Comparing Robust Regression Estimators to Detect Data Clusters: A Case Study

Alessandra Durio and Ennio Isaia

Abstract It is well known that in all situations involving the study of large data sets where a substantial number of outliers or clustered data are present, regression models based on M-estimators are likely to be unstable. Resorting to the inherent properties of robustness of the estimates based on the Integrated Square Error criterion we compare the results arising from L_2 estimates with those obtained from some common M-estimators. The discrepancy between the estimated regression models is measured by means of a new concept of similarity between functions and a system of statistical hypothesis. A Monte Carlo Significance test, is introduced to test the similarity of the estimates. Whenever the hypothesis of similarity between models is rejected, a careful investigation of the data structure is necessary to check for the presence of clusters, which can lead to the consideration of a mixture of regression models. Concerning this, we shall see how L_2 criterion can be applied in fitting a finite mixture of regression models. The requisite theory is outlined and the whole procedure is applied to a case study concerning the evaluation of the risk of fire and the risk of electric shocks of electronic transformers.

Keywords Minimum integrated square error • Mixture of regression models • Robust regression • Similarity between functions

1 Introduction

Regression is one of the widespread tools used to establish the relationship between a set of predictor variables and a response variable. However, in many circumstances, careful data preparation may not be possible and hence data may

A. Durio (✉)
Department of Economics & Statistics "S. Cognetti de Martiis", University of Turin, Lungo Dora Siena 100/A, 10124 Turin, Italy
e-mail: alessandra.durio@unito.it

G. MacKenzie and D. Peng (eds.), *Statistical Modelling in Biostatistics and Bioinformatics*, Contributions to Statistics, DOI 10.1007/978-3-319-04579-5_10,

be heavily contaminated by a substantial number of outliers. In these situations, the estimates of the parameters of the regression model obtained by the Maximum Likelihood criterion are fairly unstable.

The development of robust methods is underlined by the appearance of a wide number of papers and books on the topic including: Huber (1981), Rousseeuw and Leroy (1987), Staudte and Sheather (1990), Davies (1993), Dodge and Jurečkova (2000), Seber and Lee (2003), Rousseeuw et al. (2004), Jurečkova and Picek (2006), Maronna et al. (2006) and Fujisawa and Eguchi (2006).

The approach based on minimizing the Integrated Square Error is particularly helpful in those situations where, due to large sample size, careful data preparation is not feasible and hence data may contain a substantial number of outliers (Scott 2001). In this sense the L_2E criterion can be viewed as an efficient diagnostic tool in building useful models.

In this paper we suggest a procedure of regression analysis whose first step consists in comparing the results arising from L_2 estimates with those obtained from some common M-estimators. Afterwards, if a particular test of hypothesis leads us to reject the conjecture of similarity between the estimated regression models, we investigate the data for the presence of clusters by analyzing the L_2 minimizing function. The third step of the procedure consists in fitting a mixture of regression models via the L_2 criterion.

Below, we introduce the Integrated Square Error minimizing criterion for regression models, define a new concept of similarity between functions and introduce a Monte Carlo Significance (M.C.S.) test. We also illustrate the whole procedure by means of some simulated examples involving simple linear regression models. Finally, we present an analysis of a case study concerning the evaluation of the risk of fire and the risk of electric shocks in electronic transformers.

2 Parametric Linear Regression Models and Robust Estimators

Let $\{(x_{i1}, \ldots, x_{ip}, y_i)\}_{i=1,\ldots,n}$ be the observed data set, where each observation stems from a random sample drawn from the $p + 1$ random variable $(X_1, \ldots, X_p, Y)$. The regression model for the observed data set being studied is $y_i = m_{\boldsymbol{\beta}}(\mathbf{x}_i) + \varepsilon_i$, with $i = 1, \ldots, n$, where the object of our interest is the regression mean

$$m_{\boldsymbol{\beta}}(\mathbf{x}_i) = \mathbb{E}[Y|\mathbf{x}_i] = \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_{ij} \tag{1}$$

and the errors $\{\varepsilon_i\}_{i=1,\ldots,n}$ are assumed to be independent random variables with zero mean and unknown finite variances.

2.1 Huber M-Estimator

The presence of outliers is a problem for regression techniques; these may occur for many reasons. An extreme situation arises when the outliers are numerous and they arise as a consequence of clustered data. For example, a large proportion of outliers may be found, if there is an omitted unknown categorical variable (e.g. gender, species, geographical location, etc.) where the data behave differently in each category. In parametric estimation, the estimators with good robustness proprieties relative to maximum likelihood are the M-estimators. The class of M-estimators of the vector $\boldsymbol{\beta}$ is defined as (e.g., Hampel et al. 2005)

$$\hat{\boldsymbol{\beta}}_M = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)\right), \tag{2}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is absolutely continuous convex function with derivative ψ.

If we assume that the r.v.s ε_i are independent and identically distributed as the r.v. $\varepsilon \sim \mathcal{N}(0, \sigma)$, the least-squares estimator gives the Maximum Likelihood Estimate (*MLE*) of the vector $\boldsymbol{\beta}$, i.e.:

$$\hat{\boldsymbol{\beta}}_{MLE} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)\right]^2.$$

Since in the presence of outliers MLEs are quite unstable, i.e., inefficient and biased, for our purpose in the class of M-estimators we shall resort to the robust Huber M-estimator (*HME*) for which

$$\rho(y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)) = \begin{cases} \dfrac{1}{2}(y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i))^2 & \text{if } |y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)| \le k, \\ k\,|y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)|\,(1 - \dfrac{k}{2}) & \text{if } |y_i - m_{\boldsymbol{\beta}}(\mathbf{x}_i)| > k, \end{cases}$$

where the tuning constant k is generally set to $1.345\,\sigma$.

2.2 L_2-Based Estimator

We investigate estimation methods in parametric linear regression models based on the minimum Integrated Square Error and the minimum L_2 metric. In the α-family of estimators proposed by Basu et al. (1998), L_2 estimator, briefly L_2E, is the more robust to outliers, even if it is less efficient than *MLE*.

Given the r.v. X, with unknown density $f(x|\boldsymbol{\theta}_0)$, for which we introduce the model $f(x|\boldsymbol{\theta})$, the estimate for $\boldsymbol{\theta}_0$ minimizing the L_2 metric will be:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{L_2E} &= \arg\min_{\boldsymbol{\theta}} \int_{\mathbb{R}} [f(x|\boldsymbol{\theta}) - f(x|\boldsymbol{\theta}_0)]^2 \, dx = \\ &= \arg\min_{\boldsymbol{\theta}} \left[\int_{\mathbb{R}} f^2(x|\boldsymbol{\theta}) \, dx - 2\,\mathbb{E}\,[f(x|\boldsymbol{\theta}_0)] \right] = \\ &= \arg\min_{\boldsymbol{\theta}} \left[\int_{\mathbb{R}} f^2(x|\boldsymbol{\theta}) \, dx - \frac{2}{n} \sum_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \right],\end{aligned} \tag{3}$$

where, the so-called expected height of the density, $\mathbb{E}\,[f(x|\boldsymbol{\theta}_0)]$ is replaced with its estimate $\hat{\mathbb{E}}\,[f(x|\boldsymbol{\theta}_0)] = n^{-1} \sum_{i=1}^{n} f(x_i|\boldsymbol{\theta})$ and where (Basu et al. 1998),

$$\int_{\mathbb{R}} f^2(x|\boldsymbol{\theta}) \, dx = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} f^2(x_i|\boldsymbol{\theta}) \, dx_i \; . \tag{4}$$

We turn now our attention to illustrate how the estimates based on L_2 criterion can be applied to parametric regression models. Assuming that the random variables $Y|\mathbf{x}$ are distributed as a $\mathcal{N}(m_{\boldsymbol{\beta}_0}(\mathbf{x}), \sigma_0)$, i.e. $f_{Y|\mathbf{x}}(y|\boldsymbol{\beta}_0, \sigma_0) = \phi(y|m_{\boldsymbol{\beta}_0}(\mathbf{x}), \sigma_0)$, the L_2 estimates of the parameters in $\boldsymbol{\beta}_0$ and σ_0 are given by Eq. (3), which in this case becomes

$$\begin{aligned}(\hat{\boldsymbol{\beta}}, \hat{\sigma})_{L_2E} &= \arg\min_{\boldsymbol{\beta},\sigma} \left[\int_{\mathbb{R}} \phi^2(y|m_{\boldsymbol{\beta}}(\mathbf{x}), \sigma) \, dy - \frac{2}{n} \sum_{i=1}^{n} \phi(y_i|m_{\boldsymbol{\beta}}(\mathbf{x}_i), \sigma) \right] \\ &= \arg\min_{\boldsymbol{\beta},\sigma} \left[\frac{1}{2\sigma\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^{n} \phi(y_i|m_{\boldsymbol{\beta}}(\mathbf{x}_i), \sigma) \right],\end{aligned} \tag{5}$$

since from Eq. (4)

$$\int_{\mathbb{R}} \phi^2(y|m_{\boldsymbol{\beta}}(\mathbf{x}), \sigma) \, dy = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} \phi^2(y_i|m_{\boldsymbol{\beta}}(\mathbf{x}_i), \sigma) \, dy_i = \frac{1}{2\,\sigma\,\sqrt{\pi}} \; .$$

Clearly Eq. (5) is a feasible computationally closed-form expression so that L_2 criteria can be performed by any standard non-linear optimization procedure, for example, the `nlm` routine in the `R` library. However, it is important to recall that, whatever the algorithm, convergence to the global optimum can depend strongly on the starting values.

3 The Similarity Index and the M.C.S Test

To compare the L_2E performance with respect to some other common estimators we resort to an index of similarity between regression models introduced in Durio and Isaia (2010). In order to measure the discrepancy between the two estimated regression models, the index of similarity takes into account the *space region*

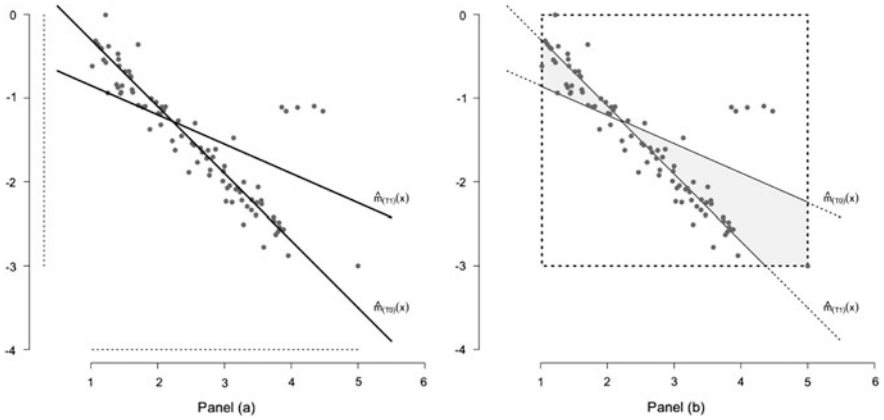

Fig. 1 Data points and two estimated regression models $\hat{m}_{T0}(x)$ and $\hat{m}_{T1}(x)$. In panel (**b**) the domains $\boldsymbol{D}^{p+1}$ and $\boldsymbol{C}^{p+1}$ upon which the $sim(T_0, T_1)$ statistic is computed

between $\hat{m}_{T_0}(\mathbf{x})$ and $\hat{m}_{T_1}(\mathbf{x})$ with respect to the space region where the whole of the data points lie. Let T_0 and T_1 be two regression estimators and $\hat{\boldsymbol{\beta}}_{T_0}$, $\hat{\boldsymbol{\beta}}_{T_1}$ the corresponding vectors of the estimated parameters. Introducing the sets:

$$\begin{aligned} \boldsymbol{I}^p &= \left[\min(x_{i1}); \max(x_{i1})\right] \times \ldots \times \left[\min(x_{ip}); \max(x_{ip})\right], \\ \boldsymbol{I} &= \left[\min(y_i); \max(y_i)\right] = [a; b], \end{aligned}$$

we define the *similarity index* as

$$\begin{aligned} sim(T_0, T_1) &\stackrel{def}{=} \frac{\int_{\boldsymbol{D}^{p+1}} d\mathbf{t}}{\int_{\boldsymbol{C}^{p+1}} d\mathbf{t}} \\ \boldsymbol{C}^{p+1} &= \boldsymbol{I}^p \times \boldsymbol{I} \\ \boldsymbol{D}^{p+1} &= \left\{(\mathbf{x}, y) \in \mathbb{R}^{p+1} : \zeta(\mathbf{x}) \leq y \leq \xi(\mathbf{x}), \mathbf{x} \in \boldsymbol{I}^p\right\} \cap \boldsymbol{C}^{p+1} \end{aligned} \tag{6}$$

with $\zeta(\mathbf{x}) = min\left(\hat{m}_{T_0}(\mathbf{x}), \hat{m}_{T_1}(\mathbf{x})\right)$ and $\xi(\mathbf{x}) = max\left(\hat{m}_{T_0}(\mathbf{x}), \hat{m}_{T_1}(\mathbf{x})\right)$.

Figure 1 shows how the similarity index given by Eq. (6) can be computed in the simple case where $p = 1$. In panel (a) we have the cloud of data points and the two estimated models $\hat{\boldsymbol{\beta}}_{T_0}$ and $\hat{\boldsymbol{\beta}}_{T_1}$. The shaded area of panel (b) corresponds to $\int_{\boldsymbol{D}^{p+1}} d\mathbf{t}$, while the integral $\int_{\boldsymbol{C}^{p+1}} d\mathbf{t}$ is given by the area of the dotted rectangle, in which data points lay.

In order to compute the integrals of Eq. (6), we employ the fast and accurate algorithm proposed by Durio and Isaia (2010).

If the vectors $\hat{\boldsymbol{\beta}}_{T_0}$ and $\hat{\boldsymbol{\beta}}_{T_1}$ are close to each other, then $sim(T_0, T_1)$ will be close to zero. On the other hand, if the estimated regression models $\hat{m}_{T_0}(\mathbf{x})$ and $\hat{m}_{T_1}(\mathbf{x})$ are

dissimilar we are likely to observe a value of $sim(T_0, T_1)$ far from zero. We therefore propose to use the $sim(T_0, T_1)$ statistic to verify the following system of hypothesis

$$\begin{cases} H_0 : \boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_{T_0} \\ H_1 : \boldsymbol{\beta}_0 \neq \hat{\boldsymbol{\beta}}_{T_0} \end{cases} \tag{7}$$

Since it is not reasonable to look for an exact form of the $sim(T_0, T_1)$ distribution, in order to check the above system of hypothesis we utilise a simplified M.C.S. test originally suggested by Barnard (1963) and later proposed by Hope (1968).

Let $sim_{T_0 T_1}$ denote the value of the $sim(T_0, T_1)$ statistic computed on the observed data. The simplified M.C.S. test consists of rejecting H_0 if $sim_{T_0 T_1}$ is the $m\,\alpha$-th most extreme statistic relative to the corresponding quantities based on the random samples of the reference set, where the reference set consists of $m - 1$ random samples, of size n each, generated under the null hypothesis, i.e., drawn at random from the model $\hat{m}_{T_0}(\mathbf{x})$ with $\sigma = \hat{\sigma}_{T_0}$. In other words we generate $m - 1$ random samples under H_0 and for each of them we compute $sim^*_{T_0 T_1}$ and we shall reject the null hypothesis, at the α significance level, if and only if the value of the test statistic $sim_{T_0 T_1}$ is greater than all the $m - 1$ values of $sim^*_{T_0 T_1}$. We remark that if we set $m\,\alpha = 1$ and fix $\alpha = 0.01$, we have $m - 1 = 99$ (while fixing $\alpha = 0.05$ would yield $m - 1 = 19$).

4 Simple Linear Regression and Examples

Since for our case study we shall consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the L_2 criterion according to Eq. (5) reduces to the following computationally closed-form expression

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma})_{L_2 E} = \underset{\boldsymbol{\beta}, \sigma}{\arg\min} \left[\frac{1}{2\sigma\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^{n} \phi(y_i | \beta_0 + \beta_1 x_i, \sigma) \right]. \tag{8}$$

In the following we introduce two simulated examples in order to demonstrate the behaviour of the L_2 criterion in the presence of outliers and in the presence of clustered data. To evaluate its performance, we shall use the Maximum Likelihood estimator and the robust Huber M estimator. Given $T_1 = L_2 E$, we shall perform the M.C.S. test two times: the first one, fixing $T_0 = MLE$, for $sim(MLE, L_2 E)$, the second one fixing $T_0 = HME$, for $sim(HME, L_2 E)$. We remark that, as $p = 1$, in both situations we have $\boldsymbol{I}^p = [\min(x_i); \max(x_i)]$ and that clearly the integrals of Eq. (6) are defined on bi-dimensional domains.

Example I. Let us consider a simulated dataset of $n = 200$ points generated according to the model $Y = X + \varepsilon$, where $X \sim \mathcal{U}(0, 10)$ and $\varepsilon \sim \mathcal{N}(0, 0.8)$. We then introduce $m = 10(30)$ points according to the model $Y = -3 + X + \varepsilon$,

Table 1 Results of simulated Example I

	$m = 10$			$m = 30$		
	MLE	*HME*	L_2E	*MLE*	*HME*	L_2E
$\hat{\beta}_0$	0.3078	0.1616	0.0353	0.2884	0.2081	0.0139
$\hat{\beta}_1$	0.9054	0.9509	0.9886	0.8635	0.8944	0.9975
$\hat{\sigma}$	0.9889	0.9972	0.7926	1.2352	1.2389	0.9712

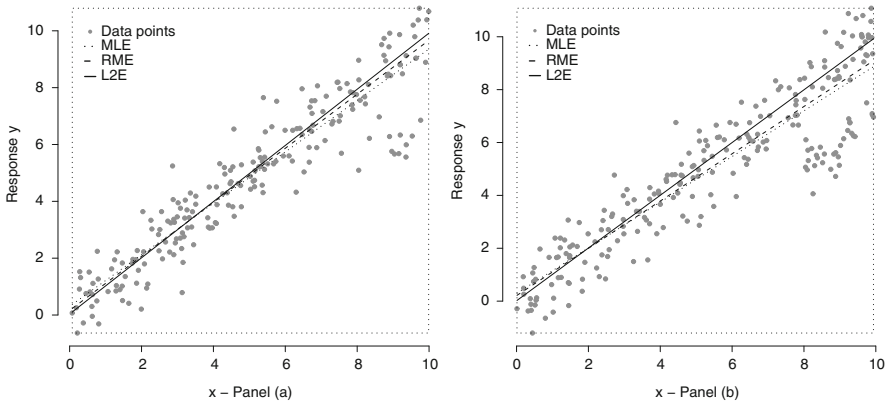

Fig. 2 Data points of Example I and estimated models $\hat{m}_{ML}(x)$, $\hat{m}_{HM}(x)$ and $\hat{m}_{L2}(x)$. In panel (**a**) we set $m = 10$ outliers while in panel (**b**) $m = 30$

where $X \sim \mathcal{U}(8, 10)$ and $\varepsilon \sim \mathcal{N}(0, 0.4)$, so that *they can be considered as outliers*. Resorting to the estimators *ML*, *HM* and L_2 we obtain the following estimates of the parameters β_0, β_1 and σ listed in Table 1 (also see Fig. 2).

Applying the M.C.S. test, with $\alpha = 0.01$, to the estimated models $\hat{m}_{ML}(x)$ and $\hat{m}_{L2}(x)$, we reject the null hypothesis of system (7) as we have $sim_{ML,L2} = 0.0203 > \max(sim^*_{ML,L2}) = 0.0128$. Turning our attention to models $\hat{m}_{HM}(x)$ and $\hat{m}_{L2}(x)$, the M.C.S. test leads us to accept the null hypothesis since $sim_{HM,L2} = 0.0091 < \max(sim^*_{HM,L2}) = 0.0123$.

In the case we add $m = 30$ outliers to the sample data, the results of the M.C.S. tests lead us to different conclusions. In both situations we reject the null hypothesis of system (7) as we have

$$sim_{ML,L2} = 0.0364 > \max(sim^*_{ML,L2}) = 0.0159$$

$$sim_{HM,L2} = 0.0289 > \max(sim^*_{HM,L2}) = 0.0103$$

When the outliers are few, the estimated regression model $\hat{m}_{HM}(x)$ and $\hat{m}_{L2}(x)$ do not differ significantly. This is not the case when the number of outliers increases; in this sense it seems that L_2 estimator can be helpful in cluster detection.

Table 2 Results of simulated Example II

	MLE	*HME*	L_2E
$\hat{\beta}_0$	2.6755	2.4956	1.7340
$\hat{\beta}_1$	0.4607	0.5086	0.6856
$\hat{\sigma}$	1.4021	1.4074	1.1633

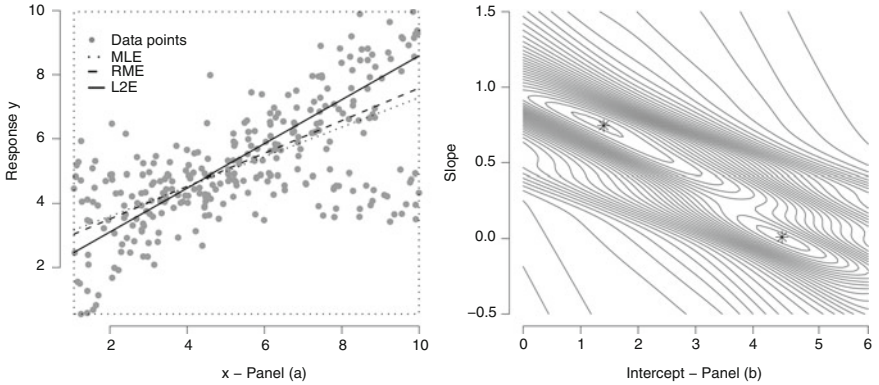

Fig. 3 (Panel **a**) Data points of Example II and estimated models $\hat{m}_{ML}(x)$, $\hat{m}_{HM}(x)$ and $\hat{m}_{L2}(x)$. (Panel **b**) Contour plot of function $g(\boldsymbol{\beta}|\sigma^*)$ of Eq. (9) evaluated at $\sigma^* = 0.5\,\hat{\sigma}_{L_2E}$

Example II. Let us consider a dataset of $n = 300$ points, 200 of which arise from model $Y = 1 + 0.8\,X + \varepsilon_1$ while the remaining from model $Y = 5 - 0.2\,X + \varepsilon_2$, where $\varepsilon_1 \sim \mathcal{N}(0, 1)$, $\varepsilon_2 \sim \mathcal{N}(0, 0.5)$ and $X \sim \mathcal{U}(1, 10)$. Again, resorting to the estimators *ML*, *HM* and L_2 we obtain the following estimates of the parameters β_0, β_1 and σ listed in Table 2 (also see Fig. 3, panel a). Considering the models $\hat{m}_{ML}(x)$ and $\hat{m}_{L2}(x)$ the M.C.S. test, with $\alpha = 0.01$, indicates that they can be considered dissimilar, as we observe $sim_{ML,L2} = 0.0582 > \max(sim^*_{ML,L2}) = 0.0210$. This is still true if we consider the estimated models $\hat{m}_{HM}(x)$ and $\hat{m}_{L2}(x)$, in fact from the M.C.S. test we have $sim_{HM,L2} = 0.0451 > \max(sim^*_{MH,L2}) = 0.0156$. Also in this situation the L_2 estimator seems to be helpful in detecting clusters of data when compared with the Maximum Likelihood and the Huber *M* estimators.

5 Mixture of Regression Models via L_2

It seems to the authors that the properties of robustness of L_2 estimates, as outlined above, can be helpful in pointing out the presence of clusters in the data, e.g. Durio and Isaia (2007).

This in the sense that whenever sample data belong to two (or more) clusters, $\hat{m}_{L2}(\mathbf{x})$ will always tend to fit the cluster with the heaviest number of data points and hence big discrepancies between $\hat{m}_{ML}(\mathbf{x})$ and $\hat{m}_{L2}(\mathbf{x})$ will be likely to be observed, as illustrated by the previous examples. Investigating more accurately function (5)

for a fixed value of σ it can be seen that in all situations where sample data are clustered it can show more than one local minimum. A simple way forward is to investigate the behaviour of the function

$$g(\boldsymbol{\beta}|\sigma^*) = \frac{1}{2\sigma\sqrt{\pi}} - \frac{2}{n}\sum_{i=1}^{n}\phi(y_i|m_{\boldsymbol{\beta}}(\mathbf{x}_i), \sigma^*) \tag{9}$$

for different values of σ^* on its parameter space, for instance, the interval $]0, 2\cdot\hat{\sigma}_{L_2E}]$. In fact, whenever sample data are clustered, function $g(\boldsymbol{\beta}|\sigma^*)$ given by Eq. (9) shows one absolute and one or more local points of minimum.

Whenever the presence of clusters of data is detected by L_2 criterion, we can use L_2 estimator assuming that the model that best fits the data is a mixture of $K \geq 2$ regression models. Assuming that each data point $(\mathbf{x}_i, y_i)$ comes from the k-th regression model $y_i = m_{\boldsymbol{\beta}_k}(\mathbf{x}_i) + \varepsilon_{ik}$ with probability p_k, we suppose that the random variables $Y|\mathbf{x}$ are distributed as a mixture of K Gaussian random variables, i.e.,

$$f_{Y|\mathbf{x}}(y|\boldsymbol{\theta}_0) = \sum_{k=1}^{K} p_k^0\,\phi(y|m_{\boldsymbol{\beta}_k^0}(\mathbf{x}), \sigma_k^0). \tag{10}$$

We are now able to derive the following closed-form expression for the estimates of $\boldsymbol{\theta}_0 = [\boldsymbol{p}^0, \boldsymbol{\beta}^0, \boldsymbol{\sigma}^0]$; in fact, according to Eq. (9) and recalling Eq. (4), we have

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{L_2E} = &\underset{\boldsymbol{p},\boldsymbol{\beta},\boldsymbol{\sigma}}{\arg\min}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\sum_{h=1}^{K} p_j\; p_h\,\phi(0|m_{\boldsymbol{\beta}_j}(\mathbf{x}_i) - m_{\boldsymbol{\beta}_h}(\mathbf{x}_i), \sigma_j^2 + \sigma_h^2) - \right.\\ &\left. - \frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} p_k\,\phi(y_i|m_{\boldsymbol{\beta}_k}(\mathbf{x}_i), \sigma_k^2)\right].\end{aligned} \tag{11}$$

Solving Eq. (11) we obtain the estimates of the vector of the weights, i.e. $\hat{\boldsymbol{p}} = [p_1, \ldots, p_K]^T$, the vector of the parameters, i.e. $\hat{\boldsymbol{\beta}} = [\beta_{0_1}, \ldots, \beta_{d_1}, \ldots, \beta_{0_K}, \ldots, \beta_{d_K}]^T$ and the vector of the standard deviations of the error of each component of the mixture, i.e. $\hat{\boldsymbol{\sigma}} = [\sigma_1, \ldots, \sigma_K]^T$.

Example II (continued). Referring to the situation of Example II, for which $\hat{\sigma}_{L_2} = 1.1633$, the contour plot of function $g(\boldsymbol{\beta}|\sigma^*)$ of Eq. (9) and displayed in Fig. 3, panel b, evaluated at $\sigma^* = 0.5\,\hat{\sigma}_{L_2E}$, shows the existence of one absolute minimum corresponding to the estimates of the parameters of the model $Y = 1 + 0.8\,X + \varepsilon_1$ and one local minimum close to the values of the parameters of the model $Y = 5 - 0.2\,X + \varepsilon_2$. We therefore consider a mixture of $K = 2$ simple linear regression models. Since in this situation Eq. (10) becomes

$$f_{Y|\mathbf{x}}(y|\boldsymbol{\theta}_0) = p_1^0\,\phi(y|\beta_{0_1}^0 + \beta_{1_1}^0\,x, \sigma_1^0) + p_2^0\,\phi(y|\beta_{0_2}^0 + \beta_{1_2}^0\,x, \sigma_2^0),$$

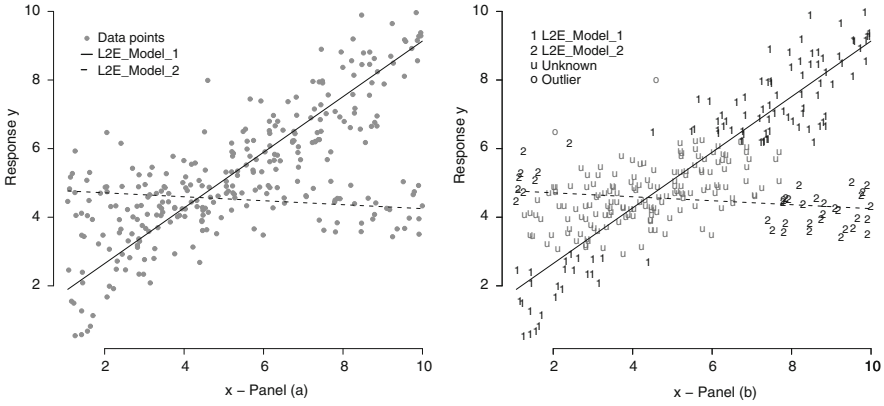

Fig. 4 (Panel **a**) Data points and estimated components of the mixture of two simple regression models via L_2. (Panel **b**) Data points assignment according to the "quick classification rule" with $\gamma = 3$

the L_2 estimates of the vector $\boldsymbol{\theta}_0$, according to Eq. (11), will be given by solving

$$\hat{\boldsymbol{\theta}}_{L_2E} = \underset{\boldsymbol{p},\boldsymbol{\beta},\boldsymbol{\sigma}}{\arg\min} \left[\frac{p_1^2\sigma_2 + p_2^2\sigma_1}{2\sigma_1\sigma_2\sqrt{\pi}} + \frac{2}{n}\sum_{i=1}^{n} p_1 p_2 \phi(0|\beta_{0_1} + \beta_{1_1}x_i - \beta_{0_2} - \beta_{1_2}x_i, \sigma_1^2 + \sigma_2^2) - \right.$$

$$\left. - \frac{2}{n}\sum_{i=1}^{n}\Big(p_1\,\phi(y_i|\beta_{0_1} + \beta_{1_1}\,x_i, \sigma_1^2) + p_2\,\phi(y_i|\beta_{0_2} + \beta_{1_2}\,x_i, \sigma_2^2)\Big)\right]. \tag{12}$$

From numerical minimization of Eq. (12), we obtain (see Fig. 4, panel a) the following estimates of the eight parameters of the mixture

$$\begin{array}{lllll} L_2E \text{ Model\_1:} & \hat{p}_1 = 0.646 & \hat{\beta}_{0_1} = 1.0281 & \hat{\beta}_{1_1} = 0.8109 & \hat{\sigma}_1 = 0.8411 \\ L_2E \text{ Model\_2:} & \hat{p}_2 = 0.354 & \hat{\beta}_{0_2} = 4.8267 & \hat{\beta}_{1_2} = -0.0576 & \hat{\sigma}_2 = 0.5854 \end{array}$$

which are quite close to the true values of the parameters.

From a practical point of view, it would be interesting to be able to highlight which data points belong to each component of the mixture; to this end we resort to a *quick classification rule* based on the assumption that the density of the errors follows a Normal distribution, i.e. $\forall\, i = 1, \ldots, n$

$$\begin{array}{ll} \text{if } |\hat{\varepsilon}_{i_1}| \leqq \gamma\,\hat{\sigma}_1 \wedge |\hat{\varepsilon}_{i_2}| > \gamma\,\hat{\sigma}_2 & \rightarrow (x_i, y_i) \in \text{ Model } L_2E - I \\ \text{if } |\hat{\varepsilon}_{i_1}| > \gamma\,\hat{\sigma}_1 \wedge |\hat{\varepsilon}_{i_2}| \leqq \gamma\,\hat{\sigma}_2 & \rightarrow (x_i, y_i) \in \text{ Model } L_2E - II \\ \text{if } |\hat{\varepsilon}_{i_1}| \leqq \gamma\,\hat{\sigma}_1 \wedge |\hat{\varepsilon}_{i_2}| \leqq \gamma\,\hat{\sigma}_2 & \rightarrow (x_i, y_i) \in \text{ Unknown model} \\ \text{if } |\hat{\varepsilon}_{i_1}| > \gamma\,\hat{\sigma}_1 \wedge |\hat{\varepsilon}_{i_2}| > \gamma\,\hat{\sigma}_2 & \rightarrow (x_i, y_i) \in \text{ Outlier,} \end{array} \tag{13}$$

where γ is an appropriate quantile of a $\mathcal{N}(0, 1)$.

Table 3 Classification I

	L_2 estimates of $\hat{\boldsymbol{p}}$	*Quick rule*	
L_2E Model_1	64.6 %	34.6 %	(103)
L_2E Model_2	35.4 %	12.7 %	(38)
Unknown model	–	52.7 %	(157)

Fixing $\gamma = 3$, if we apply the *quick rule* and drop two points that are classified as outliers we obtain (see Fig. 4, panel b) the following classification table, see Table 3. Clearly, the high percentage of not assigned points (52.7%) is due to the specific structure of the two clusters which are quite confused.

6 The Case Study

A firm operating in the field of diagnosis and decontamination of *electronic transformers fluids* assesses the risks of fluid degradation, electric shocks, fire or explosion, PCB contamination, decomposition of cellulosic insulation, etc. With the aid of well-known models and relying on the results of chemical analysis, the firm's staff estimate the value of the risk on continuous scales.

In order to determine if their methods of assigning risk values are independent of specific characteristics of the transformers (age, voltage, fluid mass, etc.) we conducted an analysis based on a database of 1,215 records of diagnosis containing oil chemical analysis, technical characteristics and risk values.

Taking into account the *risk of fire (Y)* and the *risk of electric shocks (X)*, it was natural to suppose a linear dependence between the two variables, i.e., we considered the simple regression model with $m_{\boldsymbol{\beta}}(x_i) = \beta_0 + \beta_1 x_i$.

Resorting to the estimators *ML*, *HM* and L_2 we obtained the following estimates of the parameters β_0, β_1 and σ listed in Table 4.

Although the estimates of the vector of the parameters $\boldsymbol{\beta}$ are quite close, the corresponding three estimated models differ in some way, e.g., Fig. 5, panel a.

Computing the values of the *sim*() statistics, the M.C.S. test led us to the conclusion that the L_2 estimated model can be considered dissimilar from both $\hat{m}_{ML}(x)$ and $\hat{m}_{HM}(x)$ models, as

$$sim_{ML,L2} = 0.0220 > \max(sim^*_{ML,L2}) = 0.0051$$

$$sim_{HM,L2} = 0.0203 > \max(sim^*_{HM,L2}) = 0.0031$$

Probing more deeply, we found that function $g(\boldsymbol{\beta}|\sigma^*)$ of Eq. (9) presents two points of minimum for $\sigma^* = 0.5\,\hat{\sigma}_{L_2E} = 0.0755$, as shown in Fig. 5, panel b.

Therefore we decided to model our data by means of a mixture of two simple regression models. Considering the L_2 criterion and solving Eq. (12), we found that about 57 %$(= \hat{p}_1\,\%)$ of the data points follow the model

$$\hat{m}_{\boldsymbol{\beta}_1}(x) = -0.4042 + 1.7705\,x \qquad \rightarrow \qquad L_2E \text{ Model\_1}$$

Table 4 Estimates of the parameters after resorting

	MLE	*HME*	L_2E
$\hat{\beta}_0$	−0.4321	−0.4423	−0.5330
$\hat{\beta}_1$	1.7110	1.7199	1.8115
$\hat{\sigma}$	0.1472	0.1471	0.1509

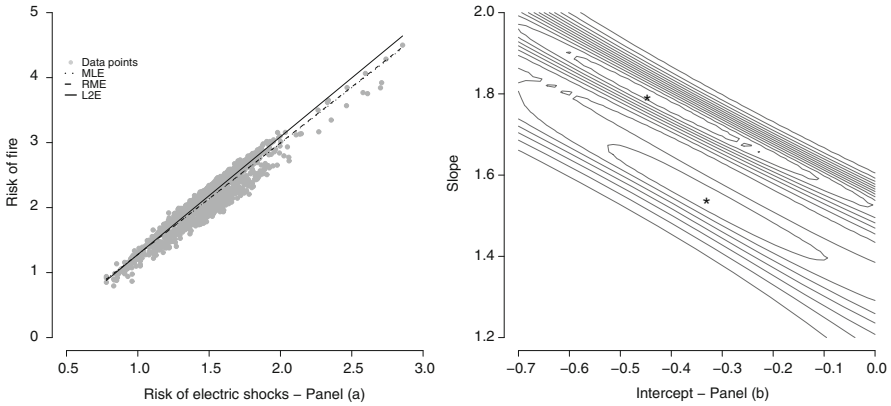

Fig. 5 Case study. (Panel **a**) Data points and estimated models $\hat{m}_{ML}(x)$, $\hat{m}_{HM}(x)$ and $\hat{m}_{L2}(x)$. (Panel **b**) Contour plot of function $g(\boldsymbol{\beta}|\sigma^*)$ of Eq. (9) evaluated at $\sigma^* = 0.5\,\hat{\sigma}_{L_2E}$, with $\hat{\sigma}_{L_2E} = 0.151$

for which $\hat{\sigma}_1 = 0.0547$, while the remaining 43 %(= $\hat{p}_2$ %) of the data points follow the model

$$\hat{m}_{\boldsymbol{\beta}_2}(x) = -0.3955 + 1.5847\,x \qquad \rightarrow \qquad L_2E \text{ Model\_2}$$

for which $\hat{\sigma}_2 = 0.0775$. Panel a of Fig. 6 shows the two estimates models.

Applying the *quick rule* we were able to classify the data according to whether they followed the first or the second regression model. From the L_2 estimates of $\hat{\boldsymbol{p}}$ and the *quick rule* (dropping two points that were classified as outliers) we obtained the following classification table, see Table 5.

In order to classify the 266 (= 22.0 %) points belonging, according to the *quick rule*, to the *Unknown Model*, we had to investigate more deeply the specific characteristics of the transformers themselves.

Examining our database, we found that 40 % of the transformers has a fluid mass $\leqq$500 kg and the L_2 criterion gave us an estimate of 43 % for the weight of points belonging to L_2E Model_1 while our *quick rule* assigned the 36.9 % of data points to L_2E Model_2.

Furthermore, our *quick classification rule* assigns 419 out of the 448 points (93.5 %) to L_2E Model_2 and these have a fluid mass less (or equal) than 500 kg, while all the 499 transformers imputed to L_2E Model_1 have a fluid mass greater than 500 kg, see Table 6.

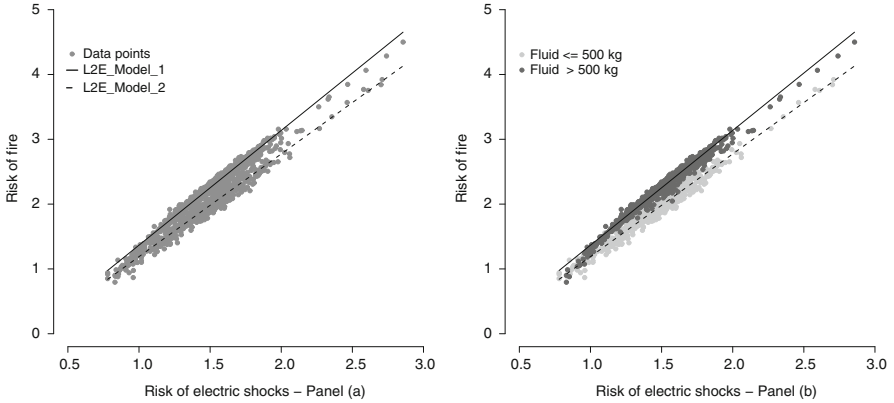

Fig. 6 Case study. (Panel **a**) Data points and estimated models $\hat{m}_{\hat{\boldsymbol{\beta}}_1}(x)$ and $\hat{m}_{\hat{\boldsymbol{\beta}}_2}(x)$. (Panel **b**) Final data points assignment according to the fluid mass of the electrical transformers

Table 5 Classification II

	L_2 estimates of $\hat{\boldsymbol{p}}$	*Quick rule*
L_2E Model_1	57.0 %	41.1 % (499)
L_2E Model_2	43.0 %	36.9 % (448)
Unknown model	–	22.0 % (266)

Table 6 Fluid mass of the model

	Fluid mass $\leqq$ 500 kg	Fluid mass $>$ 500 kg
L_2E Model_1	0 (0.0 %)	499 (100 %)
L_2E Model_2	419 (93.5 %)	29 (6.5 %)
Unknown model	65 (24.4 %)	201 (75.6 %)

From the above, we decided to use the *fluid mass as clustering variable* and so we assigned the transformers with a fluid mass equal or less than 500 kg to Model L_2E Model_2 while the transformers with a fluid mass greater than 500 kg were assigned to the L_2E Model_1 regression line. The final assignment is shown in Fig. 6, panel b.

These results allowed us to state that, at fixed level of risk of electric shocks, the risk of fire was evaluated in a different way for the two groups of transformers, i.e., the relationship between the two variables depended on the fluid mass of the transformers.

However, the chemical staff of the firm could not find any scientific reason to explain the different risks of fire in the two types of transformers, so they decided to change the model used by assigning different weights to the hydrocarbon variable in order to better reflect the differential risks of fire.

Acknowledgements The authors are indebted to the coordinating editors and to the anonymous referees for carefully reading the manuscript and for their many important remarks and suggestions.

References

Barnard, G. A. (1963). Contribution to the discussion of paper by m.s. bartlett. *Journal of the Royal Statistical Society, B, 25*, 294.

Basu, A., Harris, I. R., Hjort, N., & Jones, M. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika, 85*, 549–559.

Davies, P. L. (1993). Aspects of robust linear regression. *Annals of Statistics, 21*, 1843–1899.

Dodge, Y., & Jurečkova, J. (2000). *Adaptive regression*. New York: Springer.

Durio, A., & Isaia, E. D. (2007). A quick procedure for model selection in the case of mixture of normal densities. *Journal of Computational Statistics & Data Analysis, 51*(12), 5635–5643.

Durio, A., & Isaia, E. D. (2010). Clusters detection in regression problems: A similarity test between estimate. *Communications in Statistics Theory and Methods, 39*, 508–516.

Fujisawa, H., & Eguchi, F. (2006). Robust estimation in the normal mixture model. *Journal of Statistical Planning Inference, 136*, 3989–4011.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, R.,J., & Stahel, W. A. (2005). *Robust regression and outlier detection*. New York: Wiley.

Hope, A. C. (1968). A simplified monte carlo significance test procedure. *Journal of the Royal Statistical Society, B, 30*, 582–598.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Jurečkova, J., & Picek, J. (2006). *Robust statistical methods with R*. Boca Raton: Chapman & Hall.

Maronna, R., Martin, D., & Yohai, V. (2006). *Robust statistics: Theory and methods*. New York: Wiley.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P. J., Van Alest, S., Van Driessen, K., & Agulló, J. (2004). Robust multivariate regression. *Technometric, 46*, 293–305.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics, 43*, 274–285.

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). New York: Wiley.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.