Developments in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 19, Ljubljana: FDV, 2003

Parametric Regression Models by Minimum L_2 Criterion. A Study on the Risks of Fire and Electric Shocks of Electronic Transformers

Alessandra Durio[1] and Ennio Davide Isaia[2]

Abstract

The purpose of our work is to investigate on the use of L_2 distance as a theoretical and practical estimation tool for parametric regression models. This approach is particularly helpful in all those situations involving the study of large data sets, handling large samples with a consistent numbers of outliers, situations in which maximum likelihood regression models are unstable. We shall also see how L_2E criterion may be applied in fitting mixture regression models and how it allows to detect clusters of data. Theory is outlined, some examples on simulated data sets are given and an application to data from investigation on risks of fire and electric shocks of electronic transformers is proposed to illustrate the use of the approach. In order to estimate the parameters of the models we implemented some routines in R computing environment.

1 Introduction

In applied statistics regression is one of the most used tool in establishing the relationship between a response and an explanatory variable. In the following we shall investigate on the use of Integrated Squared Error (ISE), or L_2 distance, as a theoretical and practical estimation tool for parametric regression models. The approach based on minimizing the Integrated Squared Error, or L_2 minimizing estimate criterion (briefly L_2E), is particularly helpful (Scott, 2001a) in all those situations where, due to large sample size, a careful data preparation is not feasible and hence data may be heavily contaminated by substantial numbers of outliers or extreme values (Reiss, 1997). L_2E criterion is better suited to treating models as approximation and it may be viewed as a practical diagnostic tool in building useful models (Scott, 2001b).

[1] Department of Statistics & Mathematics, University of Turin, durio@econ.unito.it
[2] Department of Statistics & Mathematics, University of Turin, isaia@econ.unito.it

We shall point out how L_2E may be used in fitting mixtures of regression models and how it allows to simultaneously estimate the parameters of each regression model, the variances of the errors and the probability that data point belongs to each regression model. For practical purposes, the authors propose a quick rule for assign data points to each regression model, based on the estimated variance of errors.

The theory is outlined, some numerical examples are given and an application to data from investigations on risks of fire and electric shocks of electronic transformers is presented to illustrate the use of the approach.

2 Genesis of L_2 minimizing estimate

The L_2E criterion originates in the derivation of the nonparametric least squares Cross-validation algorithm for choosing the bandwidth h for the kernel estimate of a density.

In fact, it is well known that in the *nonparametric case*, given the r.v. X with unknown density $f(x)$, the latter is estimated on the basis of a random sample $X_1, \ldots, X_n$, by the kernel

$$\hat{f}_h(x) = (n\,h)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where the optimal bandwidth h is the one minimizing the ISE, i.e.

$$\begin{aligned} h &= \arg\min_h \int_{I\!R} \left[\hat{f}_h(x) - f(x)\right]^2 dx = \\ &= \arg\min_h \left[\int_{I\!R} \hat{f}_h(x)^2\, dx - 2 \int_{I\!R} \hat{f}_h(x)\, f(x)\, dx + \int_{I\!R} f(x)^2\, dx\right] \end{aligned} \tag{2.1}$$

Observing that in (2.1) $\int_{I\!R} f(x)^2$ does not depend on h and using an appropriate estimator, say $\widehat{I\!E}\left[\hat{f}_h(x)\right]$, for $I\!E\left[\hat{f}_h(x)\right] = \int_{I\!R} \hat{f}_h(x)\, f(x)\, dx$, we have the estimate

$$\hat{h} = \arg\min_h \left[\int_{I\!R} \hat{f}_h(x)^2\, dx - 2\, \widehat{I\!E}\left[\hat{f}_h(x)\right]\right] \tag{2.2}$$

An unbiased estimate for h may be obtained resorting to the Cross-validation method (Rudemo, 1982; Wand and Jones, 1995).

In the *parametric case*, given the r.v. X, with unknown density $f(x|\theta_0)$, depending on the unknown parameter θ_0, for which we introduce the model $f(x|\theta)$, we may rewrite equation (2.1) with θ replacing h, i.e.

$$\begin{aligned}\theta &= \arg\min_{\theta} \int_{I\!R} [f(x|\theta) - f(x|\theta_0)]^2 \, dx = \\ &= \arg\min_{\theta} \left[\int_{I\!R} f(x|\theta)^2 \, dx - 2 \int_{I\!R} f(x|\theta) \, f(x|\theta_0) \, dx \right] = \\ &= \arg\min_{\theta} \left[\int_{I\!R} f(x|\theta)^2 \, dx - 2 \, I\!E \left[f(x|\theta_0) \right] \right]\end{aligned} \tag{2.3}$$

If we replace the so called *expected height of the density* $I\!E\left[f(x|\theta_0)\right]$ with the estimator $\widehat{I\!E}\left[f(x|\theta_0)\right] = n^{-1}\sum_{i=1}^{n} f(x_i|\theta)$, the proposed estimator for θ_0 minimizing the L_2 distance will be

$$\hat{\theta}_{L_2E} = \arg\min_{\theta} \left[\int_{I\!R} f(x|\theta)^2 \, dx - \frac{2}{n} \sum_{i=1}^{n} f(x_i|\theta) \right] \tag{2.4}$$

The following examples may illustrate the situation.

Example I: if we suppose that $X \sim \mathcal{U}(0, b)$, then equation (2.4) becomes

$$\hat{b}_{L_2E} = \arg\min_{b} \left[\frac{1}{b} - \frac{2}{n\,b} \sum_{i=1}^{n} I(x_i \leqq b) \right]$$

where I is the indicator function.

Example II: if we suppose that $X \sim \mathcal{N}(\mu, 1)$, then, letting $\theta = \mu$, equation (2.4) becomes

$$\hat{\mu}_{L_2E} = \arg\min_{\mu} \left[\frac{1}{2\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^{n} \phi(x_i|\mu, 1) \right]$$

where ϕ denotes the normal density.

Example III: if we suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$, letting $\boldsymbol{\theta} = [\mu, \sigma^2]^t$ be the vector of the unknown parameters, equation (2.4) becomes

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{L_2E} &= \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{2\,\sigma\,\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^{n} \phi(x_i|\boldsymbol{\theta}) \right] = \\ &= \arg\min_{\mu,\sigma} \left[\frac{1}{2\,\sigma\,\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^{n} \phi(x_i|\mu, \sigma^2) \right]\end{aligned}$$

3 Parametric linear regression according to the L_2E criterion

Let us consider the observed data set $\{(x_i, y_i)\}_{i=1,\ldots,n}$, where each (x_i, y_i) pair stems from a bivariate random sample drawn from the bivariate r.v. (X, Y). The regression model for the observed data set we study is

$$Y_i = m_{\boldsymbol{\beta}}(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n \tag{3.1}$$

where the regression mean $m_{\boldsymbol{\beta}}(x) = I\!E[Y|x]$ is the object of our interest, $\boldsymbol{\beta}$ is the vector of the parameters to be estimated and the error random variables $\{\varepsilon_i\}_{i=1,\ldots,n}$ are assumed to be *independent with zero mean and unknown variances*.

Now, it is well known that if the errors ε_i are *independent and identically distributed* (i.i.d.) as the r.v. ε with normal density $\phi(\varepsilon|0, \sigma_\varepsilon)$, the Maximum Likelihood Estimate (MLE) is given by

$$(\hat{\boldsymbol{\beta}})_{MLE} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} [y_i - m_{\boldsymbol{\beta}}(x_i)]^2$$

and it is important to recall that the estimate of σ_ε^2 implies the knowledge of $\boldsymbol{\beta}$ (*and not vice versa*) and that the check of the assumptions on the density of the errors must be conducted by analyzing (through formal and/or informal tools) the residuals from the regression itself.

We now turn to illustrate how the estimate based on the L_2 criterion may be applied to parametric regression models, observing that it plugs in directly the parametric model of the residuals' density, not necessarily normal, and it simultaneously provides an estimate for the parameters of the regression model as well as for those of the density of the errors.

If we now suppose that the random errors ε_i in (3.1) are *independent and identically distributed* as the r.v. ε with density $f(\varepsilon|0, \sigma_0)$ and we observe that, $\forall i = 1, \ldots, n$, $\varepsilon_i = Y_i - m_{\boldsymbol{\beta}}(x_i)$, the parameters in $\boldsymbol{\beta}$ and σ_0 may be estimated simultaneously by L_2 criterion resorting to equation (2.4), which will assume the form

$$\begin{aligned}(\hat{\boldsymbol{\beta}}, \hat{\sigma})_{L_2E} =& \arg\min_{\boldsymbol{\beta},\sigma} \left[\int f(\varepsilon|0,\sigma)^2 \, d\varepsilon - \frac{2}{n} \sum_{i=1}^{n} f(\varepsilon_i|0,\sigma) \right] = \\ =& \arg\min_{\boldsymbol{\beta},\sigma} \left[\int f(\varepsilon|0,\sigma)^2 \, d\varepsilon - \frac{2}{n} \sum_{i=1}^{n} f(y_i - m_{\boldsymbol{\beta}}(x_i)|0,\sigma) \right]\end{aligned} \tag{3.2}$$

Referring to regression model (3.1), if we assume that the r.vs. ε_i are i.i.d. as the r.v. $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, then, since $\int_{I\!R} \phi(\varepsilon|0, \sigma^2)^2 \, d\varepsilon = (2\,\sigma\,\sqrt{\pi})^{-1}$, it is easy to check that equation (3.2) reduces to

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma})_{L_2E} = \arg\min_{\boldsymbol{\beta},\sigma} \left[\frac{1}{2\sigma\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^{n} \phi(y_i - m_{\boldsymbol{\beta}}(x_i)|0, \sigma^2) \right] \tag{3.3}$$

If we consider the simple linear regression model $Y_i = \beta_0 + \beta_1 \, x_i + \varepsilon_i$, then equation (3.3) becomes

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma})_{L_2E} = \arg\min_{\boldsymbol{\beta},\sigma} \left[\frac{1}{2\sigma\sqrt{\pi}} - \frac{2}{n} \sum_{i=1}^{n} \phi(y_i - \beta_0 - \beta_1 \, x_i|0, \sigma^2) \right] \tag{3.4}$$

Since closed forms for equations (3.2), (3.3) and (3.4) are unlikely to be found, we have to resort to numerical minimization algorithms (for example the `optim` or `nlm` routines implemented in R environment). Furthermore, in some situations, e.g. in presence of outliers or contaminated data, there may exist more than one local minimum. In these cases we shall construct two or more regression models.

3.1 Numerical example I

The ideas introduced above may be explained on a simulated dataset of $n = 200$ points such that they belong to two clusters, more precisely

Cluster 1	$y = x + \varepsilon$	$n_1 = 125$	$x \sim \mathcal{U}(1, 10)$	$\varepsilon \sim \mathcal{N}(0, 1)$
Cluster 2	$y = 1 + 0.2\,x + \varepsilon$	$n_2 = 75$	$x \sim \mathcal{U}(3, 9)$	$\varepsilon \sim \mathcal{N}(0, 1)$

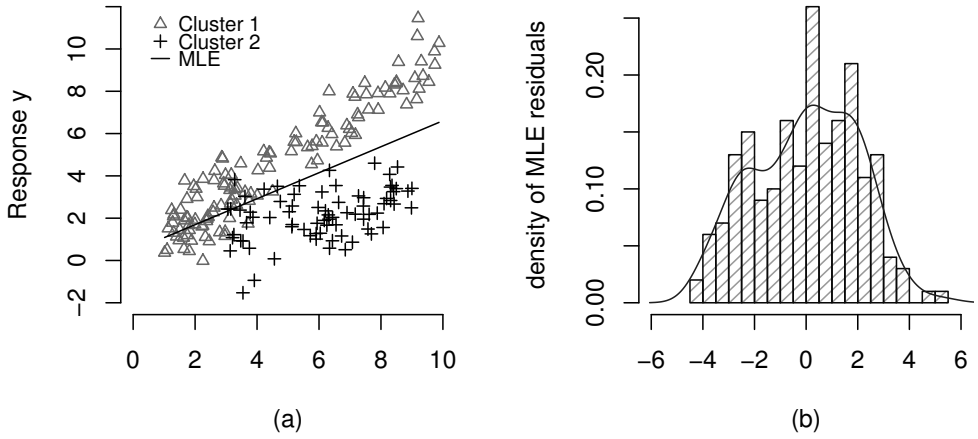

Figure 1: Panel (a): data points and MLE regression line. Panel (b): histogram of residuals from MLE regression line and kernel density estimate.

Figure 1, panel (a), shows the generated data points and the MLE regression line for which $\beta_{0_{MLE}} = 0.463$, $\beta_{1_{MLE}} = 0.614$ and $R^2 = 0.363$, with an associated p-value of the F statistics $< 2.2e - 16$.

Figure 1 , panel (b), displays the histogram and the kernel density estimate of the residuals from the MLE regression line. It clearly suggests that the errors are not normally distributed and so the simple linear regression model seems to be inadeguate. Furthermore, the kernel estimate of the density of residuals shows that we may be in presence of clustered data (the kernel estimated density seems to be *bimodal*).

It should be observed that, given the ML estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we have $\hat{\sigma}_{\varepsilon_{MLE}} = 2.024$, which is quite far away from the true value $\sigma_\varepsilon = 1$

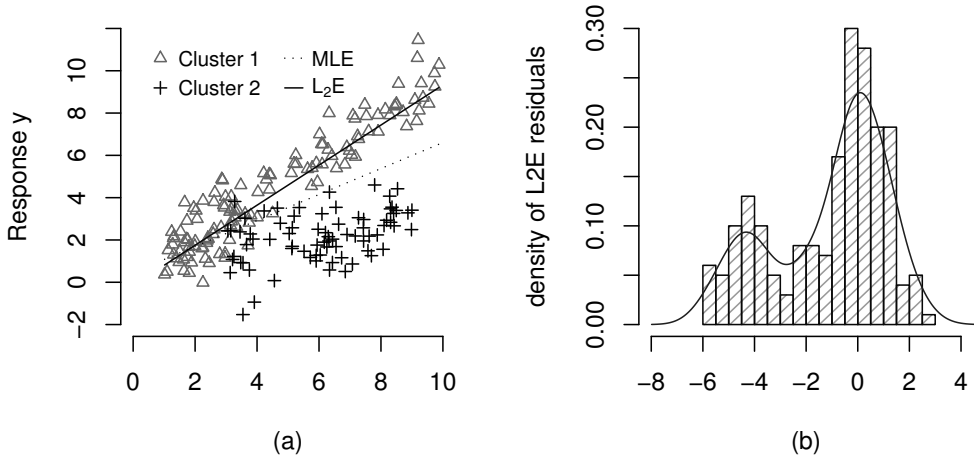

Figure 2: Panel (a): data points, MLE and L_2E regression lines. Panel (b): histogram of residuals from L_2E regression line and kernel density estimate.

Turning now to the L_2E criterion, according to (3.4), we obtain the estimates

$$\hat{\sigma}_{\varepsilon_{L_2E}} = 1.511 \qquad \hat{\beta}_{0_{L_2E}} = -0.146 \qquad \hat{\beta}_{1_{L_2E}} = 0.947$$

Figure 2, panel (a), shows the estimate of the simple linear regression model according to L_2 criterion (note that estimate of σ_ε is still inflated). Figure 2, panel (b), shows the histogram and the kernel estimate of the density of residuals from L_2E regression line. It is evident that the distribution of residuals is bimodal and this suggests that data points belong to two clusters, while L_2E gives an accurate estimate for the simple linear model applied to cluster 1.

Digging a little deeper, it is interesting to have a look at the function

$$g(\beta_0, \beta_1 | \hat{\sigma}_{\varepsilon_{L_2E}}) = \frac{1}{2\sigma_{\varepsilon_{L_2E}}\sqrt{\pi}} - \frac{2}{n}\sum_{i=1}^{n} \phi(y_i - \beta_0 - \beta_1 x_i | 0, \hat{\sigma}^2_{\varepsilon_{L_2E}}) \tag{3.5}$$

which corresponds to the function to be minimized in (3.4) with σ fixed at $\hat{\sigma}_{\varepsilon_{L_2E}}$.

Its "level plot", displayed in Figure 3, panel (a), suggests that there are two local minima and hence we may construct two regression lines, one for each cluster. From numerical minimization of equation (3.4) at $\sigma = \hat{\sigma}_{\varepsilon_{L_2E}}$, we obtain the two L_2E regression lines

$$\hat{y}_{L_2E-1} = -0.146 + 0.947\, x \qquad \hat{y}_{L_2E-2} = 0.947 + 0.147\, x$$

which are displayed in Figure 3, panel (b).

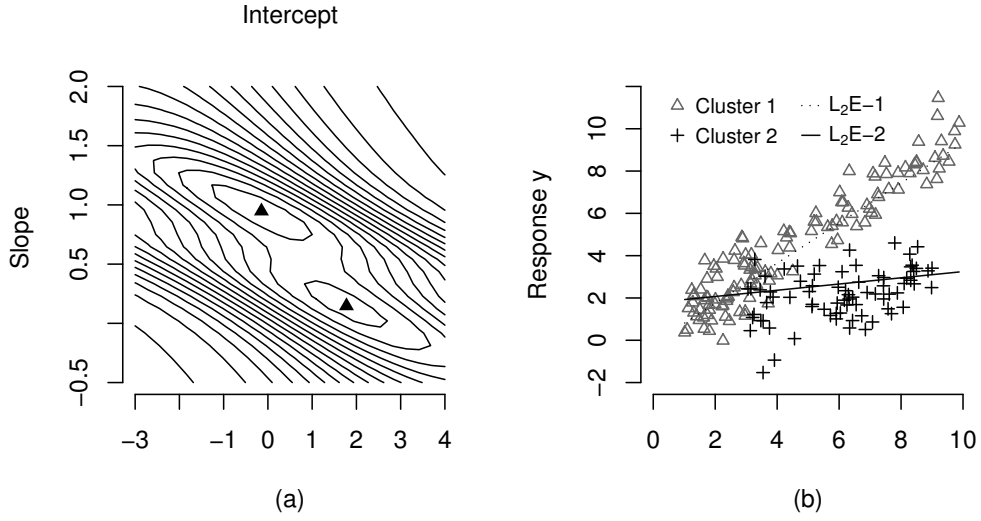

Figure 3: Panel (a): "level plot" of function (3.5) at $\sigma = \hat{\sigma}_{\varepsilon_{L_2E}}$. Panel (b): data points, $L_2E - 1$ and $L_2E - 2$ regression lines.

4 Mixture of Regression Models by L_2E

As outlined in Example I, the L_2E criterion allows us to detect the presence of two or more clusters in a dataset, through the analysis of function (3.4) for a given $\hat{\sigma}_{\varepsilon_{L_2E}}$, and it may lead to the estimate of two or more regression models.

However, it should be pointed out that the regression models are the same for each cluster and their parameters are estimated assuming a common variance of residuals while, in addition, we can not have an estimate of the size of each cluster. These problems may be overcome if we consider fitting mixture of regression models by L_2E.

Let us now consider a more complex model which assumes that each data point (x_i, y_i) comes from the k-th regression model; i.e., $\forall\, i = 1, \ldots, n$ and $\forall\, k = 1, \ldots, K$

$$Y_i = m_{\boldsymbol{\beta}_k}(x_i) + \varepsilon_k \qquad \text{with prob. } p_k \tag{4.1}$$

where $\sum_{k=1}^{K} p_k = 1$. This is to say that we assume that the model that best fits the data is a *mixture of $K \geqq 2$ regression models*.

If we furthermore assume that $\varepsilon_k \sim \mathcal{N}(0, \sigma_k^2)$, we may then state that each (x_i, y_i) stem from a bivariate r.v. (X, Y) and that the response Y conditioned on x follows a mixture of Normal r.vs. In other words, $\forall\, k = 1, \ldots, K$, with probability p_k such that $\sum_{k=1}^{K} p_k = 1$, we have $Y|x \sim \mathcal{N}(m_{\boldsymbol{\beta}_k}(x), \sigma_k^2)$, and hence

$$f(y_i|x_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k\, \phi(y_i|m_{\boldsymbol{\beta}_k}(x), \sigma_k^2)$$

We are now able to derive a close form for the estimates of the parameters

$(p_k, \boldsymbol{\beta}_k, \sigma_k)$, for a given set of K regression models, according to the ISE criterion as outlined in Section 2.

In fact, if we observe that (Basu et al., 1998)

$$\int_{I\!R} f(y|\boldsymbol{\theta})^2 \, dy = \frac{1}{n} \sum_{i=1}^{n} \int_{I\!R} f(y_i|x_i, \boldsymbol{\theta})^2 \, dy_i$$

where

$$\begin{aligned} f(y_i|x_i, \boldsymbol{\theta})^2 &= \left[\sum_{k=1}^{K} p_k \, \phi(y_i|m_{\boldsymbol{\beta}_k}(x_i), \sigma_k^2) \right]^2 = \\ &= \sum_{j=1}^{K} \sum_{l=1}^{K} p_j \, p_l \, \phi(y_i|m_{\boldsymbol{\beta}_l}(x_i), \sigma_l) \; \phi(y_i|m_{\boldsymbol{\beta}_j}(x_i), \sigma_j) \end{aligned}$$

and recalling that $\int_{I\!R} \phi(y|\mu_1, \sigma_1^2) \; \phi(y|\mu_2, \sigma_2^2) \, dy = \phi(0|\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$, then condition (2.4) becomes

$$\begin{aligned} &(\hat{p}_k, \hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k)_{L_2E} = \\ &= \underset{p_k, \boldsymbol{\beta}_k, \sigma_k}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} \sum_{l=1}^{K} p_j \, p_l \, \phi(0|m_{\boldsymbol{\beta}_j}(x_i) - m_{\boldsymbol{\beta}_l}(x_i), \sigma_j^2 + \sigma_l^2) - \right. \\ &\left. - \frac{2}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} p_k \, \phi(y_i|m_{\boldsymbol{\beta}_k}(x_i), \sigma_k^2) \right] \end{aligned} \tag{4.2}$$

4.1 The mixture of K simple linear regression models

An interesting case arises when we consider a mixture of K simple linear regression models. This is to say that, $\forall \, i = 1, \ldots, n$ and $\forall \, k = 1, \ldots, K$, $m_{\boldsymbol{\beta}_k}(x_i) = \beta_{0_k} + \beta_{1_k} \, x_i$, with probability p_k. Since in this situation equation (4.1), with probability p_k, becomes $Y_i = \beta_{0_k} + \beta_{1_k} \, x_i + \varepsilon_k$, it is easy to check that L_2E criterion for the $4\,K$ parameters $(p_k, \beta_{0_k}, \beta_{1_k}, \sigma_k)$, according to (4.2), leads to

$$\begin{aligned} &(\hat{p}_k, \hat{\beta}_{0_k}, \hat{\beta}_{1_k}, \hat{\sigma}_k)_{L_2E} = \underset{\hat{p}_k, \hat{\beta}_{0_k}, \hat{\beta}_{1_k}, \hat{\sigma}_k}{\arg\min} \\ &\left[\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} \sum_{l=1}^{K} p_j \, p_l \, \phi(0|\beta_{0_j} + \beta_{1_j} \, x_i - \beta_{0_l} - \beta_{1_l} \, x_i, \sigma_j^2 + \sigma_l^2) - \right. \\ &\left. - \frac{2}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} p_k \, \phi(y_i|\beta_{0_k} + \beta_{1_k} \, x_i, \sigma_k^2) \right] \end{aligned} \tag{4.3}$$

If we furthermore think to the mixture of $K = 2$ simple linear regression models, then, $\forall \, i = 1, \ldots, n$ and $\forall \, k = 1, 2$, we have

$$\begin{aligned} Y_i &= \beta_{0_1} + \beta_{1_1}\, x_i + \varepsilon_{i_1} \qquad \text{with prob. } p_1 \\ Y_i &= \beta_{0_2} + \beta_{1_2}\, x_i + \varepsilon_{i_2} \qquad \text{with prob. } p_2 = 1 - p_1 \end{aligned}$$

Since in this case

$$f(y_i|x_i, \boldsymbol{\theta}) = p_1\, \phi(y_i|\beta_{0_1} + \beta_{1_1}\, x_i, \sigma_1^2) + p_2\, \phi(y_i|\beta_{0_2} + \beta_{1_2}\, x_i, \sigma_2^2)$$

and

$$\begin{aligned} &\int_{I\!R} f(y|x, \boldsymbol{\theta})^2\, dy = \\ &= \sum_{j=1}^{2}\sum_{l=1}^{2} p_j\, p_l\, \phi(y_i|\beta_{0_l} + \beta_{1_l}\, x_i, \sigma_l^2)\phi(y_i|\beta_{0_j} + \beta_{1_j}\, x_i, \sigma_j^2) = \\ &= p_1^2\, \phi(0|0, 2\,\sigma_1^2) + p_2^2\, \phi(0|0, 2\,\sigma_2^2) + \\ &+ 2\, p_1\, p_2 \phi(0|\beta_{0_1} + \beta_{1_1}\, x_i - \beta_{0_2} - \beta_{1_2}\, x_i, \sigma_1^2 + \sigma_2^2) = \\ &= \frac{p_1^2}{2\,\sigma_1\sqrt{\pi}} + \frac{p_2^2}{2\,\sigma_2\sqrt{\pi}} + 2\, p_1\, p_2\, \phi(0|\beta_{0_1} + \beta_{1_1}\, x_i - \beta_{0_2} - \beta_{1_2}\, x_i, \sigma_1^2 + \sigma_2^2) \end{aligned}$$

then condition (4.3) reduces to

$$\begin{aligned} &(\hat{p}_1, \hat{p}_2, \hat{\beta}_{0_1}, \hat{\beta}_{1_1}, \hat{\beta}_{0_2}, \hat{\beta}_{1_2}, \hat{\sigma}_1, \hat{\sigma}_2)_{L_2E} = \underset{p_1, p_2, \beta_{0_1}, \beta_{1_1}, \beta_{0_2}, \beta_{1_2}, \sigma_1, \sigma_2}{\arg\min} \\ &\left[\frac{p_1^2\, \sigma_2 + p_2^2\, \sigma_1}{4\, \sigma_1\, \sigma_2\, \pi} + \frac{2}{n}\sum_{i=1}^{n} p_1\, p_2\, \phi(0|\beta_{0_1} + \beta_{1_1}\, x_i - \beta_{0_2} - \beta_{1_2}\, x_i, \sigma_1^2 + \sigma_2^2) - \right. \\ &\left. - \frac{2}{n}\sum_{i=1}^{n} p_1\, \phi(y_i|\beta_{0_1} + \beta_{1_1}\, x_i, \sigma_1^2) + p_2\, \phi(y_i|\beta_{0_2} + \beta_{1_2}\, x_i, \sigma_2^2) \right] \end{aligned} \tag{4.4}$$

Again closed forms for equations (4.3) and (4.4) are unlikely to be found, so we have to resort to numerical minimization algorithms.

4.2 Numerical example II

Let us consider a simulated dataset of $n = 200$ points, 100 of which come from model $y = 1 + 0.5\, x + \varepsilon_1$ and the remaining from model $y = 5 - 0.2\, x + \varepsilon_2$, where $\varepsilon_1 \sim \mathcal{N}(0, 1)$ while $\varepsilon_2 \sim \mathcal{N}(0, 0.5)$ and $x \sim \mathcal{U}(1,\, 10)$.

Figure 4, panel (a), shows the data points as well as the estimate of the simple linear regression model for which $\hat{\boldsymbol{\beta}}_{MLE} = [2.9195,\, 0.1593]^t$ and $R^2 = 0.1026$ with an associated p value of the F statistics equal to $3.784e - 06$, and $\hat{\sigma}_\varepsilon = 1.219$.

Clearly the MLE regression line seems not to be satisfying in fitting the data and we may think that, for a fixed x, the response Y comes from a *mixture of two simple linear regression models*, i.e.

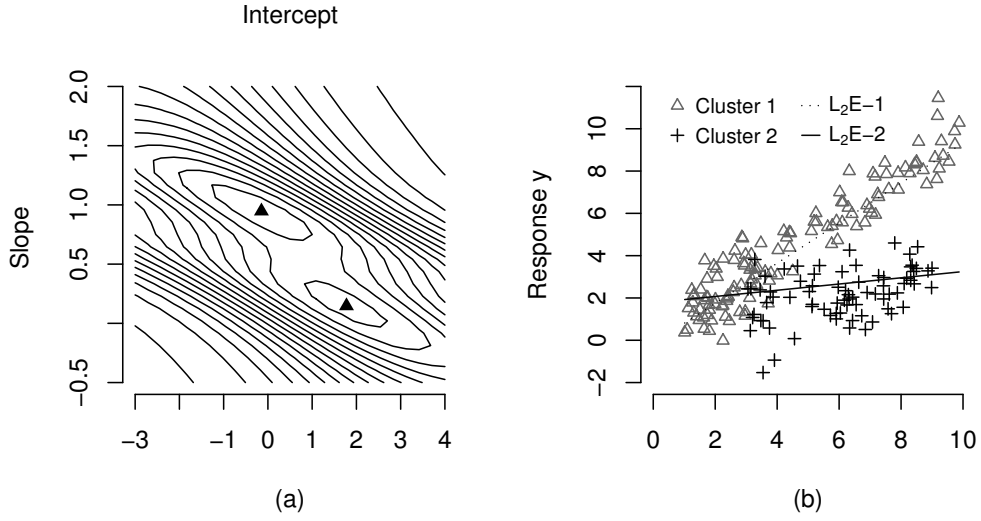

Figure 4: Panel (a): data points and MLE regression line. Panel (b): data points, L_2E-1 and L_2E-2 regression lines.

$$\begin{aligned} Y_i &= \beta_{0_1} + \beta_{1_1}\, x_i + \varepsilon_{i_1} \qquad \text{with prob. } p_1 \\ Y_i &= \beta_{0_2} + \beta_{1_2}\, x_i + \varepsilon_{i_2} \qquad \text{with prob. } p_2 \end{aligned}$$

In this case we simply have to find the solutions of (4.4) with respect to seven parameters $p_1, \beta_{0_1}, \beta_{1_1}, \beta_{0_2}, \beta_{1_2}, \sigma_1$ and σ_2, since $p_2 = 1 - p_1$. From numerical minimization, we obtain the following L_2 estimates

$$\begin{array}{llll} \hat{p}_1 = 0.415 & \hat{\beta}_{0_1} = 1.043 & \hat{\beta}_{1_1} = 0.511 & \hat{\sigma}_1 = 0.961 \\ \hat{p}_2 = 0.585 & \hat{\beta}_{0_2} = 4.996 & \hat{\beta}_{1_2} = -0.191 & \hat{\sigma}_2 = 0.541 \end{array}$$

Data points and L_2 regression lines are displayed in Figure 4, panel (b). Figure 5, panel (a) and (b), displays the histogram and the kernel density estimate of the distribution of residuals from each L_2E regression lines, i.e.

$$\begin{aligned} \varepsilon_{i_1} &= y_i - 1.043 - 0.511\, x_i \qquad \text{with prob. } p_1 = 0.415 \\ \varepsilon_{i_2} &= y_i - 4.996 - 0.191\, x_i \qquad \text{with prob. } p_2 = 0.585 \end{aligned}$$

It is worthwhile to observe that the range of residuals is quite wide and at the same time the behaviour of the ties of the kernel density estimates is "bad". This is due to the fact that the errors are calculated over all the sample points.

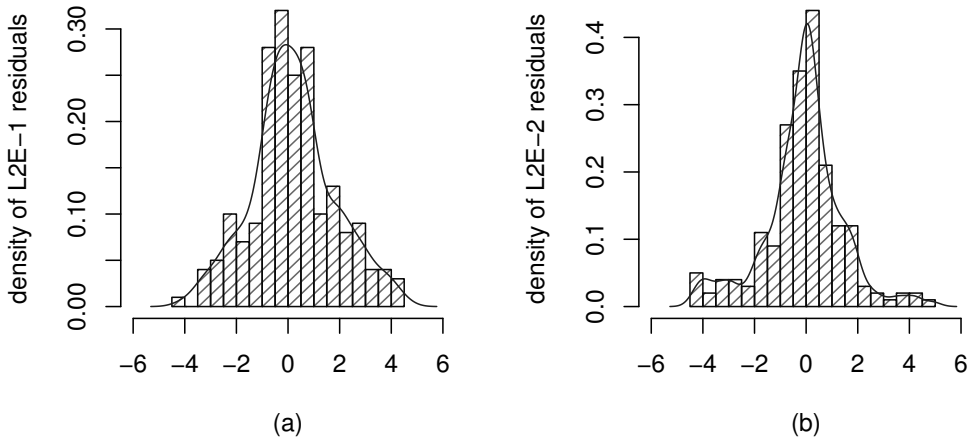

Figure 5: Panel (a): histogram and kernel density estimate of residuals from L_2E-1 regression line. Panel (b): histogram and kernel density estimate of residuals from L_2E-2 regression line.

5 Clusters identification

In the previous sections we outlined how L_2E may be used in fitting mixtures of regression models and how it allows us to simultaneously estimate the parameters of each regression model, the variances of the errors $\sigma^2_{\varepsilon_k}$ and the frequency p_k of data points belonging to each cluster.

Wishing to highlight which data point (x_i, y_i) belongs to each cluster we suggest a "quick rule", based on Tchebychev's inequality.

With regard to a mixture of two regression models, we state the following

$$\text{if } |\varepsilon_{i_1}| \leqq \gamma\,\sigma_{\varepsilon_1} \text{ and } |\varepsilon_{i_2}| > \gamma\,\sigma_{\varepsilon_2} \quad \rightarrow \quad (x_i, y_i) \in \text{Cluster I}$$

$$\text{if } |\varepsilon_{i_1}| > \gamma\,\sigma_{\varepsilon_1} \text{ and } |\varepsilon_{i_2}| \leqq \gamma\,\sigma_{\varepsilon_2} \quad \rightarrow \quad (x_i, y_i) \in \text{Cluster II}$$

$$\text{if } |\varepsilon_{i_1}| \leqq \gamma\,\sigma_{\varepsilon_1} \text{ and } |\varepsilon_{i_2}| \leqq \gamma\,\sigma_{\varepsilon_2} \quad \rightarrow \quad (x_i, y_i) \in \text{Unknown cluster}$$

$$\text{if } |\varepsilon_{i_1}| > \gamma\,\sigma_{\varepsilon_1} \text{ and } |\varepsilon_{i_2}| > \gamma\,\sigma_{\varepsilon_2} \quad \rightarrow \quad (x_i, y_i) \in \text{Outliers cluster}$$

Figure 6 shows how the previous rule may work in practice, applied to data set of the two numerical examples of Section 3.1 and 4.2.

For completeness, if we apply condition (4.4) to the dataset of Example I, from numerical minimization we obtain the following estimates

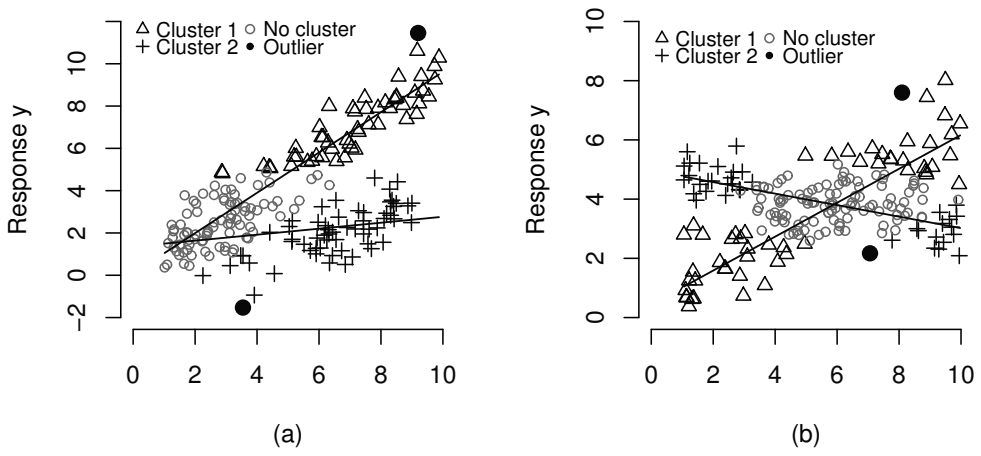

Figure 6: Clusters identification with data of Example I, panel (a), data from Example II, panel (b).

$$\hat{p}_1 = 0.521 \qquad \hat{\beta}_{0_1} = 0.085 \qquad \hat{\beta}_{1_1} = 0.953 \qquad \hat{\sigma}_1 = 0.800$$
$$\hat{p}_2 = 0.479 \qquad \hat{\beta}_{0_2} = 1.354 \qquad \hat{\beta}_{1_2} = 0.141 \qquad \hat{\sigma}_2 = 1.078$$

Clearly, the more the regression lines are far away from each others the less are the data points we are not able to assign. In other to highlight exactly two clusters a deeper analysis of phenomena under study is needed that eventually takes into account other explanatory variables.

6 The case study

A firm operating in the field of diagnosis and decontamination of electronic transformers fluids gives a judgment about risks of fluid degradation, electric shocks, firing or explosion, PCB contamination and decomposition of cellulosic insulation.

With the aid of well known mathematical models, based on the results of chemical analysis of the oil, the firm's staff achieve risk's values on continuous scales.

To verify that their methods of assigning risk's values are independent of specific characteristics of transformers (age, voltage, fluid quantity, ...) we investigated on the relation between risk's values.

In order to achieve this goal, we worked on a database of $1,215$ records of diagnosis, conducted on mineral oil distribution transformers, containing: oil chemical analysis, transformers' technical characteristics and risk's values.

Considering the risks of fire and of electric shocks, we obtained the scatterplot displayed in Figure 7. It was natural to suppose a linear dependance between the

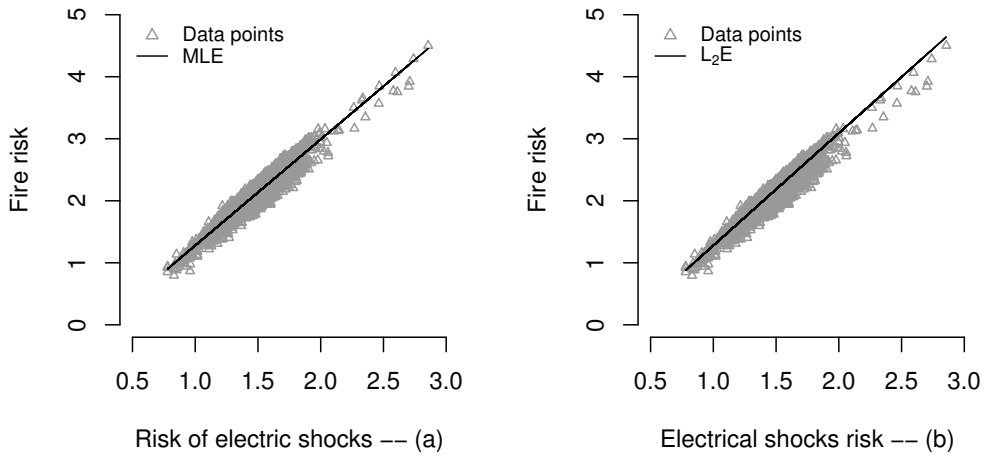

Figure 7: Panel (a): data points and MLE regression line. Panel (b): data points L_2E regression line.

two variables. Figure 7, panel (a), displays the MLE regression line while panel (b) displays the L_2E regression line. The lines estimated with the two methods are different and it can be noted that the one obtained with the L_2E criterion has a slope greater than the MLE's one.

A simple analysis on residuals from MLE regression line (Figure 8 panel (a)) suggests that errors are not normally distributed and, since the kernel estimated density is clearly *bimodal*, we may argue to be in presence of clustered data.

For this reason we supposed that the model that best fits our data is a mixture of two linear regression models. The results, according to the L_2E criterion introduced in Section 4, told us, Figure 8 panel (b), that about 43% of the data points follow the model $y = -0.369 + 1.568\,x$ ($L_2E - 1$), for which $\hat{\sigma}_\varepsilon = 0.078$ and that the remaining data points follow the model $y = -0.373 + 1.750\,x$ ($L_2E - 2$), for which $\hat{\sigma}_\varepsilon = 0.054$.

At this point, resorting to the rule proposed in Section 5, we were able to classify the data points according to the fact that they may follow the first or the second regression model. The results are summarized in Figure 9, panel (a). In this way, using $\gamma = 3$, we assigned 453 points (37.3%) to the $L_2E - 1$ regression line (cluster 1), 488 points (40.2%) to the $L_2E - 2$ regression line (cluster 2), 4 points were classified as outliers and 270 (22.2%) were the points we were not able to assign.

To assign this 270 points to one of the two clusters we had to investigate on specific characteristics of transformers. We observe that in our database 40% of the transformers had an amount of fluid $\leqq 500$ kg and that the L_2E criterion gave us an estimate of 43% of point for $L_2E - 1$ regression line. Furthermore, in our classification 423 out of the 453 points belonging to cluster 1 had an amount of fluid less (or equal) than 500 kg and all 488 transformers belonging to Cluster II had an amount of fluid greater than 500 kg. We thought to use the amount of fluid as

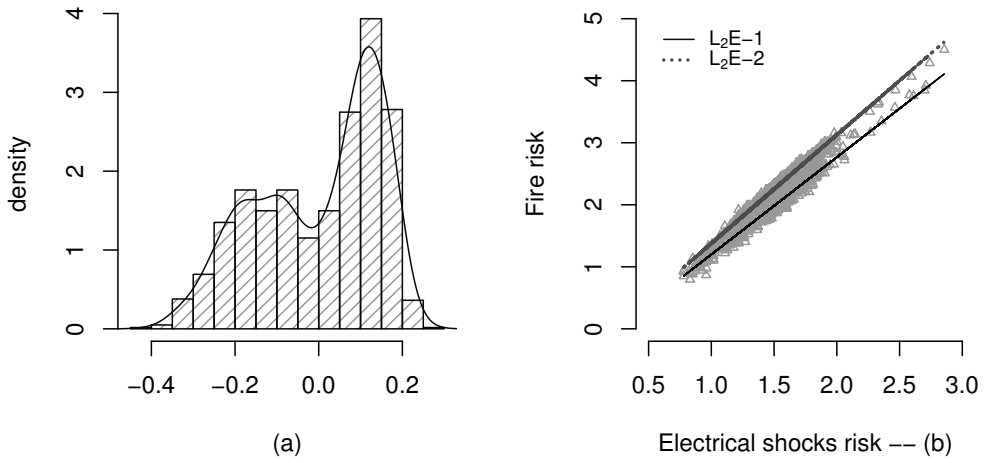

Figure 8: Panel (a): residuals from MLE regression line. Panel (b): data points and the two L_2E regression lines.

stratification variable and so we assigned the transformers with an amount of fluid less (or equal) than 500 kg to the L_2E-1 regression line and the transformer with an amount of fluid greater than 500 kg to the L_2E-2 regression line as shown in Figure 9, panel (b).

These results allowed us to say that, for a fixed level of risk of electric shocks, the risk of fire was evaluated in a different way for the two groups of transformers. This is to say that the relationship between the two variables depended on the amount of fluid contained in the transformers. The chemical staff of the firm did not find any realistic justification in explaining the different behaviour of risk of fire for the two type of transformers, so they decided to change the model used to give the risk of fire assigning different weights to hydrocarbon variable.

Considering the L_2E regression line of Figure 7 we are now able to understand why its slope is greater than the one of MLE regression line. Accordingly to Scott (2001b) this is due to the fact that L_2E regression line best fits the high dimension cluster.

7 Conclusion

In this paper we outlined an approach in diagnostic and building useful regression models based on L_2 estimate, investigating also on fitting mixture of regression models. This procedure allows to simultaneously estimate the parameters of each regression model, the variances of the errors and the probability that data point belongs to each regression model. Furthermore, we proposed a simple rule based on Tchebychev's inequality to identify the presence of clusters in the data.

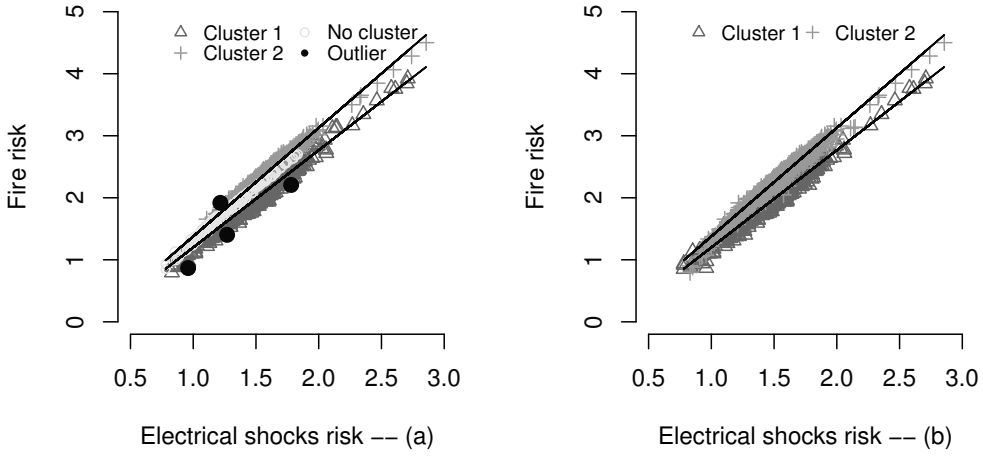

Figure 9: Panel (a): clusters identification. Panel (b): data clustering.

We found that this approach well suits a preliminary phase of data analysis. In our case study we have been able to suggest to chemical staff that the population of electrical tranformers had to be treated in different way.

The same framework in model building may be extended to complex forms, such as Generalized Linear Models and functional data analysis.

References

[1] Basu, A., Harris, I. R, Hjort, N.L., and Jones, M.C. (1998): Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, **85**.

[2] Reiss, R.D. and Thomas, M. (1997): *Statistical Analysis of Extreme Values*. Basel: Birckhäuser.

[3] Rudemo, M. (1982): Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**.

[4] Scott, D.W. (2001a): From kernel to mixtures. *Technometrics*, **43**.

[5] Scott, D.W., (2001b): Parametric statistical modeling by minimum integrated square error. *Technometrics*, **43**.

[6] Wand, M.P. and Jones, M.C. (1995): *Kernel Smoothing*. London: Chapman & Hall.