Computational Statistics and Data Analysis 64 (2013) 153–164

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Linear regression models with slash-elliptical errors

Izabel Cristina Alcantara [a,b], Francisco José A. Cysneiros [b,*]

[a] Departamento de Estatística - UFPB, Cidade Universitária, Campus 1, 58051-900, João Pessoa - PB, Brazil

[b] Departamento de Estatística - UFPE, Cidade Universitária, 50740-540, Recife - PE, Brazil

ARTICLE INFO

Article history:
Received 12 May 2012
Received in revised form 18 November 2012
Accepted 25 February 2013
Available online 15 March 2013

Keywords:
Heavy-tailed distribution
Leverage
Local influence
Residuals
Robust models
Slash-elliptical distribution

ABSTRACT

We propose a linear regression model with slash-elliptical errors. The slash-elliptical distribution with parameter q is defined as the ratio of two independent random variables Z and $U^{\frac{1}{q}}$, where Z has elliptical distribution and U has uniform distribution in (0, 1). The main feature of the slash-elliptical distribution is to have greater flexibility in the degree of kurtosis when compared to the elliptical distributions. Other advantages of this distribution are the properties of symmetry, heavy tails and the inclusion of the elliptical family as a limit case when $q \to \infty$. We develop the methodology of estimation, hypothesis testing, generalized leverage and residuals for the proposed model. In the analysis of local influence, we also develop the diagnostic measures based on the likelihood displacement under the some perturbation schemes. Finally, we present a real example where slash-Student-t model is more stable than other considered models.

1. Introduction

Last decades, similar works have been developed based on symmetrical distribution. A review of some areas where symmetric distributions are applied is described in Chmielewski (1981). In many situations of statistical modeling there is a need of searching for less sensitive models of outlying observations. Galea et al. (2003) developed diagnostic methods for linear symmetrical models and Cysneiros and Paula (2005) developed restricted methods in symmetrical linear models. Paula et al. (2009) introduced the class of linear models with first-order autoregressive elliptical errors and diagnostic methods were derived.

Rogers and Tukey (1972) presented the slash distribution as the probability distribution of a standard normal variable divided by an independent standard uniform variable. In the general case, we say that a random variable S has standard slash distribution with parameter $q > 0$ if it can be expressed as the ratio of two independent random variables Z and $U^{\frac{1}{q}}$, where Z has standard normal distribution $N(0, 1)$ and U has uniform distribution in (0, 1). The slash distribution has the properties of symmetry, heavy tails and converges to the normal distribution when $q \to \infty$.

Rogers and Tukey (1972) and Mosteller and Tukey (1977) discussed slash distribution and its properties. Kafadar (1982) proposed maximum likelihood estimators for location-scale parameters considering the slash distribution obtained through linear transformation $Y = \mu + \sqrt{\phi}S$. In the work of Andrews et al. (1972), Gross (1973) and Morgenthaler and Tukey (1991), the slash distribution is mainly used in simulation studies whose scenario involves extreme situations. Wang and Genton (2006) defining a skewed version of the slash distribution, assumed that the random variable Z has a multivariate skew normal distribution. Arslan (2008) introduced a new class of multivariate skew-slash distributions using the normal variance-mean mixture approach. Later, Arslan and Genç (2009) generalized the family of distributions proposed by Wang

* Corresponding author. Tel.: +55 81 21267422; fax: +55 81 21268422.
E-mail addresses: izabel@de.ufpb.br (I.C. Alcantara), cysneiros@de.ufpe.br, francisco.cysneiros@gmail.com (F.J.A. Cysneiros).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.02.029

and Genton (2006), constructed a family of multivariate distributions as a scale mixtures of the multivariate symmetric Kotz-type distribution and the uniform distribution. Lachos et al. (2008) derived diagnostic methods based on the local influence on scale-mixture models. Ferreira et al. (2011) developed EM algorithm for estimation of the parameters on scale-mixture models. Gómez et al. (2007) considered the standard slash distribution in the form:

$$Z = U^{1/q} S \sim N(0, 1)$$

and generalized slash distribution by replacing the distribution of Z by the family of univariate and multivariate elliptical distributions. This new family of distributions proposed by Gómez et al. (2007), known as slash-elliptical distribution, has the property of symmetry and greater flexibility in the degree of kurtosis when compared to the elliptical distribution. These properties were also observed by Genç (2007) to slash-power-exponential distribution. Another advantage of this family of distributions is to contain the elliptical family as a limiting case.

Given this new family of distributions, the linear regression model with error distribution in a univariate slash-elliptical family is developed, which will be called only by slash-elliptical distribution. The aim of this paper is to derive a methodology for estimating, hypothesis testing, generalized leverage, residual and diagnostic analysis based on the local influence approach. Section 2 introduces the slash-elliptical regression model and procedures for estimation are presented. Simulation studies of the proposed residual are presented. In addition, diagnostic measures based on the local influence approach are developed in Section 3. Section 4 is devoted to analysis of a real data set using slash-elliptical regression model and finally, some conclusions are presented in the final section.

2. Slash-elliptical regression model

Definition 1. A random variable Y has slash-elliptical distribution with location parameter $\mu \in \Re$ and scale $\phi > 0$ if Y can be expressed as

$$Y = \mu + \sqrt{\phi}\frac{V}{U^{1/q}},$$

where V and U are independent random variables, V has standard elliptical distribution, U has uniform distribution in $(0, 1)$ and q is the specific parameter of slash-elliptical distribution.

Definition 2. A random variable Y is called slash-elliptical variable and denoted by $\sim SEL(\mu, \phi, q, g(\cdot))$ if its density is given by:

$$f(y; \mu, \phi, q) = \frac{1}{\sqrt{\phi}} \begin{cases} \dfrac{q}{2|z|^{q+1}} H(z^2), & z \neq 0 \\ \dfrac{q}{q+1} g(0), & z = 0, \end{cases} \tag{1}$$

where $z = (y - \mu)/\sqrt{\phi}$ and $H(z^2) = \int_0^{z^2} t^{(q-1)/2} g(t)dt$, for some generating function of density $g(\cdot)$, with $g(t) > 0$ for $t > 0$ and $\int_0^{\infty} t^{-1/2} g(t)dt = 1$.

Definitions 1 and 2 can be found in Gómez et al. (2007) and Gómez and Venegas (2008). Whereas, the Definition 1 provides a method for generating slash-elliptical random variables based on elliptical and uniform distributions, the Definition 2 characterizes a slash-elliptical random variable by its density function.

Let $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ independent random variables, where $\epsilon_i \sim SEL(0, \phi, q, g(\cdot))$. We define the linear regression model with slash-elliptical errors and parameter q is fixed or known, which abbreviated to slash-elliptical model, by:

$$y_i = \mathbf{x}_i^t \beta + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

where for each ith observation y_i is the response variable, $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{ip})^t$ is the regressor vector $(p \times 1)$, and $\beta = (\beta_1, \ldots, \beta_p)^t$ is a vector of unknown parameters $(p \times 1)$.

2.1. Parameter estimation

To estimate the parameter vector $\theta = (\beta^t, \phi)^t$, we can use the maximum likelihood method, which is to maximize the log-likelihood function of the slash-elliptical model given by:

$$L(\theta) = -\frac{n}{2}\log(\phi) + \sum_{i \in A} a(z_i) + K, \tag{3}$$

with respect to θ, where $z_i = \frac{y_i - \mu_i}{\sqrt{\phi}}$, $a(z_i) = \log(H(z_i^2)) - (q + 1)\log|z_i|$ is a function that depends on θ through z_i, $A = \{i : z_i \neq 0\}$, $K = (n - n_A)\log\left(\frac{q}{q+1}g(0)\right) + n_A \log(\frac{q}{2})$ is constant with respect to θ and n_A is the number of