UNIVERSITÀ
DEGLI STUDI
DI UDINE

DIPARTIMENTO
DI SCIENZE ECONOMICHE
E STATISTICHE

Simple outlier labelling based on quantile regression, with application to the steelmaking process

Ruggero Bellio, Mauro Coletto

Luglio 2014

n. 2 / 2014

Statistica Metodologica e Sperimentale

Simple outlier labelling based on quantile regression, with application to the steelmaking process

Ruggero Bellio, University of Udine (Italy) *
Mauro Coletto, IMT Institute for Advanced Studies, Lucca (Italy)

Abstract

This paper proposes some methods for outlier identification in the regression setting, motivated by the analysis of steelmaking process data. Modern steelmaking processes produce a large body of data, and it is essential to analyze them for monitoring the production process. Here we focus in particular on settings where the response variable is given by the energy level consumed at the plant, in relation with another variable, such as the oxygen level. The paper proposes a methodology that extends to the regression setting the boxplot rule, commonly used for outlier screening with univariate data. The focus here is on bivariate settings with a single covariate, but the method is defined more generally. The proposal is based on quantile regression, including an additional transformation parameter for selecting the best scale for linearity of the conditional quantiles. The resulting method is used to perform labeling of potential outliers, with a quite low computational complexity, allowing for simple implementation within statistical software as well as simple spreadsheets. Some simulation experiments and application to real life examples investigate and illustrate the methodology.

Keywords: Boxplot rule; Outlier; Quantile regression; Single-index model; Steelmaking process.

*Address for correspondence: Ruggero Bellio, Department of Economics and Statistics, University of Udine, Via Tomadini 30/A, I-33100 Udine (Italy) ruggero.bellio@uniud.it

1 Introduction

Outlier detection is a fundamental task of data analysis in virtually any field of application. The statistical literature on the subject is very extensive, starting from classical references (e.g Barnett & Lewis, 1994; Rousseeuw & Leroy, 1987), and including essentially all the texts on regression models (e.g. Cook & Weisberg, 1982; Fox & Weisberg, 2011). To some extent also nearly all the literature on robust statistics has covered the task (e.g. Atkinson & Riani, 2000; Huber & Ronchetti, 2009; Maronna, Martin, & Yohai, 2006), not to mention that approaches developed in machine learning and related fields have also treated the subject (e.g. Aggarwal, 2013; Han, Kamber, & Pei, 2006; Hodge & Austin, 2004).

Modern steelmaking companies make intensive usage of process data, collected and analysed at the production plant. The main purposes of this activity is to monitor the stability of the production process, to evaluate the quality of the production, to increase production volumes and to prevent failures, increasing the overall efficiency of the plant. Anomalous conditions are monitored since they can lead to dangerous failures and large production losses, therefore outlier detection is an essential task. For basic information about the steelmaking process, see Turkdogan (1996) and the references therein. Some introductory general information can also be found at `http://www.steeluniversity.org`.

In our study we consider data coming from an Electric Arc Furnace (EAF) plant. The EAF is a plant that produces steel from ferrous scrap and various kind of iron units. It consumes a huge amount of resources, resulting in a quite energy-intensive process, therefore a variable of primary interest is the energy consumed in the melting process. Indeed, monitoring the consumed energy is an important task leading to improvements in energy efficiency and reduced production costs. The actual plant at which the data used in this paper were obtained is not disclosed to preserve proprietary information, but to some extent the data employed in this paper are representative of a broad class of similar datasets.

Usage of statistical process control techniques, such as control charts, has an important role in steelmaking (e.g. Kano & Nakagawa, 2008). Univariate plots, such as control charts, may be used to monitor the energy consumptions and detect anomalous points, yet multivariate approaches are essential to monitor the consumed energy level with respect to other variables characterizing the melting process, such as the O_2 level (Turkdogan, 1996, Chap. 8). Figure 1 displays an illustrative data set, showing the energy consumption (Unit of measurement: kWh/ton - kilowatt-hour per charged ton of scrap material in the furnace) against the O_2 level (Unit of measurement: SCF/ton - standard cubic foot per charged ton) for a sample of $n = 1216$ heats collected in an EAF. For the remainder of this paper, this data set will be denoted as the D_1 data set. A plot of studentized deletion residuals obtained from simple linear regression (e.g. Cook & Weisberg, 1982, §2.2) quickly detects some outlying observations, which would not be flagged by univariate displays.

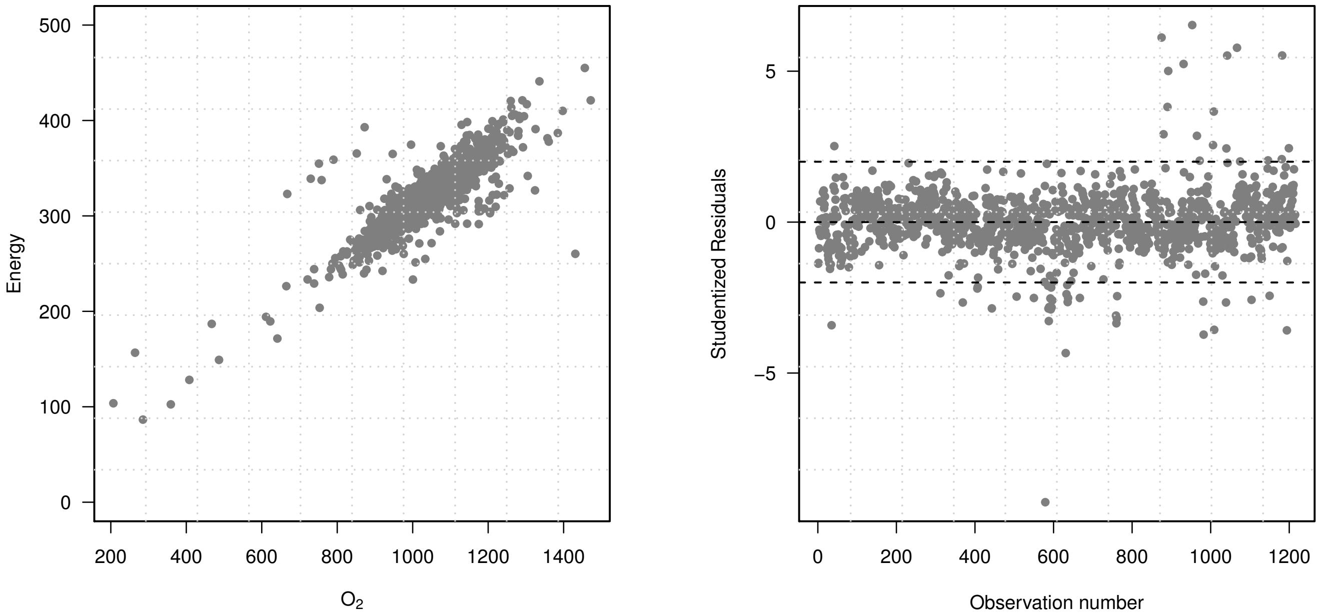

Figure 1: D_1 data. Left panel: plot of energy level against O_2 level for a sample of $n = 1216$ heats. Right panel: index plot of studentized deletion residuals, with horizontal lines at ±2 added.

Regression diagnostics (Belsley, Kuh, & Welsch, 1980; Cook & Weisberg, 1982) can be effective in outlier detection, and they have the advantage of being based on a method as simple as linear regression. At the same time, the latter fact is the main limitation of such approach, as the lack of robustness of linear regression prevents its routine application for semi-automatic outlier identification.

The aim of this paper is to propose a simple and robust method for outlier labeling based on quantile regression, first introduced by Koenker and Bassett (1978) and for which Koenker (2005) provides a broad overview. The method is based on a single index assumption, achieving semi-automatic transformation to the scale of the response variable where linearity of conditional quantiles is supported by the available data. The methodology extends to the regression setting the boxplot rule for univariate settings (Tukey, 1977). The resulting method seems quite effective and, at the same time, very convenient for practical implementation. We stress that the method implements *outlier labeling*, meaning that we endorse screening for potential outliers rather than formal detection based on statistical testing.

The plan of the paper is as follows. Section 2 summarizes some useful theory on outlier detection, whereas Section 3 provides the methodology proposed in this paper. Results from some simulation studies are summarized in Section 4, while results from empirical examples are given in Section 5. Some final remarks conclude the paper.

2 Background on outlier identification

As mentioned in the introduction, the existing body of literature on outlier detection is very extensive, and here we confine the attention to those methods which are close to the approach proposed here. In the univariate setting, some authors such as Brant (1990) and Barnett and Lewis (1994, Chap. 2) make a distinction between *classical* and *resistant* outlier rules, and to some extent the same principle can be extended to multivariate settings.

The boxplot rule (Tukey, 1977) is probably the most commonly used resistant rule for outlier screening. Given a sample of size n, if q_L and q_U are the upper and lower quartiles, the inner and outer fences are given by

$$\mathrm{IF}_L = q_L - k\,(q_U - q_L)$$

and

$$\mathrm{IF}_U = q_U + k\,(q_U - q_L)\,,$$

for a certain choice of the constant k. Tukey defined any observations falling below IF_L or above IF_U for $k = 1.5$ as *outside outliners*, while those falling outside the fences for $k = 3$ as *far out outliners*. The boxplot rule is appealingly simple and quickly gained a widespread usage in applied statistics, linked to the popularity of the boxplot for exploratory data analysis. Hoaglin, Iglewicz, and Tukey (1986) were among the first that made a careful study of this procedure. They noted that, being based on robust measures, in most cases it avoids the problem of masking, which arises where the presence of multiple outliers makes them difficult to detect. The same authors made also a study of the swamping properties of the procedure, that is the tendency to misclassify observations as outliers. They made a distinction between the *outside rate per observation*, that is the misclassification rate for a single observation, and the *some-outside rate per sample*, which corresponds to the sample-wise error rate in simultaneous inference. Hoaglin et al. (1986) noted that, under some parametric distribution for the data, both the two error rates for the boxplot rule have a strong dependence on the sample size. Brant (1990) made a thorough comparison between the boxplot rule and the classical ESD rule (Rosner, 1983). He focused in particular on the some-outside rate per sample, coming to the conclusion that classical rules and the boxplot rule may be equally effective, with perhaps an exception for normal data where classical rules may be preferable. We note in passing that Brant (1990) made the important remark that "the outlier problem is sufficiently ill-posed that optimal solutions are not feasible", implying that the best that can be done is to investigate on how some proposed methods perform in some situations of interest.

The widespread usage of the boxplot rule lead several authors to propose adjustments designed to make it more suitable for some specific situations. Among the many proposals, the median rule introduced by Carling (2000) replaces the fences with $q_M \pm k_1\,(q_U - q_L)$, where q_M is the sample median. Similarly, Kimber (1990) proposed the use of semi-interquartile ranges to adapt to asymmetric samples, obtaining the fences

$$\mathrm{IFS}_L = q_M - 2\,k_2\,(q_M - q_L)\,, \qquad\qquad \mathrm{IFS}_U = q_M + 2\,k_2\,(q_U - q_M)\,.$$

The paper by Carling (2000) is noteworthy also because a method for determining k and k_1 is proposed, for the boxplot and the median rule respectively. Carling's method is designed for achieving a predetermined outside rate per observation for a given sample size, using some information about the data distribution. The method by Schwertman and de Silva (2007), instead, extends the boxplot rule to control the some-outside rate per sample for a given sample size. Similarly to the ESD rule, their method can be used to identify multiple outliers. The Carling and Schwertman and de Silva methods are compared in Carter, Schwertman, and Kiser (2009).

Moving from the univariate to the regression setting, things get quickly more complicated, as multivariate outliers are surely more challenging (Rousseeuw & Leroy, 1987). Our aim is to obtain a resistant rule, in the same spirit of the boxplot rule. Direct extension of the boxplot to more than one dimension do exist, at least in the bivariate case (Rousseeuw, Ruts, & Tukey, 1999), but they are not quite as practical as the univariate version. In a sense, the most direct way to obtain a resistant rule in a regression setting employs instead quantile regression (e.g. Koenker, 2005). The idea is actually very simple, and it is best presented in the bivariate case, though extensions to several regressors are feasible, and they will be touched in passing later on.

Let $Q_{Y|x}(\tau|x)$ the conditional quantile function of the response variable for a given value x of the regressor, for any $\tau \in (0,1)$. Merely by analogy with the unidimensional case, we can simply define the fences for outlier labeling as

$$\begin{aligned}\mathrm{IF}_L(x) &= Q_{Y|x}(0.25|x) - k\,\{Q_{Y|x}(0.75|x) - Q_{Y|x}(0.25|x)\} \qquad (1)\\ \mathrm{IF}_U(x) &= Q_{Y|x}(0.75|x) + k\,\{Q_{Y|x}(0.75|x) - Q_{Y|x}(0.25|x)\}\,.\end{aligned}$$

The method was actually proposed by Cho, Kim, Jung, Lee, and Lee (2008) for mass spectrometry experiments and Eo, Hong, and Cho (2014) for censored data. Both these two papers were supported by some `R` software (R Core Team, 2014), based on the `quantreg` library for quantile regression (Koenker, 2013).

A resistant rule based on quantile regression is appealing for the data of interest in this paper, as outlying values are more likely to occur in the response variable, rather than in the regressors. Quantile regression suits well this kind of situations, having a satisfactory breakdown point for fixed design points (Koenker, 2005, §2.3.2). The main problem for the application of outlier labeling based on (1) is the specification of the form for $Q_{Y|x}(\tau|x)$. It is tempting to specify a linear form $Q_{Y|x}(\tau|x) = \beta_0(\tau) + x\,\beta_1(\tau)$, but some attention is required. Figure 2 exemplifies two possible situations. The left panel shows the application of the resistant rule based on (1) with $k = 1.5$ for an artificial data set, simulated from the regression line fitted to the data of Figure 1 assuming normal error with constant variance, whereas the right panel displays the same procedure applied to the original data.

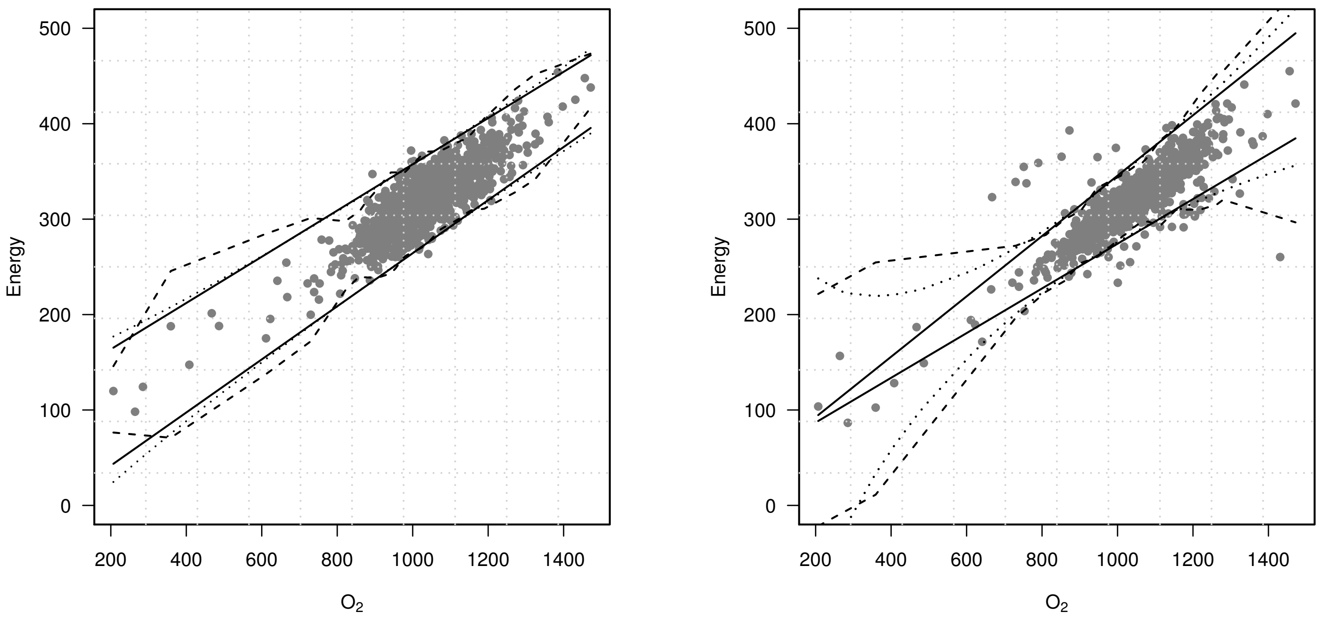

Figure 2: Plot of energy level against O_2 level for a simulated data set (left panel) and for the D_1 data set (right panel). Fences based on quantile regression computed with $k = 1.5$ are superimposed, for linear quantiles (solid lines), nonparametric quantiles (dashed lines) and for the single-index method introduced in Section 3 (dotted).

The plots report also the fences based on a nonparametric specification of the conditional quantiles, adopting the smoothing splines specification implemented by the `rqss` function of the `quantreg` library. For the artificial data set, the two specifications agree reasonably well, with perhaps a minor discrepancy for lower values of oxygen, where data sparsity prevents an accurate estimation of the regression quantiles for the nonparametric method. When $k = 1.5$, both methods flag few observations as potential outliers, 12 in the linear case and 19 in the nonparametric one respectively i.e. around 1% of the cases for either method. No observation is flagged by either method with $k = 3$. Things are quite different for the real data. They exhibit the presence of some outlying observations, and a substantial amount of heterogeneous variability, resulting in somewhat different fences for the linear and nonparametric quantiles. The method based on linear quantiles flags 53 and 16 observations with $k = 1.5$ and $k = 3$ respectively, whereas for the nonparametric quantiles these numbers are down to 38 and 12 respectively. Essentially, the discrepancy is due to the fact that the fences based on the linear specification converge to a single point for decreasing values of oxygen, flagging some points with low oxygen values that are not labelled by the nonparametric version. The fences based on the latter method are once again wider when the sample information is smaller, a rather sensible property.

The fact that with the linear specification of quantiles the upper and lower fences may tend to converge, and even cross for an x-value included in the sample, is a direct consequence of the fact that estimated linear quantiles for different values of τ will typically have different slopes $\widehat{\beta}_1(\tau)$. The problem of quantile crossing is well known in quantile regression (e.g Koenker, 2005, §2.5). Some solutions that avoid the crossing do exist, such as the constrained method proposed by Bondell, Reich, and Wang (2010). Such method, however, would not have any effect in cases like those of the right panel of Figure 2, where the crossing does not actually occur within the convex hull of the observed data points.

The nonparametric specification of the fences is not ideal either. In fact, whereas the linear quantile regression is very simple, as the linear coefficients for any τ can be easily estimated by standard linear programming (Koenker, 2005, Chap. 6), nonparametric specifications are much more challenging. Even when well-tested and dedicated software is employed, like for the examples of Figure 2, some care is required in order to obtain sensible estimates. For both the data sets the smoothness parameter λ has to be set to a value much larger than the default value of 1 for `rqss` to obtain smooth estimated quantiles; the plots were actually obtained with $\lambda = 50$. Even so, the obtained fences are marginally unsatisfactory, with some local nonlinearities which may require some further attention. The remainder of this paper will be devoted to a method for implementing the resistant rule (1), which combines computational simplicity with effectiveness for the kind of data of interest here.

3 A practical method for outlier labelling

It can be argued that quantile regression provides the best route for extending the boxplot rule for outlier labeling to regression settings, nonetheless, as mentioned in the previous section, some care is required for its practical implementation.

3.1 Quantile regression based on a single-index assumption

The main idea of the proposal of this paper is to look for a scale for which the linearity assumption for the conditional quantiles is supported from the data. This has been rather customary for standard linear regression based on ordinary least squares, since the pioneeristic paper by Box and Cox (1964) on parametric response transformations. For quantile regression, the same idea has found application in econometrics.

Powell (1991) proposed a nonlinear specification for the conditional quantiles, employing a single-index assumption. In particular, given a linear index $x^T\beta(\tau)$ for a generic vector of covariate values x, the form assumed for $Q_{Y|x}(\tau|x)$ is

$$Q_{Y|x}(\tau|x) = g\left\{x^T\beta(\tau), \lambda(\tau)\right\} \,, \tag{2}$$

where $g(\cdot)$ is a strictly increasing function in $x^T\beta(\tau)$, and it depends on the unknown parameter $\lambda(\tau)$ as well. A possible choice is to set the inverse of $g\left\{x^T\,\beta(\tau), \lambda(\tau)\right\}$ in the first argument equal to the Box-Cox transformation, given by

$$g^{-1}(y, \lambda) = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0\,. \end{cases} \tag{3}$$

To be precise, Powell (1991) assumed a more general power transformation than (3), but as a matter of fact all the published applications of this methodology for quantile regression have employed the Box-Cox transformations; see Fitzenberger, Wilke, and Zhang (2009) and the references therein. As $g\{\cdot, \lambda(\tau)\}$ is strictly increasing, and due to the fact that quantiles are order-preserving, by applying the inverse transformation $g^{-1}\{\cdot, \lambda(\tau)\}$ to both sides of (2) it follows that

$$Q_{g^{-1}\{Y,\lambda(\tau)\}|x}(\tau|x) = x^T\beta(\tau)\,,$$

namely the linear assumption holds for the quantiles of the response variable transformed according to $g^{-1}\{\cdot, \lambda(\tau)\}$.

In terms of the original response variable, when the Box-Cox transformation (3) is employed, from the form of the inverse function (3) we obtain

$$Q_{Y|x}(\tau|x) = \begin{cases} \{\lambda(\tau)\, x^T\beta(\tau) + 1\}^{1/\lambda(\tau)} & \text{if } \lambda(\tau) \neq 0 \\ \exp\{x^T\beta(\tau)\} & \text{if } \lambda(\tau) = 0\,. \end{cases} \tag{4}$$

The above form shows that the proposal implies a nonlinear form for the quantiles of the original response and, perhaps more importantly, that by letting the parameter $\lambda(\tau)$ indexing the transformation depend on τ we are not imposing to work on the same scale for all the quantiles of interest, thus achieving a noteworthy amount of flexibility.

3.2 Parameter estimation

Chamberlain (1994) suggested a two-step estimation procedure to estimate the parameters $\beta(\tau)$ and $\lambda(\tau)$ for a given value of τ. Using standard notation for quantile regression, let $\rho_\tau(u)$ denote the piecewise linear loss function

$$\rho(u) = u\left\{\tau - I(u < 0)\right\}.$$

It is a basic result that the coefficients of standard quantile regression for $Q_{Y|x}(\tau|x)$ can be estimated by minimizing $\sum_{i=1}^{n} \rho_\tau\{y_i - x_i^T \beta(\tau)\}$ (Koenker, 2005, Chap. 1). This fact can be used to define the two-step procedure for estimating $\{\beta(\tau), \lambda(\tau)\}$, for a given τ, reported below for a general transformation g.

1. For a given value $\lambda = \lambda(\tau)$, obtain the constrained estimate of $\widehat{\beta}_\lambda(\tau)$ as

$$\widehat{\beta}_\lambda(\tau) = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau\left\{g^{-1}(y_i, \lambda) - x_i^T \beta\right\} \tag{5}$$

2. Select the transformation parameter $\widehat{\lambda}(\tau)$ such that

$$\widehat{\lambda}(\tau) = \arg\min_{\lambda} \sum_{i=1}^{n} \rho_\tau\left[y_i - g\{x_i^T \widehat{\beta}_\lambda(\tau), \lambda\}\right] \tag{6}$$

   The final estimate of $\beta(\tau)$ is then given by $\widehat{\beta}(\tau) = \widehat{\beta}_{\widehat{\lambda}(\tau)}$.

The crucial point of the above algorithm is that both Step 1. and 2. are rather simple to implement. Indeed, (5) amounts to a standard linear quantile regression with response values given by $g^{-1}(y_i, \lambda)$, whereas (6) only requires the evaluation of the objective function over a grid of points for λ. As noted by Fitzenberger et al. (2009, p. 163), when the Box-Cox transformation is used it typically suffices to consider the interval $[-1.5, 2]$ as set of possible values for $\lambda(\tau)$.

3.3 Alternative transformations

The Box-Cox transformation (3) is a natural choice for $g^{-1}\{\cdot, \lambda(\tau)\}$, but it is not totally free of pitfalls. They are essentially two: (i) the method requires the response variable to be strictly positive; (ii) the inverse Box-Cox transformation may not be defined for all the observations and for every value $\lambda(\tau)$ of interest. The first problem is not an issue when the response variable corresponds to energy level, like for the data considered here, but it prevents the usage of the method for monitoring relative changes or other transformed variables. The second problem arises when the condition

$$\lambda\, x^T \widehat{\beta}_\lambda(\tau) + 1 > 0\,, \tag{7}$$

deriving from (4) and arising at Step 2. of the estimation algorithm, is not fulfilled. The problem is less serious than it may seem, as one could decide that those values of λ for which the condition (7) fails to be satisfied for all the data points are ruled out by the data or, alternatively, the procedure proposed by Fitzenberger et al. (2009) could be employed. The latter solution requires to select an interval $[\underline{\lambda}, \overline{\lambda}]$ for which the condition (7) holds for all the data points, and then solve for $\widehat{\lambda}(\tau)$ in (6) by restricting the search to such interval.

Even if the limitations of the Box-Cox transformation do not appear serious enough to prevent its utilization in (2), nonetheless some alternatives that may be slightly preferable do exist, and will be introduced here. The first one is the transformation introduced in Yeo and Johnson (2000), that has gained some popularity in applied statistics (e.g Yee, 2004). Usage of the Yeo-Johnson transformation as an alternative to (3) gives

$$g^{-1}(y,\lambda) = \begin{cases} \{(y+1)^\lambda - 1\}/\lambda & \text{if } y \geq 0, \lambda \neq 0 \\ \log(y+1) & \text{if } y \geq 0, \lambda = 0 \\ -\{(1-y)^{2-\lambda} - 1\}/(2-\lambda) & \text{if } y < 0, \lambda \neq 2 \\ -\log(1-y) & \text{if } y < 0, \lambda = 2\,. \end{cases} \tag{8}$$

As well described in Yeo and Johnson (2000) and Yee (2004), the transformation is symmetric in the first argument, and for positive y it amounts to a shifted Box-Cox transformation. The inversion formula gives the corresponding specification for the quantiles of the response variable

$$Q_{Y|x}(\tau|x) = \begin{cases} \{\lambda(\tau)\, x^T\beta(\tau) + 1\}^{1/\lambda(\tau)} - 1 & \text{if } x^T\beta(\tau) \geq 0, \lambda(\tau) \neq 0 \\ \exp\{x^T\beta(\tau)\} - 1 & \text{if } x^T\beta(\tau) \geq 0, \lambda(\tau) = 0 \\ 1 - \left[1 - x^T\beta(\tau)\{2-\lambda(\tau)\}\right]^{1/\{2-\lambda(\tau)\}} & \text{if } x^T\beta(\tau) < 0, \lambda(\tau) \neq 2 \\ 1 - \exp\{-x^T\beta(\tau)\} & \text{if } x^T\beta(\tau) < 0, \lambda(\tau) = 2 \end{cases} \tag{9}$$

The Yeo-Johnson transformation is a good proposal for replacing the Box-Cox transformation in the method sketched in §3.1. First, the transformation (8) is defined on the entire real line. Secondly, provided that $\lambda(\tau)$ is selected in the interval $[-2, 2]$, the existence of the inverse formula

(9) may be an issue only for $x^T\beta(\tau) \geq 0$ and $\lambda < 0$, whereas for the Box-Cox transformation the condition (7) may not hold also for some instances with $x^T\beta(\tau) < 0$ and $\lambda > 0$.

A further possibility is given by the dual power transformation proposed by Yang (2006). Like the Box-Cox transformation, the corresponding function $g^{-1}(y, \lambda)$ is defined only for $y > 0$. The expressions replacing (3) and (4) for such proposal are given by

$$g^{-1}(y, \lambda) = \begin{cases} (y^{\lambda} - y^{-\lambda})/(2\,\lambda) & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0\,, \end{cases} \tag{10}$$

and

$$Q_{Y|x}(\tau|x) = \begin{cases} \left[\lambda(\tau)\, x^T\beta(\tau) + \sqrt{1 + \lambda(\tau)^2\,\{x^T\beta(\tau)\}^2}\right]^{1/\lambda(\tau)} & \text{if } \lambda(\tau) \neq 0 \\ \exp\{x^T\beta(\tau)\} & \text{if } \lambda(\tau) = 0\,. \end{cases} \tag{11}$$

The dual power transformation (10) does not include the identity transformation, and it is symmetric in λ. Differently from both the Box-Cox and the Yeo-Johnson transformations, there is never an issue with the existence of (11) for any value of $\lambda(\tau)$ and $x^T\beta(\tau)$. Therefore, also the dual power transformation can be a sensible proposal for extending the boxplot rule for outlier labeling using quantile regression, as described in §3.1.

To give a flavor about the method, the results of the application of the boxplot rule adopting the quantile specification (2) are also included in Figure 2, using the Yeo-Johnson transformation. For the simulated data, the estimated transformation parameters $\widehat{\lambda}(\tau)$ are all close to 1 for $\tau = 0.25, 0.5, 0.75$, therefore the resulting fences track very closely those obtained with the linear specification of quantiles. For the D_1 data, we obtain $\widehat{\lambda}(0.25) = 1.38$, $\widehat{\lambda}(0.50) = 1.30$ and $\widehat{\lambda}(0.75) = 0.77$. The resulting fences are markedly different from the linear ones, and closer to the nonparametric ones, though some local differences exist. This seems an appealing result, as the computational effort of the single-index method is much smaller than that of the nonparametric method. The accuracy of proposed methodology will be investigated in the next sections.

4 Simulation studies

In order to investigate the properties of the proposed methodology, some simulation studies have been carried out. In particular, four different simulation settings were investigated. The case with a single covariate was considered, and the covariate values were randomly selected (without replacement) from the observed oxygen levels of the D_1 data. For each settings, the two sample sizes $n = 100$ and $n = 1,000$ were considered, and $5,000$ Monte Carlo replications were performed for a given sample size of each setting. In broad generality, the design of simulation studies for outlier labeling is a very challenging task, as there exist endless possibilities for introducing outliers in a given model. Here we attempt to focus on somewhat high rates of contaminated data. Although the resulting designs are probably not totally realistic, they allowed us to study the masking properties of the various methods. The features of the simulation settings are given as follows.

i. *Uncontaminated normal data.* Data for the response variable were generated from the linear regression model

$$y_i = 55 + 0.26\, x_i + 18\, \varepsilon_i \qquad \varepsilon_i \sim N(0,1)$$

where the model parameters were obtained from the model fitted to the D_1 data.

ii. *Contamination in the error term.* Same as Setting i., but the error term for 15% (on the average) of the observations was replaced by $\varepsilon_i + 4\,\mathrm{sgn}(\varepsilon_\mathrm{i})$.

iii. *Mixture of regression lines.* Same as Setting i., but a subset of 15% (on the average) of the observations were generated by a regression line with slope 0.36.

iv. *Contamination in the error term for lognormal data.* The log response was generated from the linear regression model

$$\log(y_i) = 0.13 + 0.81\, \log(x_i) + 0.06\, \varepsilon_i \qquad \varepsilon_i \sim N(0,1)$$

where once again the model parameters were obtained from the fitted model to the D_1 data on the log scale. A subset of the 10% of the observations (on the average) had their error term replaced by $\varepsilon_i + \mathrm{sgn}(\varepsilon_\mathrm{i})\ \max(4, |\mathrm{v_i}|)$, with $v_i \sim t(3)$. Here the logs of covariate values were taken as equally spaced values in the sample range of log oxygen for the D_1 data.

Four different procedures for outlier labeling where considered, comparing three resistant rules with a classical one. In particular, two different versions of the resistant rule (1) based on the formulation (2) were assessed, considering as the g-transformation both the Yeo-Johnson transformation (9) and the dual power transformation (11). We also included in the experiment the rule (1) applied without any transformation and adopting a linear specification for the quantiles. For all the three resistant rules, three different vales of k were considered, namely $k = 1.5, 2.0, 3.0$. Finally, the classical outlier test based on studentized deletion residuals (Cook & Weisberg, 1982, §2.2) was also considered, with 1% and 5% significant levels, including for the latter also a Bonferroni correction for multiple testing. Notice that the linear regression model was not considered for the original data, but indeed after transforming the response values using the Yeo-Johnson transformation (8), selecting the value of λ that makes the residuals of the regression as close to be normally distributed as possible. To this end, some `R` functions from the `car` package were employed (Fox & Weisberg, 2011).

4.1 Results for Setting i.

The first simulation study is entirely about the swamping properties of the various procedures, so both the *outside rate per observation* and the *some-outside rate per sample* were estimated by simulation. The results are reported in Table 1.

Table 1: Simulation results under uncontaminated normal data: outside rates expressed in percentages for various methods. For the outlier test, k is the significance level. 'O rate' is the *outside rate per observation*, 'S rate' is the *some-outside rate per sample.*

Algorithm	k	O rate	S rate	O rate	S rate
		$n = 100$		$n = 1000$	
Resistant (Y-J)	1.5	1.38	67.7	0.76	99.8
	2.0	0.47	35.4	0.10	60.8
	3.0	0.21	19.6	0.01	6.4
Resistant (Dual)	1.5	1.34	66.8	0.76	99.8
	2.0	0.44	33.3	0.10	60.9
	3.0	0.19	17.8	0.01	6.4
Resistant (Linear)	1.5	1.17	60.2	0.74	99.6
	2.0	0.28	21.5	0.09	56.9
	3.0	0.05	4.8	<0.01	0.68
Outlier test	0.05	5.01	100.0	5.00	100.0
	0.01	0.98	68.1	1.00	100.0
	$0.05/n$	0.05	4.6	<0.01	4.7

The results for the three resistant procedures are largely comparable, with a slight edge for the linear specification, which was advantaged by the simulation scenario. For what concerns the outside rate per observation, the resistant procedures offer a good protection even with $k = 1.5$, for which for $n = 100$ the rate is around 1%. Only the choice $k = 3.0$ offer a strong protection against swamping at the sample-wise level. However, the choice $k = 2.0$ does offer some protection at the sample-wise level. Indeed, the some-outside rates per sample for the three resistant methods with $k = 2.0$ are even better than the results for the univariate rule with $k = 1.5$, and the latter choice was deemed as satisfactory by Hoaglin et al. (1986, Table 2). Even the results for the three resistant rules proposed here with $k = 1.5$ are not terrible for $n = 100$ when compared with the univariate results. Finally, the results for the outlier test are totally as expected. When the 1% significance level is adopted, the swamping properties of the test are roughly similar to those of the resistant rules with $k = 1.5$. The Bonferroni adjustment gives the best performances, and they are roughly not so different from the results obtained with the resistance rules with $k = 3.0$. We notice in passing that modern theory of multiple testing, starting from the seminal contribution of Benjamini and Hochberg (1995), would offer some more satisfactory alternatives than the Bonferroni adjustment to be used in conjunction with the classical outlier test; see, for instance, Cerioli and Farcomeni (2011) and the references therein. At any rate, rather than a doable alternative, the Bonferroni-adjusted test is included here as a sort of extreme benchmark, useful for making a comparison with the resistant rules with $k = 3.0$.

4.2 Results for Setting ii.

This setting has a sizable percentage of outliers, which are generally well separated from the bulk of the data. The results for the various methods are reported in Table 2. The three resistant rules provide quite similar results, which differ instead from those of the classical outlier test. The most noticeable trend is that the resistant rules with $k = 1.5$ are generally quite effective in labelling the outliers, without losing much in terms of false detections, which are generally low for all the methods. The resistant rules with $k = 2.0$ are less effective, though they perform much better than the resistant rules with $k = 3.0$. The classical outlier tests are much more affected by masking, and actually only the test with significance level equal to 0.05 is able to identify around the 80% of the outliers. The Bonferroni-adjusted test is overwhelmingly conservative. In this respect, the resistant rules with $k = 3.0$ are less conservative, despite the similar behavior for what concerns swamping noted in Setting i.

4.3 Results for Setting iii.

The outliers for this setting are generally more overlapping than those of Setting ii., and for this reason this is a more challenging scenario. In fact, the results for this setting, reported in Table 3, are less good for the resistant rules with $k = 1.5$, but only slightly so. Once again, the three resistant rules perform quite similarly, and they greatly outperform the classical outlier test, which provides very little control over outlier masking.

4.4 Results for Setting iv.

This setting was designed for obtaining a situation where the issue about converging fences noticed for the resistant rule with linear conditional quantiles for the D_1 data could occur frequently. Therefore, a certain degree of variance heterogeneity was introduced by generating log-normally distributed responses. At the same time the percentage of genuine outliers was lowered, as the

Table 2: Simulation results under Setting ii., rates expressed in percentages for various methods. For the outlier test, k is the significance level. 'O rate' is the *outside rate per observation*, *True detection* and *False detection* are computed observation-wise.

Algorithm	k	O rate	True detection	False detection	O rate	True detection	False detection
			$n = 100$			$n = 1000$	
Resistant (Y-J)	1.5	14.3	93.6	0.51	15.1	99.6	0.13
	2.0	10.5	71.4	0.23	13.0	86.9	0.02
	3.0	2.5	17.1	0.16	1.4	9.5	0.01
Resistant (Dual)	1.5	14.3	93.9	0.50	15.1	99.6	0.16
	2.0	10.6	71.9	0.23	13.0	86.9	0.03
	3.0	2.5	16.9	0.16	1.4	9.5	0.01
Resistant (Linear)	1.5	14.4	95.1	0.37	15.1	99.9	0.12
	2.0	10.6	72.8	0.12	13.1	87.5	0.01
	3.0	2.3	16.4	0.05	1.4	9.3	<0.01
Outlier test	0.05	11.6	80.6	0.03	13.5	90.2	0.01
	0.01	3.0	23.7	<0.01	2.6	17.5	<0.01
	$0.05/n$	0.1	1.1	0.00	<0.01	<0.01	0.00

log-normal distribution would tend to generate some outliers in any case, though such outliers were not considered as such in computing the detection rates. Indeed, here some difference between the resistant rules based on the method of Section 3 and those based on the linear specification come out, showing the higher adaptive nature of the methodology which introduces the additional transformation parameters $\lambda(\tau)$. Like in the previous setting, the classical outlier test performs poorly.

Table 3: Simulation results under Setting iii., rates expressed in percentages for various methods. For the outlier test, k is the significance level. 'O rate' is the *outside rate per observation*, *True detection* and *False detection* are computed observation-wise.

Algorithm	k	O rate	True detection	False detection	O rate	True detection	False detection
			$n = 100$			$n = 1000$	
Resistant (Y-J)	1.5	12.8	85.2	0.52	14.3	94.8	0.13
	2.0	10.2	70.2	0.24	12.4	82.0	0.02
	3.0	4.3	31.0	0.17	4.4	29.4	0.01
Resistant (Dual)	1.5	13.0	86.1	0.51	14.3	94.8	0.13
	2.0	10.3	70.9	0.24	12.3	82.1	0.02
	3.0	4.3	30.8	0.17	4.4	29.4	0.01
Resistant (Linear)	1.5	13.1	87.7	0.40	14.3	95.0	0.11
	2.0	10.4	72.5	0.16	12.3	82.5	0.01
	3.0	4.3	31.2	0.09	4.4	29.6	<0.01
Outlier test	0.05	8.4	52.6	1.36	9.1	58.3	0.53
	0.01	1.6	11.4	0.32	1.4	7.9	0.27
	$0.05/n$	0.04	0.40	0.01	0.09	0.10	0.02

Table 4: Simulation results under Setting iv., rates expressed in percentages for various methods. For the outlier test, k is the significance level. 'O rate' is the *outside rate per observation*, *True detection* and *False detection* are computed observation-wise.

Algorithm	k	O rate	True detection	False detection	O rate	True detection	False detection
			$n = 100$			$n = 1000$	
Resistant (Y-J)	1.5	9.0	83.7	0.81	10.2	99.2	0.32
	2.0	5.3	53.2	0.24	5.7	57.3	0.04
	3.0	1.1	11.6	0.05	0.2	2.0	<0.01
Resistant (Dual)	1.5	9.0	83.7	0.81	10.2	99.2	0.32
	2.0	5.3	53.2	0.24	5.7	57.3	0.04
	3.0	1.1	11.6	0.05	0.2	2.0	<0.01
Resistant (Linear)	1.5	7.9	74.5	0.62	9.0	87.5	0.28
	2.0	4.3	43.5	0.13	4.4	44.1	0.03
	3.0	0.7	7.6	0.01	0.1	1.2	<0.01
Outlier test	0.05	6.3	50.8	1.48	5.9	49.6	1.06
	0.01	3.6	34.6	0.31	3.7	35.2	0.17
	$0.05/n$	1.1	12.3	0.02	0.2	2.3	<0.01

5 Application to some real data sets

In the following, the application of the methodology to some data sets will be illustrated, in order to focus on some points of particular interest.

5.1 D_2 data

The first data set, named here as the D_2 data set, is somewhat similar to the D_1 data, but of smaller size. It includes a sample of $n = 208$ observations about the same two variables, O_2 level and energy level. Here the assumption of constant variability is substantially tenable, and outlier labeling based on the procedure (2) with linear specification of the conditional quantiles provides results similar to those of the single-index approach. Two things are worth noting about this example. The first one is that the approach based on the nonparametric specification of quantiles requires a substantial amount of trial-and-error for data sets of this size, and actually in order to avoid too many wiggles in the resulting fences we had to set the smoothness parameter to a very large value (e.g. 800 for the results in Figure 3). Furthermore, this examples confirms once again that classical procedures based on least-squares are unsuitable for outlier labeling, even when the response variable is properly transformed. To this end, the right panel of Figure 3 display the normal quantile-quantile plot for the studentized residuals of the simple linear regression model where the response variable is given by the Yeo-Johnson transformation. It is apparent that the residuals from the linear regression are rather far from the theoretical $t(\nu)$ distribution with $\nu = 205$, which is very close to normality, suggesting that finding a normalizing transformation for the response is quite challenging for data of this type. Masking is substantial when the classical outlier test is adopted. Indeed, the test flags a number of potential outliers comparable with that labelled the resistant rules with $k = 2.0$ only when a significance level equal to 0.05 is chosen. Yet, such a significance level offers very little protection against swamping, as seen in the simulation study of §4.1.

5.2 D_3 data

The second data set includes data from a sample of size $n = 4998$, including data on the energy level and the CH_4 level, for two different process practices. The data are highly unbalanced, with only 23 points for one of the two practices. For such small group, the estimated coefficient of CH_4 in the linear quantile regression for the energy level is very close to zero for any value

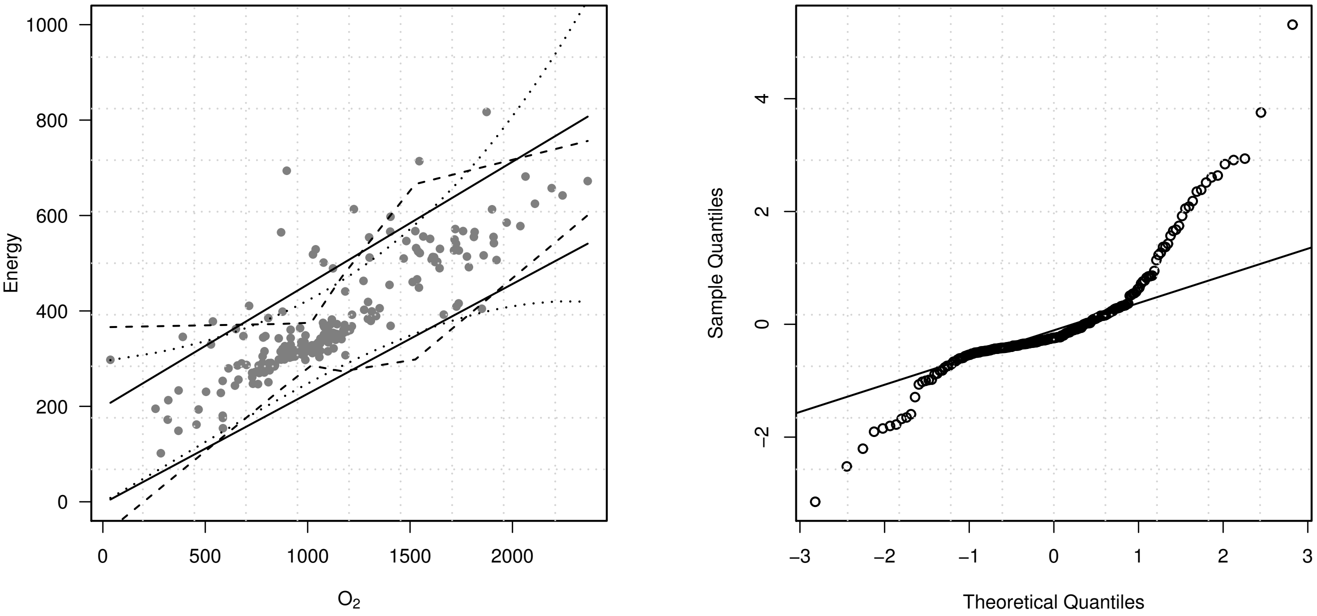

Figure 3: D_2 data set. Left panel: Plot of energy level against O_2 level. Fences based on quantile regression computed with $k = 2.0$ are superimposed, for linear quantiles (solid lines), nonparametric quantiles (dashed lines) and for the single-index method based on the Yeo-Johnson transformation (dotted). Right panel: normal quantile-quantile plot for studentized deletion residuals of the linear model for transformed response.

of τ, not attaining statistical significance at any reasonable level. In such case the single index method is not appropriate, as the transformation parameter $\lambda(\tau)$ is barely identifiable. The left panel of Figure 4 displays the results, showing that the fences from the single index method are totally overlapping with those of the linear specification for the conditional quantiles. In such case, the most sensible approach would be to monitor the energy level alone, without considering any regression model.

Things are rather different for the large portion of data corresponding to the other process practice. Here the conditional distribution of the energy level depends strongly on the CH_4 level, and outlier labeling based on quantile regression is meaningful. The resistant rule based on the linear specification of the conditional quantiles is not the best choice, as quantile crossing occurs, causing the fences to cross as well. The single-index method is more satisfactory, as the fences diverge rather than converge, resulting in a more conservative labeling of the observations with large values of CH_4. The resistant rule based on the nonparametric specification of the quantiles provides perhaps the best outcome, though once again a careful choice of the smoothness parameter is required. The nice-looking fences reported in the right panel of Figure 4 were obtained with the smoothness parameter for the `qss` function set to 200. Interesting enough, with a value of the smoothness parameter around 150 the resulting fences would be quite similar to those obtained with the single-index method.

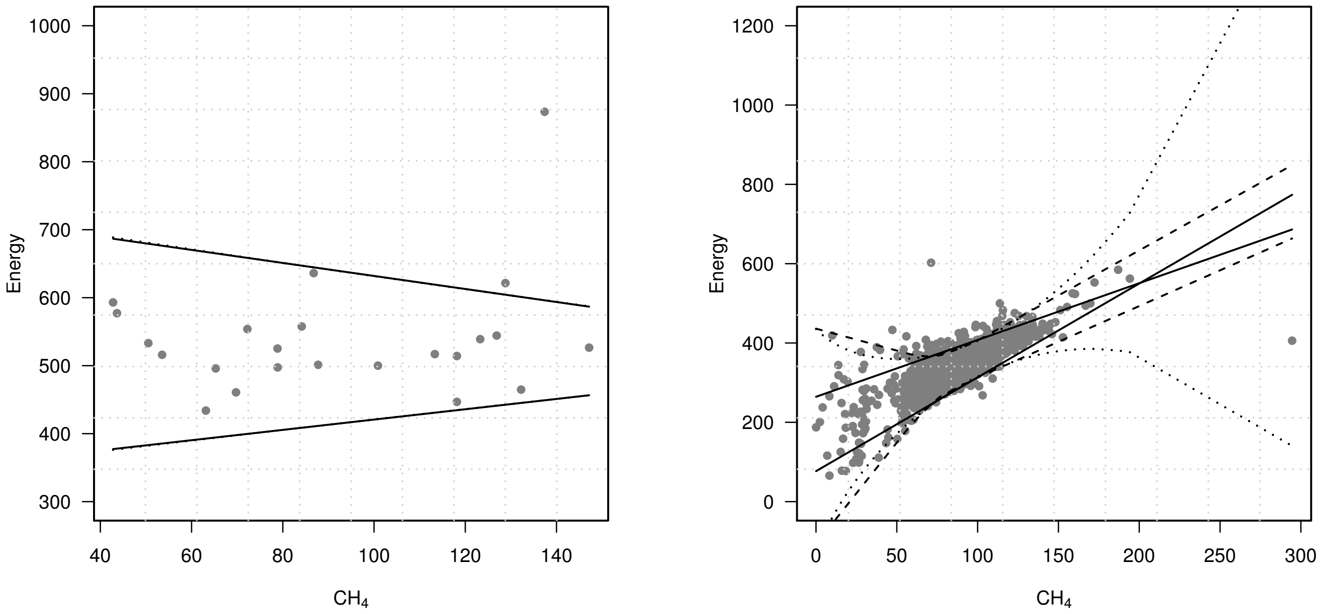

Figure 4: D_3 data set. Left panel: Plot of energy level against CH_4 level, for the first process practice. Fences based on quantile regression computed with $k = 1.5$ are superimposed, for linear quantiles (solid lines) and for the single-index method based on the Yeo-Johnson transformation (dotted). Right panel: Plot of energy level against CH_4 level, for the second process practice. Fences based on quantile regression computed with $k = 1.5$ are superimposed, for linear quantiles (solid lines), nonparametric quantiles (dashed lines) and for the single-index method based on the Yeo-Johnson transformation (dotted).

5.3 D_4 data

The last data set is similar to the previous one, but it includes data from four different process practices, and it will be used here to illustrate the possibility of extending outlier labeling to some multiple regression settings. The data include observations on a sample of size $n = 340$, with data on the energy level and the CH_4 level for four different process practices. The group sizes for the four different practices are 199, 2, 7 and 132 respectively. The sample size for Practice 1 and 4 are large enough to permit a separate analysis. For Practice 1 the results for two resistant rules are reported in the left panel of Figure 5, showing that in this case the single-index method seems to be slightly preferable over the linear specification. The sample sizes for Practice 2 and 3 are so small that we could ignore them, but it is actually possible to consider all the observations together by a multiple regression specification. Therefore, as a linear predictor for the i-th observation we take

$$x_i^T\beta(\tau) = \beta_0(\tau) + d_{i1}\beta_1(\tau) + d_{i2}\beta_2(\tau) + d_{i4}\beta_3(\tau) + z_i\beta_4(\tau) + z_i d_{i1}\beta_5(\tau) + z_i d_{i4}\beta_6(\tau) \qquad (12)$$

where $d_{ij},\ j = 2, 3, 4$ are the dummy variables employed for coding the process practice, and z_i the value of the CH_4 level. The results of the application of the resistant rule (1) with the single-index method and the linear method for the conditional quantiles are reported in the right panel of Figure 5. Though both methods flag a small subset of observations as outliers, the linear method flags also a few observations that are not flagged by the single-index method, similarly to what happens for the data of Practice 1 only. At any rate, for either method the multivariate extension (12) provides results similar to what obtained by considering the data of Practice 1 and 4 separately, suggesting that multivariate versions of the method may be effective.

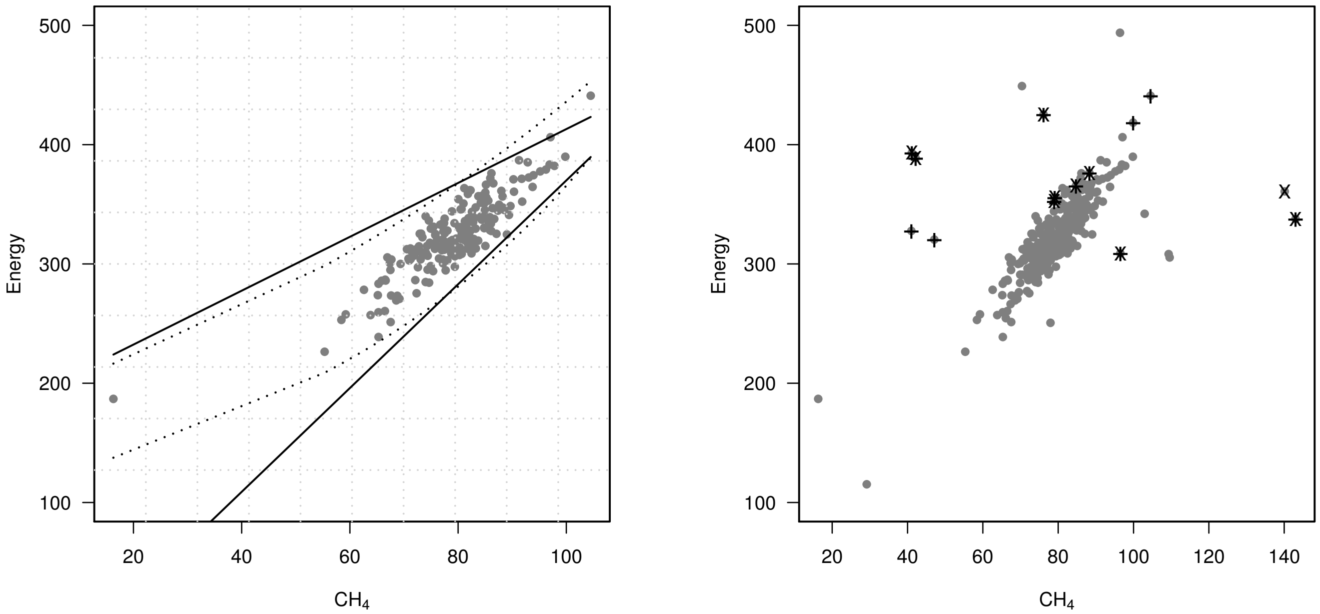

Figure 5: D_4 data set. Left panel: Plot of energy level against CH_4 level, for Practice 1. Fences based on quantile regression computed with $k = 1.5$ are superimposed, for linear quantiles (solid lines) and for the single-index method based on the Yeo-Johnson transformation (dotted). Right panel: Plot of energy level against CH_4 level, for all practices. Observations flagged as potential outliers by the resistant rule with $k = 1.5$ extended to multiple regression are highlighted, for the single index method ('x') and the linear specification of conditional quantiles ('+').

6 Discussion

The purpose of this paper is to show that the boxplot rule for outlier labeling can be extended to the regression setting within the framework of quantile regression, and that by introducing an additional transformation parameter some additional flexibility can be gained. Indeed, several results obtained in the literature for the univariate setting can be transferred to the regression setting. We mention in particular the fact that resistant rules based on quantile regression are generally well resistant to masking, especially with choices of the constant k such as $k = 1.5$ or $k = 2.0$. Good protection against swamping, as the sample-wise level, is instead provided by larger values of k, such as $k = 3.0$. Similarly to what noted by Hoaglin et al. (1986) for the univariate case, such choice is rather conservative, and it can be seen as a supplementary rule for joint usage with the basic resistant rules that have $k = 1.5$ or $k = 2.0$. A recommendable strategy is probably given by the application of the resistant rule with different values of k. We also notice that variations of the resistant rule, such as extension of the median rule by Carling (2000) or the rule based on semi-interquartile ranges proposed by Kimber (1990), provide performances in the simulation studies very similar to those of the basic method studied here, so that their usage was not investigated further.

In this paper we endorse an informal approach to outlier screening, for which the resistant rules extending the boxplot rule are indeed well suited. In our opinion, such informal approach is the most suitable to the ill-posed nature of the outlier problem, leaving the final decision about the nature of the flagged point to the expert of the data-generating process. However, some researchers may prefer a more formal way of operating. In such case, theory of multiple testing might be adapted, within a parametric or a semiparametric setting, and it seems likely that the resulting methodology could provide a better control of swamping than the resistant rules employed here.

The methodology developed here has some potential for being useful in many settings, but it has been motivated by the analysis of steelmaking process data. We have focused on energy consumption for the sake of clarity, but the methodology could be used for other process variables as well. At any rate, an appealing feature of the methodology is its simplicity, as the method is entirely based on linear quantile regression, iterated over a grid of points for the transformation parameter. Implementation of linear quantile regression is possible by a simple linear programming software, so that the methodology can be implemented not only within statistical software, but also within commonly used spreadsheets that are able to solve linear programming problems. In many applied settings, this makes the proposed methods preferable to more sophisticated nonlinear quantile regression methods, including nonparametric versions.

Two things seem worth stressing about the single-index method proposed here. The first one is that we could not find any substantial difference among the various transformations employed, so that our preference is for the Yeo-Johnson and the dual power transformations over the Box-Cox transformation, for the theoretical reasons given in Section 3. The second one, and perhaps more important, is that although the linear specification of quantiles may be occasionally afflicted by quantile crossing, which is aesthetically unpleasant at best, the quantitative consequences of such problem are typically limited, and indeed in the simulation experiments the resistant rule with linear quantiles provides quite acceptable performances. All in all, though the single-index method seems preferable, the linear specification of the conditional quantiles can still provide useful results, at the lowest computational price.

This article focuses mainly on the bivariate setting, with a single explanatory variable. The multivariate setting has been only touched in passing, and it would require a more thorough analysis. For the setting of interest here, visualization of the fences was deemed as an important feature of the method. Such visualization becomes impossible, or at best quite unpractical, with more than one covariate, therefore some alternative approaches would be required. An investigation along the lines of Eo et al. (2014), who proposed a specific score for the identification of outliers in multivariate regression settings, could be of some interest.

Acknowledgments

We are grateful to `beanTech s.r.l.` for supporting this work with a grant, and for providing the data for research purposes.

References

Aggarwal, C. C. (2013). *Outlier Analysis.* Springer, New York.

Atkinson, A. A. C., & Riani, M. (2000). *Robust Diagnostic Regression Analysis.* Springer, New York.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data (3rd Edition).* Wiley, New York.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics.* Wiley, New York.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289–300.

Bondell, H. D., Reich, B. J., & Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika*, *97*, 825–838.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, *26*, 211–252.

Brant, R. (1990). Comparing classical and resistant outlier rules. *Journal of the American Statistical Association*, *85*, 1083–1090.

Carling, K. (2000). Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis*, *33*, 249–258.

Carter, N. J., Schwertman, N. C., & Kiser, T. L. (2009). A comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry. *Statistical Methodology*, *6*, 604–621.

Cerioli, A., & Farcomeni, A. (2011). Error rates for multivariate outlier detection. *Computational Statistics & Data Analysis*, *55*, 544–553.

Chamberlain, G. (1994). Quantile regression, censoring, and the structure of wages. In *Advances in Econometrics: Sixth World Congress* (Vol. 2, pp. 171–209). Econometric Society Monograph. Cambridge University Press, Cambridge.

Cho, H., Kim, Y.-j., Jung, H. J., Lee, S.-W., & Lee, J. W. (2008). OutlierD: an R package for outlier detection using quantile regression on mass spectrometry data. *Bioinformatics*, *24*, 882–884.

Cook, R. D., & Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

Eo, S.-H., Hong, S.-M., & Cho, H. (2014). Identification of outlying observations with quantile regression for censored data. *arXiv preprint:1404.7710*.

Fitzenberger, B., Wilke, R. A., & Zhang, X. (2009). Implementing Box–Cox quantile regression. *Econometric Reviews*, *29*, 158–181.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression (2nd edition).* Sage, Thousand Oaks, CA.

Han, J., Kamber, M., & Pei, J. (2006). *Data Mining: Concepts and Techniques.* Morgan Kaufmann, Waltham MA.

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *81*, 991–999.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*, 85–126.

Huber, P. J., & Ronchetti, E. (2009). *Robust Statistics (2nd edition).* Wiley, New York.

Kano, M., & Nakagawa, Y. (2008). Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry. *Computers & Chemical Engineering*, *32*, 12–24.

Kimber, A. (1990). Exploratory data analysis for possibly censored data from skewed distributions. *Applied Statistics*, *39*, 21–30.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Koenker, R. (2013). quantreg: Quantile regression [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=quantreg` (R package version 5.05)

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics.* Wiley, New York.

Powell, J. L. (1991). Estimation of monotonic regression models under quantile restrictions. In W. Barnett, J. Powell, & G. Tauchen (Eds.), *Nonparametric and Semiparametric Methods in Econometrics* (pp. 357–384). Cambridge University Press, New York.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rosner, B. (1983). Percentage points for a generalized esd many-outlier procedure. *Technometrics*, *25*, 165–172.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, *53*, 382–387.

Schwertman, N. C., & de Silva, R. (2007). Identifying outliers with sequential fences. *Computational statistics & data analysis*, *51*, 3800–3810.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, Mass.

Turkdogan, E. T. (1996). *Fundamentals of Steelmaking.* The Institute of Materials, London.

Yang, Z. (2006). A modified family of power transformations. *Economics Letters*, *92*, 14–19.

Yee, T. W. (2004). Quantile regression via vector generalized additive models. *Statistics in Medicine*, *23*, 2295–2315.

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*, 954–959.