Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 821–834
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

The use of mixtures for dealing with non-normal regression errors

F. Bartolucci[a,*], L. Scaccia[b]

[a] *Dipartimento di Economia, Università di Urbino, Via A. Saffi 42, 61029 Urbino, Italy*
[b] *Dipartimento di Scienze Statistiche, Università di Perugia, Via A. Pascoli, 06100 Perugia, Italy*

Received 31 January 2003; received in revised form 8 April 2004; accepted 9 April 2004

Available online 5 May 2004

Abstract

In many situations, the distribution of the error terms of a linear regression model departs significantly from normality. It is shown, through a simulation study, that an effective strategy to deal with these situations is fitting a regression model based on the assumption that the error terms follow a mixture of normal distributions. The main advantage, with respect to the usual approach based on the least-squares method is a greater precision of the parameter estimates and confidence intervals. For the parameter estimation we make use of the EM algorithm, while confidence intervals are constructed through a bootstrap method.

Keywords: EM algorithm; Kurtosis; Location-scale mixtures; Normal probability plot; Residual analysis; Skewness; Switching regression

1. Introduction

A basic assumption of the linear regression model is that the error terms have a normal distribution; most of the inferential procedures currently used are based on this assumption. A wide literature now exists on the detection of violations of this assumption on the basis the ordinary least-squares (OLS) residuals. The usual technique is based on the *normal probability plot* that allows us to compare the observed residuals

* Corresponding author. Tel.: +39-075-5855202; fax: +39-0722-305550.
E-mail addresses: Francesco.Bartolucci@uniurb.it (F. Bartolucci), luisa@stat.unipg.it (L. Scaccia).
URL: http://www.econ.uniurb.it/bartolucci/index.htm

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.04.005

with those expected under normality; a more formal test, known as *correlation test for normality*, may be carried out through the Pearson correlation coefficient between these residuals (Blom, 1958; Looney and Gulledge, 1985). Other formal tests have been set up; most of them are tailored to the detection of skewness of the error terms, which naturally arises when measurements are based on extremes. Tests of this type are those of Anscombe (1961), based on the third moment of the OLS residuals, and Boos (1987); other tests have been proposed by Huang and Bolch (1974) and Quesenberry (1986). Obviously, departures from normality of the error terms may have negative effects on the inferential procedures. Cox and Hinkley (1968), for instance, showed that the variance of the OLS estimator is much larger than that of the maximum-likelihood estimator (MLE) when the error terms follow the extreme value distribution. In this situation, we also have an overestimation of the standard errors of the parameter estimates that leads to tests with significance level larger than the nominal one and to wider confidence intervals than necessary; a similar problem affects prediction intervals (see also, Boos, 1987, Section 1). The literature on remedial measures for departure from normality is not so developed as that on its detection, partially because, with large samples, the usual inferential procedures are approximately valid. A well-known method for dealing with non-normal errors is based on transformations of the response variables (Box and Cox, 1964). Also the Huber M-estimator (Huber, 1981) may be effectively used in some situations (Boos, 1987).

In this paper, we illustrate the use of mixtures as a remedial measure for non-normal errors. According to our approach, whenever we detect departure from normality, we will fit a linear regression model with the same structure as the original one, apart from the assumption that the error terms follow a mixture of normal distributions with a finite number of components. To detect departure from normality, we follow the approach of Looney and Gulledge (1985) based on the correlation coefficient between observed and expected residuals under normality (see also, Blom, 1958; Gan and Koehler, 1990; Gan et al., 1991). Note that the use of mixture models in the linear regression context is well-known, even if with other aims: to deal with two different regression functions, the so-called *switching regression* (see, for instance, Quandt and Ramsey, 1978), and to deal with outliers (Aitkin and Wilson, 1980). In this paper, instead, mixtures are exploited as a convenient semiparametric method, which lies between parametric models and kernel density estimators, to model the unknown distributional shape of the errors. In this perspective, the choice of a mixture of normal distributions seems to be a natural one, given its tractability and flexibility. An example of the use of normal mixtures to represent a wide variety of density shapes can be found in Marron and Wand (1992).

The paper is organized as follows. In Section 2 we introduce some preliminary notation and describe the correlation test for normality proposed by Blom (1958) and extensively investigated by Looney and Gulledge (1985). Then, in Section 3, we illustrate the proposed approach, based on a mixture model, to deal with non-normal errors and the estimation of the parameters of such a model through the EM algorithm. Finally, in Section 4, we illustrate the results of a simulation study that shows the advantages of the proposed approach.

2. Notation and preliminary results

The linear regression model is based on the assumption

$$Y_i = \mu_i + \varepsilon_i, \quad \mu_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, \ldots, n, \tag{1}$$

where Y_i represents the response variable for the ith sample unit, x_{ij} is the ith observation of the jth predictor variable and β_j's are regression parameters of the model. Finally, ε_i is the error term corresponding to the ith observation; under the assumption of normality, we have that ε_i, $i = 1, \ldots, n$, are i.i.d. $N(0, \sigma^2)$.

The usual approach for detecting lack of normality of the ε_i's is based on the normal probability plot. A common method to set up this plot consists in the following steps:

(1) compute the residuals of the regression, $e_1, \ldots, e_n$, where

$$e_i = y_i - \tilde{\beta}_0 - \sum_{j=1}^{p} x_{ij}\tilde{\beta}_j,$$

with y_i denoting the observed value of Y_i and $\tilde{\beta}_j$ the OLS estimate of β_j;

(2) sort the residuals $e_1, \ldots, e_n$ in ascending order and let $e_{(i)}$ be the ith smallest residual;

(3) plot the points of coordinates $(\hat{e}_{(i)}, e_{(i)})$, $i = 1, \ldots, n$, where

$$\hat{e}_{(i)} = \sqrt{s^2}\Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right)$$

is the expected value of $e_{(i)}$ under the assumption of normality,

$$s^2 = \frac{1}{n - p - 1}\sum_{i=1}^{n} e_i^2$$

is the unbiased estimate of σ^2 and Φ^{-1} denotes the inverse of the standard normal distribution.

It is clear that the closer the points are to a straight line, the more reliable the assumption of normality. A more formal assessment is based on the use of the correlation index between the observed and the expected residuals, namely

$$r = \frac{\sum_{i=1}^{n} e_{(i)}\hat{e}_{(i)}}{\sqrt{\sum_{i=1}^{n} e_{(i)}^2 \sum_{i=1}^{n} \hat{e}_{(i)}^2}}.$$

A low level of this index indicates departure from normality. Critical values for r, which depend only on n, may be found through a Monte Carlo simulation. These have been tabulated, for several values of n, by Looney and Gulledge (1985).

3. The proposed approach

When, for a given sample, the p-value of the correlation test illustrated in the previous section is less than a certain level, say 0.05, we propose to estimate a model based on the same regression function as the original one and on the assumption that the error terms follow a mixture of normal distributions. In a previous version of this paper we restricted the attention on *location mixtures*, for which all the components have the same variance. In this way, however, the paper naturally focused on the skewness of the errors as a departure from normality. Following the suggestion of a reviewer, we now consider a more general framework in which the components of the mixture are not constrained to have the same variance (*location-scale mixtures*). These normal mixtures with heteroscedastic components are expected to perform better than those with homoscedastic components when the error terms have a leptokurtic distribution. More precisely, the model considered is based on assumptions (1) and

$$\varepsilon_i \sim \sum_{h=1}^{k} \pi_h N(v_h, \tau_h^2), \tag{2}$$

where π_h's are weights adding to 1 and the v_h's satisfy the identifiability constraint $\sum_{h=1}^{k} \pi_h v_h = 0$. In the following, we illustrate how maximum-likelihood estimation of the parameters of this model may be carried out through the well-known EM algorithm (see, Dempster et al., 1977). For a gentle tutorial on the EM algorithm and its application to parameter estimation for mixture models see also Bilmes (1998).

First of all consider that the density of the ith observation of Y, y_i, is

$$\sum_{h=1}^{k} \pi_h \phi(y_i; \mu_{ih}, \tau_h^2), \qquad \mu_{ih} = v_h + \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j,$$

where $\phi(y_i; \mu_{ih}, \tau_h^2)$ denotes the density at y_i of the normal distribution $N(\mu_{ih}, \tau_h^2)$. So, the log-likelihood of the model at issue is given by

$$l = \sum_{i=1}^{n} \log \left[\sum_{h=1}^{k} \pi_h \phi(y_i; \mu_{ih}, \tau_h^2) \right].$$

To maximize l with respect to the parameters of the model we proceed through the EM algorithm, normally used in the presence of missing data. Let z_{ih} be a binary variable equal to 1 when the ith observation has been generated from the hth component and to 0 otherwise. Obviously, the variables z_{ih}'s are unknown, but, if they were known, the so-called *complete* log-likelihood would be, up to a constant factor,

$$l_c = \sum_{i=1}^{n} \sum_{h=1}^{k} z_{ih}[\log \pi_h + \log \phi(y_i; \mu_{ih}, \tau_h^2)] = l_{c1} + l_{c2}, \tag{3}$$

where

$$l_{c1} = \sum_{h=1}^{k} z_{\cdot h} \log \pi_h \quad \text{and} \quad l_{c2} = -\frac{1}{2} \sum_{h=1}^{k} z_{\cdot h} \log \tau_h^2 - \frac{1}{2} \sum_{i=1}^{n} \sum_{h=1}^{k} z_{ih} \frac{(y_i - \mu_{ih})^2}{\tau_h^2},$$

with $z_{\cdot h} = \sum_{i=1}^{n} z_{ih}$. To maximize the first component simply let π_h equal to $\hat{\pi}_h = z_{\cdot h}/n$, $h = 1, \ldots, k$. Instead, to show how l_{c2} can be maximized, it is convenient to express such a function in matrix notation; so, let $\boldsymbol{y} = (y_1 \cdots y_n)'$ be the vector of observed data and $\boldsymbol{\gamma} = (\boldsymbol{\alpha}' \ \boldsymbol{\beta}')'$ be the $k + p$-dimensional vector of identifiable parameters, where

$$\boldsymbol{\alpha} = (\beta_0 + v_1 \cdots \beta_0 + v_k)' \quad \text{and} \quad \boldsymbol{\beta} = (\beta_1 \cdots \beta_p)'.$$

Moreover, let $\boldsymbol{z}_h = (z_{1h} \cdots z_{nh})'$ and $\boldsymbol{\mu}_h = (\mu_{1h} \cdots \mu_{nh})'$ and note that the latter may be expressed as $\boldsymbol{\mu}_h = \boldsymbol{X}_h \boldsymbol{\gamma}$, where $\boldsymbol{X}_h = (\boldsymbol{O}_h \ \boldsymbol{X})$, $\boldsymbol{O}_h$ is a matrix of dimension $n \times k$ with all the elements equal to 0 apart from those of column h which are equal to 1 and $\boldsymbol{X}$ is the $n \times p$ matrix with entries x_{ij}; consequently, we have

$$l_{c2} = -\frac{1}{2} \sum_{h=1}^{k} z_{\cdot h} \log \tau_h^2 - \frac{1}{2} \sum_{h=1}^{k} \frac{1}{\tau_h^2} (\boldsymbol{y} - \boldsymbol{X}_h \boldsymbol{\gamma})' \operatorname{diag}(\boldsymbol{z}_h)(\boldsymbol{y} - \boldsymbol{X}_h \boldsymbol{\gamma}).$$

The first derivative of this function with respect to $\boldsymbol{\gamma}$ is equal to

$$\frac{\partial l_{c2}}{\partial \boldsymbol{\gamma}} = \sum_{h=1}^{k} \frac{1}{\tau_h^2} \boldsymbol{X}_h' \operatorname{diag}(\boldsymbol{z}_h)(\boldsymbol{y} - \boldsymbol{X}_h \boldsymbol{\gamma}),$$

which, for fixed $\boldsymbol{\tau}^2 = (\tau_1^2 \cdots \tau_k^2)'$, is solved by

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{M}^{-1} \boldsymbol{N} \boldsymbol{y},$$

where

$$\boldsymbol{M} = \sum_{h=1}^{k} \frac{1}{\tau_h^2} \boldsymbol{X}_h' \operatorname{diag}(\boldsymbol{z}_h) \boldsymbol{X}_h \quad \text{and} \quad \boldsymbol{N} = \sum_{h=1}^{k} \frac{1}{\tau_h^2} \boldsymbol{X}_h' \operatorname{diag}(\boldsymbol{z}_h). \tag{4}$$

From $\hat{\boldsymbol{\gamma}}$ we directly obtain $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$; we may also obtain $\hat{\beta}_0$ as $\frac{1}{k} \sum_{h=1}^{k} \hat{\pi}_h \hat{\alpha}_h$ and, for $h = 1, \ldots, k$, $\hat{v}_h$ as $\hat{\alpha}_h - \hat{\beta}_0$. Finally, for fixed $\boldsymbol{\gamma}$, the derivative of l_{c2} with respect to $\boldsymbol{\tau}^2$ is solved by $\hat{\boldsymbol{\tau}}^2$ whose elements are

$$\hat{\tau}_h^2 = \frac{1}{z_{\cdot h}} (\boldsymbol{y} - \boldsymbol{X}_h \boldsymbol{\gamma})' \operatorname{diag}(\boldsymbol{z}_h)(\boldsymbol{y} - \boldsymbol{X}_h \boldsymbol{\gamma}), \quad h = 1, \ldots, k. \tag{5}$$

The EM algorithm consists in iterating the following two steps until convergence:

(E) On the basis of the current estimate of the parameters, compute the expected value of the complete log-likelihood given the observed data, $\mathrm{E}(l_c|\boldsymbol{y})$. In practice, this consists in substituting to any z_{ih} in (3) its conditional expected value

$$p_{ih} = \mathrm{E}(z_{ih}|\boldsymbol{y}) = \frac{\pi_h \phi(y_i; \mu_{ih}, \tau_h^2)}{\sum_{g=1}^{k} \pi_g \phi(y_i; \mu_{ig}, \tau_g^2)}.$$

(M) Maximize $\mathrm{E}(l_c|\boldsymbol{y})$ with respect to the parameters of the model as follows:
(i) for any h update the estimate of π_h with $\frac{1}{n} \sum_{i=1}^{n} p_{ih}$;
(ii) iteratively update, until convergence, the estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}^2$ through, respectively, (4) and (5) where any z_{ih} has been substituted with p_{ih}.

A well-known problem is that the likelihood of a mixture of normal distributions with heteroscedastic components is unbounded. To avoid this problem, we constrained the variances τ_h^2 so that the ratio between the largest and the smallest variance is less than a certain value, say 100 as in the simulation presented in the next section. Note however, that the algorithm above usually converges to a local maximum of the likelihood and so this bound is seldom reached in practice; this reduces the problem of the arbitrariness of the bound.

Since the log-likelihood may have more than one local maximum, a crucial point concerns the choice of the starting values of the EM algorithm. We suggest the following strategy. For $\boldsymbol{\beta}$ simply use the corresponding OLS estimate. For $\boldsymbol{\alpha}$ use the vector with elements $\tilde{\beta}_0 + \tilde{v}_h$, $h = 1, \ldots, k$, where $\tilde{\beta}_0$ denotes the OLS estimate of β_0 and $\tilde{v}_h$ the estimate of v_h obtained by fitting the mixture model (2) to the OLS residuals, which also provides the initial values of the parameters τ_h^2 and π_h, $h = 1, \ldots, k$. This may be performed by using an EM algorithm similar to the one described in the previous section. We suggest to initialize this algorithm through a preliminary hierarchical clustering of the residuals based on the complete linkage distance measure between clusters (see Gordon, 1999, Section 4.2.2).

4. Simulation study

To assess the advantages of the proposed approach, we carried out a simulation study in which, apart from the standard normal, the standardized versions of the following distributions are considered for the error terms:

Distribution	Density function $f(\varepsilon)$	Skewness γ_1	Kurtosis γ_2
Extreme value	$\exp[\varepsilon - \exp(-\varepsilon)]$	1.14	2.40
Gamma $(\alpha = 2, \beta = 1)$	$\varepsilon \exp(-\varepsilon)$	1.41	3.00
Lognormal $(\mu = 0, \sigma^2 = 1)$	$\frac{1}{\varepsilon\sqrt{2\pi}} \exp\{-[\log(\varepsilon)]^2/2\}$	6.18	110.94
t (5 d.f.)	$\frac{2}{\Gamma(2.5)\sqrt{5\pi}}(1 + \varepsilon^2/5)^{-3}$	0.00	6.00
Mixture	$0.5\phi(\varepsilon; -1.75, 1) + 0.5\phi(\varepsilon; 1.75, 1)$	0.00	−1.14

We also considered two different regression functions. The first one, indicated hereafter by R_1, involves two non-random predictors ($p = 2$) defined as $x_{i1} = t_i$ and $x_{i2} = t_i^2$, $i = 1, \ldots, n$, where $t_i = (2i - 1)/n - 1$. The true values of the parameters are $\beta_0 = -1$, $\beta_1 = 2$ and $\beta_2 = 3$. In the second case, instead, we have only one random predictor since the regression function, which will be referred to as R_2, is based on an autoregressive structure of order one (AR1). The true values of the parameters are $\beta_0 = 1$ and $\beta_1 = 0.95$. In summary, we have

$$R_1\colon\ Y_i = -1 + 2x_{i1} + 3x_{i2} + \varepsilon_i,$$

$$R_2\colon\ Y_i = 1 + 0.95Y_{i-1} + \varepsilon_i.$$

Table 1
Power of the test of Looney and Gulledge (1985) for regression function R_1

Nominal level of significance	Normal			Extreme value			Gamma		
	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
0.50	0.498	0.518	0.502	0.782	0.924	0.996	0.886	0.992	1.000
0.25	0.257	0.261	0.253	0.582	0.841	0.981	0.760	0.975	1.000
0.10	0.112	0.101	0.112	0.396	0.725	0.954	0.592	0.922	0.999
0.05	0.059	0.043	0.067	0.310	0.640	0.907	0.477	0.864	0.999
0.01	0.015	0.016	0.014	0.155	0.408	0.783	0.267	0.660	0.983
Nominal level of significance	Lognormal			t			Mixture		
	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
0.50	0.987	1.000	1.000	0.688	0.755	0.918	0.714	0.909	0.995
0.25	0.970	1.000	1.000	0.470	0.605	0.827	0.461	0.748	0.984
0.10	0.933	0.999	1.000	0.305	0.464	0.745	0.193	0.499	0.939
0.05	0.888	0.995	1.000	0.217	0.359	0.682	0.085	0.328	0.857
0.01	0.761	0.993	1.000	0.102	0.206	0.491	0.007	0.081	0.529

Table 2
Power of the test of Looney and Gulledge (1985) for regression function R_2

Nominal level of significance	Normal			Extreme value			Gamma		
	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
0.50	0.495	0.528	0.479	0.764	0.955	0.999	0.914	0.994	1.000
0.25	0.262	0.252	0.230	0.617	0.870	0.988	0.795	0.978	1.000
0.10	0.102	0.101	0.091	0.442	0.742	0.961	0.637	0.935	1.000
0.05	0.061	0.046	0.042	0.347	0.625	0.931	0.529	0.887	0.999
0.01	0.010	0.012	0.010	0.197	0.411	0.807	0.308	0.749	0.984
Nominal level of significance	Lognormal			t			Mixture		
	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
0.50	0.996	1.000	1.000	0.669	0.770	0.900	0.782	0.942	0.999
0.25	0.982	1.000	1.000	0.478	0.616	0.816	0.501	0.802	0.990
0.10	0.951	1.000	1.000	0.333	0.495	0.695	0.246	0.562	0.953
0.05	0.921	0.999	1.000	0.246	0.401	0.618	0.119	0.394	0.899
0.01	0.825	0.994	1.000	0.134	0.234	0.459	0.014	0.108	0.607

For each distribution of the error terms and each regression function, we generated 1000 samples of three different sizes ($n = 25, 50, 100$) from the resulting model. For any sample, the hypothesis of normality of the error terms has been assessed through the correlation test described in Section 2. The power of such a test (i.e. the relative frequency of times that the null hypothesis is rejected) is shown in Tables 1 and 2 for several nominal levels of significance.

Table 3
Comparison between OLS, ML and proposed estimator of the parameters for the regression function R_1

			Normal			Extreme value			Gamma		
			$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
OLS	$\hat{\beta}_0$	Bias	−0.0133	−0.0187	0.0076	−0.0115	0.0090	−0.0014	−0.0174	−0.0018	−0.0015
		MSE	0.0890	0.0462	0.0221	0.0914	0.0408	0.0212	0.0910	0.0437	0.0215
	$\hat{\beta}_1$	Bias	0.0259	0.0013	−0.0078	−0.0021	−0.0073	−0.0013	0.0035	0.0120	−0.0018
		MSE	0.1203	0.0594	0.0281	0.1201	0.0576	0.0297	0.1191	0.0598	0.0320
	$\hat{\beta}_2$	Bias	0.0428	0.0221	−0.0120	0.0326	−0.0282	0.0030	0.0163	0.0122	0.0018
		MSE	0.4330	0.2178	0.1119	0.4727	0.2069	0.1040	0.4478	0.2161	0.1032
MLE	$\hat{\beta}_0$	Bias	−0.0133	−0.0187	0.0076	−0.0473	−0.0095	−0.0100	−0.0599	−0.0178	−0.0139
		MSE	0.0890	0.0462	0.0221	0.0725	0.0324	0.0168	0.0631	0.0282	0.0134
	$\hat{\beta}_1$	Bias	0.0259	0.0013	−0.0078	−0.0011	−0.0002	0.0028	0.0072	0.0044	0.0011
		MSE	0.1203	0.0594	0.0281	0.0903	0.0376	0.0194	0.0502	0.0181	0.0076
	$\hat{\beta}_2$	Bias	0.0428	0.0221	−0.0120	0.1125	0.0146	0.0229	0.1438	0.0602	0.0389
		MSE	0.4330	0.2178	0.1119	0.3315	0.1440	0.0654	0.2074	0.0803	0.0319
Mixture ($k=2$)	$\hat{\beta}_0$	Bias	−0.0143	−0.0184	0.0074	−0.0211	0.0038	−0.0058	−0.0328	−0.0042	−0.0039
		MSE	0.0912	0.0468	0.0226	0.0885	0.0390	0.0197	0.0880	0.0394	0.0173
	$\hat{\beta}_1$	Bias	0.0231	0.0009	−0.0079	−0.0025	−0.0041	0.0015	−0.0001	0.0052	0.0042
		MSE	0.1211	0.0601	0.0286	0.1186	0.0551	0.0239	0.1081	0.0408	0.0164
	$\hat{\beta}_2$	Bias	0.0458	0.0211	−0.0111	0.0613	−0.0127	0.0163	0.0624	0.0195	0.0091
		MSE	0.4519	0.2260	0.1143	0.4600	0.1978	0.0875	0.4231	0.1647	0.0677
Mixture ($k=3$)	$\hat{\beta}_0$	Bias	−0.0141	−0.0176	0.0078	−0.0231	0.0073	−0.0065	−0.0297	−0.0023	−0.0070
		MSE	0.0909	0.0465	0.0229	0.0912	0.0412	0.0204	0.0878	0.0389	0.0177
	$\hat{\beta}_1$	Bias	0.0231	0.0027	−0.0071	−0.0011	−0.0023	0.0042	−0.0069	−0.0015	0.0026
		MSE	0.1218	0.0613	0.0291	0.1232	0.0616	0.0268	0.1114	0.0450	0.0155
	$\hat{\beta}_2$	Bias	0.0452	0.0185	−0.0125	0.0674	−0.0230	0.0184	0.0530	0.0138	0.0183
		MSE	0.4511	0.2248	0.1182	0.4851	0.2164	0.0947	0.4251	0.1693	0.0672

			Lognormal			t			Mixture		
			$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
OLS	$\hat{\beta}_0$	Bias	−0.0121	0.0069	−0.0044	0.0081	−0.0025	−0.0002	0.0070	−0.0028	0.0028
		MSE	0.0746	0.0425	0.0211	0.0851	0.0483	0.0215	0.0843	0.0443	0.0239
	$\hat{\beta}_1$	Bias	−0.0062	0.0010	0.0023	−0.0064	−0.0044	0.0029	−0.0102	−0.0026	−0.0016
		MSE	0.0873	0.0581	0.0261	0.1207	0.0586	0.0297	0.1099	0.0595	0.0310
	$\hat{\beta}_2$	Bias	0.0133	−0.0045	−0.0010	−0.0142	0.0111	−0.0105	−0.0218	0.0022	−0.0077
		MSE	0.3683	0.2198	0.0988	0.4416	0.2340	0.1046	0.4572	0.2182	0.1143
MLE	$\hat{\beta}_0$	Bias	−0.0613	−0.0297	−0.0207	0.0032	0.0006	−0.0026	0.0004	−0.0021	−0.0002
		MSE	0.0309	0.0144	0.0070	0.0731	0.0388	0.0176	0.0447	0.0184	0.0090
	$\hat{\beta}_1$	Bias	0.0002	−0.0016	0.0015	−0.0136	−0.0027	0.0025	−0.0053	−0.0009	−0.0010
		MSE	0.0070	0.0025	0.0008	0.1070	0.0514	0.0251	0.0805	0.0302	0.0128
	$\hat{\beta}_2$	Bias	0.0593	0.0275	0.0141	−0.0041	0.0047	−0.0021	−0.0038	0.0042	−0.0014
		MSE	0.0336	0.0108	0.0033	0.3904	0.1944	0.0869	0.3251	0.1165	0.0501
Mixture ($k=2$)	$\hat{\beta}_0$	Bias	−0.0184	0.0039	−0.0042	0.0082	−0.0006	−0.0009	0.0062	−0.0008	0.0046
		MSE	0.0491	0.0263	0.0124	0.0840	0.0465	0.0213	0.0851	0.0421	0.0193
	$\hat{\beta}_1$	Bias	0.0062	−0.0016	0.0008	−0.0095	−0.0014	0.0030	−0.0106	−0.0025	0.0000
		MSE	0.0359	0.0157	0.0063	0.1206	0.0559	0.0281	0.1108	0.0558	0.0188
	$\hat{\beta}_2$	Bias	0.0321	0.0042	−0.0015	−0.0144	0.0053	−0.0085	−0.0194	−0.0039	−0.0131
		MSE	0.1451	0.0564	0.0237	0.4314	0.2251	0.1010	0.4603	0.2033	0.0756
Mixture ($k=3$)	$\hat{\beta}_0$	Bias	−0.0222	0.0019	−0.0056	0.0078	−0.0019	−0.0035	0.0058	0.0006	0.0053
		MSE	0.0475	0.0232	0.0109	0.0866	0.0483	0.0225	0.0857	0.0430	0.0201
	$\hat{\beta}_1$	Bias	0.0067	−0.0009	0.0017	−0.0050	−0.0010	0.0003	−0.0120	0.0008	−0.0018
		MSE	0.0360	0.0078	0.0032	0.1260	0.0620	0.0302	0.1116	0.0575	0.0227
	$\hat{\beta}_2$	Bias	0.0438	0.0103	0.0027	−0.0135	0.0091	−0.0008	−0.0184	−0.0080	−0.0151
		MSE	0.1444	0.0309	0.0125	0.4470	0.2418	0.1073	0.4640	0.2127	0.0858

Table 4
Comparison between OLS, ML and proposed estimator of the parameters for the regression function R_2

			Normal			Extreme value			Gamma		
			$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
OLS	$\hat{\beta}_0$	Bias	0.2617	0.2015	0.1756	0.2451	0.2144	0.1668	0.2374	0.1780	0.1951
		MSE	0.3010	0.1905	0.1312	0.3670	0.2248	0.1667	0.3357	0.1927	0.1672
	$\hat{\beta}_1$	Bias	−0.0298	−0.0157	−0.0109	−0.0241	−0.0151	−0.0099	−0.0247	−0.0129	−0.0120
		MSE	0.0040	0.0011	0.0005	0.0038	0.0012	0.0006	0.0036	0.0010	0.0006
MLE	$\hat{\beta}_0$	Bias	0.2617	0.2015	0.1756	0.1742	0.1415	0.1204	0.1139	0.0832	0.0941
		MSE	0.3010	0.1905	0.1312	0.2564	0.1340	0.1032	0.1677	0.0709	0.0547
	$\hat{\beta}_1$	Bias	−0.0298	−0.0157	−0.0109	−0.0175	−0.0100	−0.0073	−0.0121	−0.0061	−0.0058
		MSE	0.0040	0.0011	0.0005	0.0027	0.0007	0.0003	0.0015	0.0003	0.0002
Mixture ($k=2$)	$\hat{\beta}_0$	Bias	0.2599	0.1999	0.1757	0.2269	0.1862	0.1285	0.2017	0.1245	0.1295
		MSE	0.3029	0.1910	0.1321	0.3466	0.1942	0.1166	0.3016	0.1416	0.0971
	$\hat{\beta}_1$	Bias	−0.0296	−0.0156	−0.0110	−0.0224	−0.0131	−0.0077	−0.0211	−0.0090	−0.0080
		MSE	0.0040	0.0011	0.0005	0.0036	0.0010	0.0004	0.0032	0.0007	0.0003
Mixture ($k=3$)	$\hat{\beta}_0$	Bias	0.2576	0.1972	0.1770	0.2278	0.1736	0.1172	0.2041	0.1145	0.1102
		MSE	0.3028	0.1888	0.1327	0.3585	0.1994	0.1235	0.3039	0.1397	0.0908
	$\hat{\beta}_1$	Bias	−0.0293	−0.0154	−0.0110	−0.0226	−0.0121	−0.0070	−0.0213	−0.0083	−0.0068
		MSE	0.0040	0.0011	0.0005	0.0038	0.0011	0.0004	0.0032	0.0007	0.0003

			Lognormal			t			Mixture		
			$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
OLS	$\hat{\beta}_0$	Bias	0.2077	0.1843	0.1381	0.2584	0.1850	0.1692	0.2296	0.1969	0.1583
		MSE	0.7479	0.2764	0.1971	0.3267	0.1808	0.1509	0.2526	0.1910	0.1396
	$\hat{\beta}_1$	Bias	−0.0186	−0.0131	−0.0083	−0.0287	−0.0140	−0.0103	−0.0251	−0.0153	−0.0099
		MSE	0.0036	0.0013	0.0006	0.0041	0.0010	0.0005	0.0032	0.0011	0.0005
MLE	$\hat{\beta}_0$	Bias	−0.0100	0.0040	0.0065	0.2204	0.1482	0.1308	0.1301	0.0904	0.0775
		MSE	0.0376	0.0192	0.0100	0.2841	0.1425	0.1077	0.1527	0.0865	0.0623
	$\hat{\beta}_1$	Bias	−0.0022	−0.0016	−0.0010	−0.0247	−0.0115	−0.0080	−0.0133	−0.0070	−0.0047
		MSE	0.0001	0.0000	0.0000	0.0036	0.0008	0.0004	0.0018	0.0005	0.0002
Mixture ($k=2$)	$\hat{\beta}_0$	Bias	0.1054	0.0604	0.0410	0.2376	0.1640	0.1493	0.2226	0.1638	0.0906
		MSE	0.5528	0.0688	0.0359	0.3039	0.1616	0.1339	0.2487	0.1671	0.0910
	$\hat{\beta}_1$	Bias	−0.0097	−0.0043	−0.0025	−0.0265	−0.0125	−0.0091	−0.0242	−0.0128	−0.0056
		MSE	0.0022	0.0002	0.0001	0.0039	0.0009	0.0005	0.0031	0.0010	0.0003
Mixture ($k=3$)	$\hat{\beta}_0$	Bias	0.0884	0.0383	0.0197	0.2442	0.1560	0.1371	0.2220	0.1643	0.0938
		MSE	0.5362	0.0483	0.0199	0.3163	0.1702	0.1346	0.2501	0.1703	0.1058
	$\hat{\beta}_1$	Bias	−0.0080	−0.0028	−0.0012	−0.0273	−0.0119	−0.0083	−0.0242	−0.0128	−0.0058
		MSE	0.0020	0.0002	0.0000	0.0041	0.0010	0.0005	0.0031	0.0010	0.0004

Table 5
Comparison between OLS confidence intervals and confidence intervals computed according to the proposed approach for the regression function R_1

			Normal			Extreme value			Gamma		
			$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
OLS	$\hat{\beta}_0$	Width	1.164	0.829	0.589	1.146	0.820	0.585	1.161	0.824	0.585
		Coverage	0.929	0.944	0.953	0.921	0.940	0.957	0.942	0.942	0.955
	$\hat{\beta}_1$	Width	1.344	0.957	0.680	1.323	0.946	0.676	1.339	0.952	0.676
		Coverage	0.945	0.955	0.950	0.934	0.944	0.951	0.939	0.947	0.952
	$\hat{\beta}_2$	Width	2.611	1.855	1.318	2.570	1.834	1.309	2.602	1.844	1.309
		Coverage	0.938	0.941	0.952	0.931	0.947	0.958	0.950	0.940	0.938
Mixture	$\hat{\beta}_0$	Width	1.166	0.832	0.592	1.149	0.827	0.597	1.157	0.807	0.552
		Coverage	0.929	0.944	0.951	0.926	0.954	0.964	0.944	0.948	0.958
	$\hat{\beta}_1$	Width	1.353	0.965	0.684	1.365	0.964	0.673	1.364	0.913	0.609
		Coverage	0.945	0.955	0.953	0.936	0.952	0.969	0.945	0.967	0.983
	$\hat{\beta}_2$	Width	2.633	1.871	1.326	2.656	1.889	1.320	2.680	1.809	1.191
		Coverage	0.937	0.944	0.949	0.940	0.954	0.971	0.953	0.963	0.978

			Lognormal			t			Mixture		
			$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
OLS	$\hat{\beta}_0$	Width	1.028	0.734	0.535	1.135	0.830	0.580	1.178	0.827	0.588
		Coverage	0.892	0.918	0.931	0.934	0.951	0.957	0.948	0.955	0.944
	$\hat{\beta}_1$	Width	1.186	0.848	0.618	1.310	0.958	0.670	1.360	0.955	0.679
		Coverage	0.955	0.954	0.958	0.943	0.946	0.925	0.944	0.941	0.945
	$\hat{\beta}_2$	Width	2.304	1.643	1.196	2.545	1.856	1.297	2.641	1.851	1.316
		Coverage	0.941	0.942	0.960	0.935	0.941	0.943	0.948	0.945	0.942
Mixture	$\hat{\beta}_0$	Width	0.888	0.593	0.422	1.141	0.837	0.587	1.176	0.807	0.537
		Coverage	0.897	0.910	0.937	0.943	0.953	0.955	0.949	0.952	0.942
	$\hat{\beta}_1$	Width	1.001	0.540	0.369	1.361	0.970	0.690	1.364	0.909	0.556
		Coverage	0.959	0.984	0.988	0.942	0.951	0.941	0.943	0.936	0.950
	$\hat{\beta}_2$	Width	1.909	1.064	0.718	2.631	1.902	1.328	2.658	1.765	1.089
		Coverage	0.964	0.976	0.984	0.945	0.952	0.951	0.946	0.939	0.942

Obviously, when the true distribution of the error terms is the normal one, the power of the correlation test for normality is very close to the nominal level of significance. In the other cases the power is greater, especially for highly skewed and/or leptokurtic distributions such as the Gamma and the Lognormal. As we may expect, the power also increases with the sample size.

For any sample, the regression parameters (β_0, β_1, β_2 for R_1 and β_0, β_1 for R_2) are estimated either through the mixture-based method described in Section 3 or the standard OLS method, according to whether the hypothesis of normality is rejected or not. We chose 0.05 as a level of significance for such a test. Tables 3 and 4 show the bias and the mean-square error (MSE) of the resulting estimator for a number of components of the mixture (k) equal to 2 and 3; these tables also show the bias

Table 6
Comparison between OLS confidence intervals and confidence intervals computed according to the proposed approach for the regression function R_2

			Normal			Extreme value			Gamma		
			$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
OLS	$\hat{\beta}_0$	Width	1.694	1.401	1.228	1.662	1.416	1.234	1.664	1.390	1.213
		Coverage	0.925	0.916	0.930	0.906	0.911	0.895	0.923	0.926	0.922
	$\hat{\beta}_1$	Width	0.186	0.105	0.073	0.181	0.103	0.072	0.176	0.101	0.072
		Coverage	0.914	0.915	0.921	0.895	0.910	0.907	0.920	0.922	0.924
Mixture	$\hat{\beta}_0$	Width	1.701	1.403	1.234	1.680	1.417	1.249	1.687	1.375	1.143
		Coverage	0.924	0.914	0.930	0.902	0.927	0.926	0.925	0.940	0.944
	$\hat{\beta}_1$	Width	0.187	0.105	0.073	0.180	0.102	0.073	0.175	0.097	0.066
		Coverage	0.914	0.913	0.921	0.887	0.927	0.938	0.910	0.950	0.944
			Lognormal			t			Mixture		
			$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$	$n=25$	$n=50$	$n=100$
OLS	$\hat{\beta}_0$	Width	1.405	1.279	1.129	1.672	1.369	1.203	1.700	1.393	1.216
		Coverage	0.891	0.896	0.902	0.917	0.914	0.946	0.919	0.932	0.936
	$\hat{\beta}_1$	Width	0.144	0.090	0.066	0.181	0.101	0.071	0.186	0.103	0.072
		Coverage	0.899	0.908	0.916	0.910	0.930	0.944	0.911	0.928	0.928
Mixture	$\hat{\beta}_0$	Width	1.319	1.033	0.929	1.690	1.395	1.186	1.663	1.294	0.964
		Coverage	0.917	0.954	0.970	0.920	0.922	0.946	0.914	0.926	0.930
	$\hat{\beta}_1$	Width	0.122	0.064	0.051	0.182	0.103	0.070	0.180	0.093	0.055
		Coverage	0.936	0.972	0.970	0.912	0.928	0.956	0.902	0.912	0.926

and the MSE of the OLS estimator as well as those of the MLE based on the true distribution of the error terms.

As we may expect, the MLE performs much better than the OLS estimator for any distribution we considered, apart from the normal one. In these situations also the proposed estimator performs better than the OLS estimator, especially when the true distribution of the error terms is highly skewed and/or leptokurtic. Obviously, our estimator generally performs worse than the MLE; the latter, however, requires the knowledge of true distribution of the error terms. Finally, note that the number of components of the mixture does not affect significantly the MSE of the proposed estimator of the regression parameters and so we suggest to use $k=2$ components. This is the most favorable situation from the point of view of the parameter estimation: the EM algorithm in Section 3 is fast and the problem of the choice of its starting values is negligible. A similar result has been obtained for the homoscedastic case ($\tau_h^2=\tau^2$, $\forall h$) considered in the previous version of the paper (see tables available at http://stat.unipg.it/~luisa) in which we have taken into account mixtures with a number of components up to 5.

For each simulated sample we also computed a 95% confidence interval for any regression parameter. When we reject the hypothesis of normality, these intervals are computed through a bootstrap method (Efron and Tibshirani, 1993) based on 500

subsamples; in this case, only mixtures of $k = 2$ components have been considered. Tables 5 and 6 show the average width of these intervals together with the coverage probability.

Note that the actual coverage probability of the confidence intervals computed on the basis of the proposed approach is generally larger than that of the intervals computed, as usual, on the basis of the OLS estimates. In many situations, the latter ones are wider; the difference is clear for the Gamma and Lognormal distributions. These results confirm that, when the true distribution of the stochastic component of the regression model departs from normality, the inference on the parameters can significantly benefit from the use of mixtures of normal distributions to model such a component.

Acknowledgements

The authors are grateful to a referee and an associated editor for very helpful comments. They also would like to acknowledge the financial support provided by the found M.I.U.R. 2002.

References

Aitkin, M., Wilson, G.T., 1980. Mixture models, outliers, and the EM algorithm. Technometrics 22, 325–331.

Anscombe, F.J., 1961. Examination of residuals. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 1–36.

Bilmes, J.A., 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, TR-97-021. Department of Electrical Engineering and Computer Science, UC Berkeley.

Blom, G., 1958. Statistical Estimates and Transformed Beta Variables. Wiley, New York.

Boos, D., 1987. Detecting skewed errors from regression residuals. Technometrics 29, 83–90.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformation (with discussion). J. Roy. Statist. Soc. Ser. B 26, 211–252.

Cox, D.R., Hinkley, D.V., 1968. A note on the efficiency of least squares estimates. J. Roy. Statist. Soc. Ser. B 30, 284–289.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–22.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, London.

Gan, F.F., Koehler, K.J., 1990. Goodness-of-fit tests based on P–P probability plots. Technometrics 32, 289–303.

Gan, F.F., Koehler, K.J., Thompson, J.C., 1991. Probability plots and distribution curves for assessing the fit of probability-models. Amer. Statist. 45, 14–21.

Gordon, A.D., 1999. Classification. Chapman & Hall, London.

Huang, C.J., Bolch, B.W., 1974. On the testing of regression disturbances for normality. J. Amer. Statist. Assoc. 69, 330–335.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Looney, S.W., Gulledge, T.R., 1985. Use of the correlation coefficient with normal probability plots. Amer. Statist. 39, 75–79.

Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Ann. Statist. 20, 712–736.

Quandt, R.E., Ramsey, J.B., 1978. Estimating mixtures of normal distributions and switching regressions. J. Amer. Statist. Assoc. 73, 730–738.

Quesenberry, C.P., 1986. Some transformation methods in goodness-of-fit. In: Stephens, M.A., D'Agostino, R.B. (Eds.), Goodness-of-Fit Techniques, Marcel Dekker, New York.