & |x| \le b \\ b(|x| - b/2), & |x| > b, \end{cases} \tag{5}$$

where $b > 0$ is a tuning parameter, is known to lead to efficient estimators under several conditions. Choosing $b = 1.345$, it delivers an M-estimator with 95% efficiency relative to the normal model and with substantial resistance to outlying