Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 49 (2005) 875–891
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Robust regression diagnostics with data transformations

Tsung-Chi Cheng*

Department of Statistics, National Chengchi University, 64 Chih-Nan Road, Section 2, Taipei 11623, Taiwan

Received 8 March 2004; received in revised form 12 June 2004; accepted 14 June 2004
Available online 6 July 2004

Abstract

The problems of non-normality or functional relationships between variables may often be simplified by an appropriate transformation. However, the evidence for transformations may sometimes depend crucially on one or a few observations. Therefore, the purpose of the paper is to develop a method that will not be influenced by potential outliers during the process of data transformations. The concepts of the least trimmed squares estimator and the trimmed likelihood estimator are used to obtain the robust transformation parameters. Furthermore, the proposed procedure unifies robust statistics and a diagnostic approach to deal with the outlier problem in the regression transformation.

Keywords: Box–Cox transformation; High breakdown estimator; Least trimmed squares estimator; Robust diagnostics; Trimmed likelihood

1. Introduction

The assumption of normality customarily provides a powerful and convenient way to analyze a linear regression problem. The problem of non-normality may often be simplified by an appropriate transformation, such as the parametric family of power transformations in Box and Cox (1964). In addition, relationships between variables can be explored or simplified by means of data transformation. However, the evidence for transformations may sometimes depend crucially on one or a few observations. Several authors have pointed out that data transformations are very sensitive to outliers. (For more information on data

* Tel.:+886229393091X81132; fax: +886229398024.
E-mail address: chengt@nccu.edu.tw (T.-C. Cheng).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.06.010

transformations, see Sakia, 1992.) The purpose of this paper is to develop a method of data transformation that will not be influenced by potential outliers.

There are two ways to deal with outlier problems in regression analysis, diagnostic approaches and robust estimators (Rousseeuw and Leroy, 1987, p. 8). Diagnostic approaches for assessing the contribution of individuals to the evidence for a transformation have been suggested by Atkinson (1985), Cook and Wang (1983), and Lawrance (1988). Tsai and Wu (1990) first take into account the deletion effect of a single observation on the Jacobian of data transformation. Kim et al. (1996) extend the Jacobian effect to multiple deletion diagnostics on Box–Cox transformations. Some disadvantages do limit the use of multiple deletion diagnostics, which include combinatorial problems, the size of the deleted subsets in practice, and lack of devices to present the results for a large sample size.

In an effort to robustify the analysis of the regression transformation, Carroll and Ruppert (1985, 1988) adapt a bounded influence estimator of Krasker and Welsch (1982), in which an estimating equation is used to allow for the effect on the estimate of the transformation parameter by the leverage points. Parker (1988) considers the L_1 estimator for the regression transformation. However, one of the shortcomings of these estimators is that they can have a low breakdown point. The (finite) sample breakdown point of an estimator is the smallest proportion of observations which when altered can cause the value of the estimator to be arbitrarily large or small. One of the desirable properties for a robust estimator is one with a high breakdown point, which is capable of handling multiple outliers. In this article, we are particular interested in developing a robust transformation in the use of high breakdown estimators.

The first high breakdown estimator, the least median of squares (LMS), was proposed by Rousseeuw (1984). Since then, robust diagnostics has been developed to solve the problem of outliers in a systematic way (Rousseeuw and Van Zomeren, 1990). Atkinson (1986a) first employs the concept of the "constructed variable" (see Atkinson (1985) for details) and shows the effect of the deletion of observations on the score statistic for power transformation. He therefore suggests a two-stage procedure to avoid the masking effect on data transformation. The first stage is an exploratory method for the identification of outliers, in which a robust analysis using LMS is performed on a series of values of transformation parameters. The second stage uses multiple-deletion diagnostic methods to serve as a confirmatory method. Fung (1993) also proposes a stepwise procedure using robust methods to confirm the outliers and leverage points without considering transformation. A similar idea is suggested in Hadi and Simonoff (1993) for the identification and test of multiple outliers for linear models.

Atkinson and Riani (2000) adapt the forward search algorithm (Atkinson, 1994) by including the score tests for the Box–Cox transformation. By giving a fixed value of the transformation parameter, they show that the forward search algorithm provides high breakdown estimates of regression parameters and monitors the effect of individual observations on the transformed data. LMS is used again as the criterion to assess the performance of the search. In terms of the "forward" search, the deletion diagnostics of Tsai and Wu (1990) and Kim et al. (1996) can be viewed as a "backward" approach.

An alternative to LMS is the least trimmed squares (LTS) estimator, which has better theoretical properties than the LMS. The details of LTS will be given in the later section. Hadi and Luceño (1997) propose the trimmed likelihood estimator, which is based on trimming

the likelihood function rather than directly trimming the data. In this article, we first connect LTS and the idea of the trimmed likelihood approach for the problem of data transformation. The robust framework is then proposed to provide a unified approach, together with the concepts of exploratory and confirmatory methods. A resampling procedure is implemented for this purpose and in order to obtain a high breakdown estimate of the transformation parameters.

The outline of this article is as follows. Section 2 shows the diagnostic approach on Box–Cox transformation for the linear regression model. Section 3 first gives a brief description of high breakdown estimators. A robust transformation is then proposed and a computing algorithm is also provided to obtain the result in Section 4. Section 5 carries out a simulation study to show the performance of the resulting approach. Some real data analyses are illustrated in Section 6. Section 7 draws some conclusions and comments.

2. Diagnostics on regression transformation

Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})$ denotes $p \times 1$ regression coefficients, $\boldsymbol{X} = (\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_n^{\mathrm{T}})$ is an $n \times p$ design matrix, and $\boldsymbol{\epsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ represents the error term. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_{p-1})$ be any estimate of parameter $\boldsymbol{\beta}$. The residuals from this estimate are $e_i(\hat{\boldsymbol{\beta}}) = y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, $i = 1, \ldots, n$. The most popular regression estimator is the least-squares estimate (LSE), which corresponds to

$$\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} e_i^2, \tag{2}$$

where $e_i = e_i(\hat{\boldsymbol{\beta}})$. We often make a certain idealized assumption about the error term, where ε_i is usually assumed to be independent and identically distributed with a normal distribution, $N(0, \sigma^2)$, for the purpose of statistical inferences. Under the assumption of normality, the LSE is the same as the maximum likelihood estimator (MLE).

When the assumption of normality for model (1) does not exist, Box and Cox (1964) consider the following transformation:

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

If the transformed observations $\boldsymbol{y}(\lambda)$ are normally distributed with mean $\boldsymbol{X}\boldsymbol{\beta}(\lambda)$ and common variance $\sigma^2(\lambda)$, for the ith observation, then the log-likelihood function based on the above transformation is

$$\ell_i = \ell(\boldsymbol{\theta}; y_i) \doteq -\frac{1}{2}\log\sigma^2(\lambda) - \frac{(y_i(\lambda) - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(\lambda))^2}{2\sigma^2(\lambda)} + (\lambda - 1)\log y_i, \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \lambda)$. The ordinary MLE of λ, $\hat{\lambda}$, hence maximizes $\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, y_i)$.

Although MLE has good statistical properties, the estimate $\hat{\lambda}$ is very sensitive to outliers. To identify those cases that influence $\hat{\lambda}$, the deletion diagnostic approach compares $\hat{\lambda}$ and $\hat{\lambda}_{(\mathscr{M})}$, where $\hat{\lambda}_{(\mathscr{M})}$ is the MLE of λ based on $n-m$ observations after deleting m observations. The set of those deleted m cases is indicated by $\mathscr{M}$. This process is computationally intensive, because the Jacobian differs for each set $\mathscr{M}$. Kim et al. (1996) extend the case-deletion model approach of Tsai and Wu (1990) to multiple deletion diagnostics on a Box–Cox transformation. If J denotes the Jacobian of the transformation from $\boldsymbol{y}$ to $\boldsymbol{y}(\lambda)$, then Kim et al. (1996) show that an approximation to $\hat{\lambda}_{(\mathscr{M})}$ of Tsai and Wu (1990), based on the $n-m$ from n cases, minimizes

$$Q_{(\mathscr{M})}(\lambda) = \tilde{\boldsymbol{z}}(\lambda)^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_E)\tilde{\boldsymbol{z}}(\lambda)\left(\prod_{i\in\mathscr{M}} y_i\right)^{2(\lambda-1)/(n-m)}, \tag{5}$$

where $\tilde{\boldsymbol{z}}(\lambda) = \boldsymbol{y}(\lambda)/J^{1/(n-m)}$ and $\boldsymbol{H}_E = \boldsymbol{X}_E(\boldsymbol{X}_E^{\mathrm{T}}\boldsymbol{X}_E)^{-1}\boldsymbol{X}_E^{\mathrm{T}}$. Here, $\boldsymbol{X}_E = (\boldsymbol{X}, \boldsymbol{E}_{\mathscr{M}})$, and $\boldsymbol{E}_{\mathscr{M}}$ is an $n \times m$ matrix containing a 1 in the position of the row and column which correspond to the set $\mathscr{M}$ and 0's elsewhere.

Kim et al. (1996) derive the one-step estimator to approximate $\hat{\lambda}_{(\mathscr{M})}$ as follows. They first define

$$\boldsymbol{w}^{\mathrm{T}}(\lambda) = \frac{\partial \boldsymbol{z}(\lambda)}{\partial\lambda},$$
$$\boldsymbol{u}^{\mathrm{T}}(\lambda) = \frac{\partial \boldsymbol{w}(\lambda)}{\partial\lambda},$$

where $\boldsymbol{z}(\lambda) = \boldsymbol{y}(\lambda)/J^{1/n}$. Now let $\boldsymbol{H}_{\mathscr{M}}$ be the $m\times m$ submatrix of $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ indexed by $\mathscr{M}$. In addition, $\boldsymbol{r}_{\boldsymbol{z},\mathscr{M}}$, $\boldsymbol{r}_{\boldsymbol{w},\mathscr{M}}$, and $\boldsymbol{r}_{\boldsymbol{u},\mathscr{M}}$ indicate $m \times 1$ subvectors of

$$\boldsymbol{r}_{\boldsymbol{z}} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{z}(\hat{\lambda}),$$
$$\boldsymbol{r}_{\boldsymbol{w}} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{w}(\hat{\lambda}),$$
$$\boldsymbol{r}_{\boldsymbol{u}} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{u}(\hat{\lambda}),$$

respectively, and let

$$S_{\boldsymbol{pq},\mathscr{M}} = \boldsymbol{r}_{\boldsymbol{p},\mathscr{M}}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_{\mathscr{M}})^{-1}\boldsymbol{r}_{\boldsymbol{q},\mathscr{M}},$$

where $\boldsymbol{p} = \boldsymbol{z}, \boldsymbol{w}$ and $\boldsymbol{q} = \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{u}$. Finally, define

$$G_{\mathscr{M}} = \log\left(\prod_{i\in\mathscr{M}} y_i \Big/ \dot{y}^m\right)\Big/(n-m),$$

where $\dot{y}$ is the geometric mean of the y_i's. To assess the effect of the m cases being deleted from the data set, the closed form of the diagnostic is then defined as

$$\hat{\lambda}^{TW}_{(\mathscr{M})} = \hat{\lambda} - \left[2G_{\mathscr{M}} + \frac{\boldsymbol{r}_{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{r}_{\boldsymbol{w}} + \boldsymbol{r}_{\boldsymbol{z}}^{\mathrm{T}}\boldsymbol{r}_{\boldsymbol{u}} - S_{\boldsymbol{ww},\mathscr{M}} - S_{\boldsymbol{zu},\mathscr{M}} - 2G_{\mathscr{M}}S_{\boldsymbol{zw},\mathscr{M}}}{G_{\mathscr{M}}(\boldsymbol{r}_{\boldsymbol{z}}^{\mathrm{T}}\boldsymbol{r}_{\boldsymbol{z}} - S_{\boldsymbol{zz},\mathscr{M}}) - S_{\boldsymbol{zw},\mathscr{M}}}\right]^{-1}. \tag{6}$$

3. Robust regression

In this section, we first summarize some issues about robust regression, which will be used in the later discussion.

3.1. High breakdown estimators

The first high breakdown estimator, the least median of squares (LMS), was proposed by Rousseeuw (1984). Let

$$e^2_{(1),n} \leqslant e^2_{(2),n} \leqslant \cdots \leqslant e^2_{(n),n} \tag{7}$$

be the ordering of the residuals e_i^2, $i = 1, \ldots, n$. LMS is defined by

$$\min e^2_{(\mathrm{med}),\mathrm{n}},$$

where med $= [(n + p + 1)/2]$, and $[\cdot]$ indicates the integer part. However, the LMS estimator converges at the low rate of $n^{-1/3}$ to a non-normal distribution. Its asymptotic efficiency approaches 0 as the sample size goes to infinity (Rousseeuw and Leroy, 1987, Section 4.4.).

Instead of adding all the squared residuals as in (2), one can limit one's attention to a "trimmed" sum of squares. If only the first q of those ordered residuals are included in the summation, then the least trimmed squares (LTS) estimator is defined as

$$\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{q} e^2_{(i),n}. \tag{8}$$

For $q = [n/2] + [(p + 1)/2]$, the LTS reaches the maximal possible value for the breakdown point $([(n - p)/2] + 1)/n$ (Rousseeuw and Leroy, 1987, p. 132), which is the same as that of the LMS estimate minimizing the "median" residual $e^2_{(q),n}$ for the same q. The LTS estimator converges to a normal distribution at the rate of $n^{-1/2}$.

The use of LTS in the application of robust regression has become more feasible and popular after the fast algorithm to find the LTS solution proposed by Rousseeuw and Van Driessen (1999). They show that after starting any approximation to the LTS estimate, it is possible to obtain another approximation yielding an even lower objective function (8). They call this a *C-step*, where *C* stands for "concentration". The resulting algorithm can quickly obtain the LTS solution. This procedure is available in S-PLUS.

Atkinson and Cheng (1999) apply the forward search algorithm to find the LTS and also discuss the choice of q. They show that one can achieve more stable results for the detection of outliers as well as highly efficient estimates when more data are fitted, provided that q is small enough to exclude outlying cases. Zaman et al. (2001) suggest that $[0.75n]$ is a reasonable value for q in most empirical studies.

3.2. Trimmed likelihood estimator

Let X be a random variable with a probability density $f(x; \theta)$ which depends on an unknown parameter θ. Let $x_1, \ldots, x_n$ be n independent realizations of X. The MLE of θ is

the value of θ that maximizes the logarithm of the likelihood function

$$L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ell(\theta; x_i), \tag{9}$$

where $\ell(\theta; x_i) = \ln f(x_i; \theta)$ is the contribution of the ith observation to the log likelihood function (9). The ML method has desirable properties. However, it depends on aggregate statistics, and it is sensitive to outlier and/or to violations of distributional assumptions.

Hadi and Luceño (1997) propose a trimmed likelihood principle based on trimming the likelihood function rather than directly trimming the data. They show that this trimming likelihood principle produces many existing estimators, such as MLE, LMS, and LTS. It is always possible to order and trim observations according to their contributions to the likelihood function, because the likelihood is scalar-valued. For any given value of θ, the likelihood ordering is

$$\ell(\theta; x_{(1)}) \geqslant l(\theta; x_{(2)}) \geqslant \cdots \geqslant l(\theta; x_{(n)}).$$

The method proposed by Hadi and Luceño (1997) replaces the log-likelihood function by the trimmed log-likelihood function

$$\sum_{i=a}^{b} w_i l(\theta; x_{(i)}), \tag{10}$$

where $a \leqslant b$, $(a, b) \in \{1, 2, \ldots, n\}$, and $w_i \geqslant 0$ are weights. The estimator $\theta(a, b, w)$ is obtained by maximizing (10). They refer to this method as the *maximum trimmed likelihood* (MTL) method and to $\hat{\theta}(a, b, w)$ as the maximum trimmed likelihood estimators (MTLE).

Consider the case $w_i = 1$, $a \leqslant i \leqslant b$. When $a = 1$ and $b = n$, $\hat{\theta}(1, n)$ is the MLE of θ, so that MLE is a special case of MTLE. When $a = b = [(n + 1)/2]$, the resulting estimator is referred to as the maximum median likelihood estimator (MMLE). When $a = 1$ and $b < n$ and data are Gaussian, then $\hat{\theta}(1, b)$ is the LTS of the location parameter θ. Hadi and Luceño (1997) also show that the MMLE of $\boldsymbol{\beta}$ and σ^2 for model (1) are the same as the LMS estimates of $\boldsymbol{\beta}$ and σ^2, respectively.

4. Robust transformations

For the robust estimation of transformation (3), Carroll and Ruppert (1987) propose an M-type estimator $\tilde{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$, which is the solution to

$$\sum_{i=1}^{n} w_i(y_i, \tilde{\boldsymbol{\theta}}) s_i(y_i, \tilde{\boldsymbol{\theta}}) = 0,$$

where $s_i(\cdot)$ is the score function of the log-likelihood function (4) for y_i, and w_i is a suitable scalar function. They also extend this idea to "transform both sides" of the regression model. This estimator is a kind of the robust bounded-influence estimators of Krasker and Welsch (1982).

For an inference about λ, it is easier to work with the normalized Box–Cox transformation

$$\boldsymbol{z}(\lambda) = \begin{cases} (\boldsymbol{y}^{\lambda} - 1)/\lambda \dot{y}^{\lambda-1}, & \lambda \neq 0, \\ \dot{y} \log \boldsymbol{y}, & \lambda = 0, \end{cases} \tag{11}$$

where $\dot{y} = \exp(\sum \log(y_i/n))$. The analogy of (4) for the ith observation is

$$\ell_i = \ell(\boldsymbol{\theta}; z_i) \doteq -\frac{1}{2} \log \sigma^2(\lambda) - \frac{(z_i(\lambda) - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(\lambda))^2}{2\sigma^2(\lambda)}. \tag{12}$$

For a particular value of λ, the MLE of $\boldsymbol{\beta}$ is equivalent to maximizing

$$L(\cdot) = -\frac{n}{2} \log \hat{\sigma}^2(\lambda).$$

Parker (1988) adapts Laplace errors for the transformation parameter, in which the log-likelihood function becomes

$$L_1(\boldsymbol{\theta}) = -n \log(2\sigma(\lambda)) - \sum_{i=1}^{n} |z_i(\lambda) - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(\lambda)|/\sigma(\lambda).$$

However, it is known that the bounded-influence estimator is not robust from the view of the breakdown point, and the L_1 estimate is not able to resist the leverage points. Some adaptive estimators to these aspects for model (1) without consideration of transformation have been proposed (e.g. Simpson et al., 1992; Coakley and Hettmansperger, 1993; Chave and Thomson, 2003). However, this direction is not covered in the scope of this paper.

Atkinson and Riani (2000) adapt the forward search algorithm using LMS for the problem of data transformation. The forward search is a powerful general method, which involves successively augmenting the subset of data until all data are included on the fit. During the process of data increment, the unidentified subsets of the data are detected and their effect on fitted models can be evaluated. The fan plot provides a forward plot of the score test statistic for a series of values of λ. It is used to present the evolution of the score statistics during the forward process. This procedure can be viewed as a unified result of Atkinson's previous works (Atkinson, 1985, 1986a). However, this approach does not provide an explicit estimate of λ.

In the following discussion, we unite the diagnostic and robust approaches to obtain the robust estimates of transformed data.

4.1. LTS and transformation

The relationships of LTS and MTLE are discussed by Müller and Neykov (2003) for a generalized linear model. In this subsection we show the LTS and MTLE solutions to the transformation parameter.

Assume that $\mathcal{Q}$ is the set including those q observations with the largest values of (12) when a particular value of λ is given. Therefore, the corresponding maximized trimmed log-likelihood function of (10) for model (1) under transformation (11) is

$$L_q(\hat{\boldsymbol{\beta}}_q) = \sum_{i \in \mathcal{Q}} \ell(\hat{\boldsymbol{\beta}}_q; z_{(i)}) \doteq -\frac{q}{2} \log \hat{\sigma}_q^2(\lambda), \tag{13}$$

where $\ell(\hat{\boldsymbol{\beta}}_q; z_{(i)})$ is the ith ordering likelihood, and $\hat{\boldsymbol{\beta}}_q = \hat{\boldsymbol{\beta}}_q(\lambda) = \hat{\boldsymbol{\beta}}(1, q)$ denotes MTLE of $\boldsymbol{\beta}$ at the value of λ. The MTLE of σ^2 at the value of λ is then

$$\hat{\sigma}_q^2(\lambda) = \sum_{i \in \mathcal{Q}} e_i^2(\lambda)/(q - p), \tag{14}$$

where the residuals are defined as usual for the transformed data, which are

$$e_i(\lambda) = e_i(\hat{\boldsymbol{\beta}}_q) = z_i(\lambda) - \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_q, \quad i = 1, \ldots, n.$$

The observations in $\mathcal{Q}$ of (14) are those cases with the smallest q residuals, which correspond to those with the largest values of likelihood for (13). Once the set $\mathcal{Q}$ is attained, the MTLE of λ at the value of q, denoted by $\hat{\lambda}_q$, can be computed by the MLE of λ based on the set $\mathcal{Q}$. Therefore, $\hat{\boldsymbol{\theta}}_q = (\hat{\boldsymbol{\beta}}_q, \hat{\sigma}_q^2, \hat{\lambda}_q)$ are the LTS estimates and also the MTLE of $\boldsymbol{\theta}$.

The LTS or MTLE of $\boldsymbol{\theta}$ is carried out by a two-stage estimation procedure, which is similar to the solution of MLE described in Atkinson (1985, pp. 86–87). Note that for a given λ, the Jacobian of the transformed variables on the trimmed cases is independent of $\boldsymbol{\beta}$, which can be seen from (5). Therefore, maximizing (13) is equivalent to minimizing (14).

4.2. Numerical computing

For a specific value of q, an approximate solution of $\hat{\boldsymbol{\theta}}_q$ can be obtained by subsampling from the data in the following way:

- *RT1*: Give an estimate of λ to transformation (11), and choose a random subsample with s cases, say $s = p + 1$, from the transformed data.
- *RT2*: Apply the *C*-step of Rousseeuw and Van Driessen (1999) to obtain a subset with q cases, which is denoted by $\mathcal{Q}$.
- *RT3*: Apply result (6) to find the estimate of λ, $\hat{\lambda}_{(\bar{\mathcal{Q}})}^{TW}$, where $\bar{\mathcal{Q}}$ denotes the complementary set of $\mathcal{Q}$.
- *RT4*: Obtain the LSE of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_q$, based on the set $\mathcal{Q}$ after a transformation using $\hat{\lambda}_{(\bar{\mathcal{Q}})}^{TW}$, and calculate the objective function (14).

Steps RT1 to RT4 denote one subsampling scheme. The subsampling scheme is repeated, and the estimate of the transformation parameter is updated to RT1 when a lower value of (14) is reached.

The details of the procedure are described as follows. For the first step, the initial value of λ is MLE of λ from the whole data set. This value is replaced by $\hat{\lambda}_q$ in the subsequent subsampling. The working data are the transformed data in the following steps. If $\mathcal{S}$ indicates the set of randomly selected s cases, then we compute LS regression coefficients, denoted by $\hat{\boldsymbol{\beta}}_{\mathcal{S}}(\hat{\lambda}_q)$, based on $\mathcal{S}$. The ordered residuals for each observation are defined as

$$e_{(1),\mathcal{S}} \leqslant e_{(2),\mathcal{S}} \leqslant \cdots \leqslant e_{(n),\mathcal{S}}, \tag{15}$$

where

$$e_{i,\mathscr{S}} = e_i(\hat{\boldsymbol{\theta}}_{q,\mathscr{S}}) = z_i(\hat{\lambda}_q) - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathscr{S}}(\hat{\lambda}_q), \quad i = 1, \ldots, n.$$

Applying the *C*-step procedure to the transformed data during step RT2, we can obtain a new subset with q observations which fulfills the necessary condition for a global minimum of the LTS objective function. If $\mathscr{Q}$ indicates the set of q cases with the smallest residuals of (15), then the estimate $\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW}$ is computed by using (6) at step RT3. This step intends to obtain the estimate of the transformation parameter excluding those potential outliers in $\bar{\mathscr{Q}}$.

For the last step, applying the estimate $\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW}$ to (11), we now work on the new transformed data set. The estimated LS regression coefficient $\hat{\boldsymbol{\beta}}_{\mathscr{Q}}(\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW})$ is attained based on the set $\mathscr{Q}$. Therefore, the corresponding result (14) is computed by

$$s_q^2(\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW}) = \sum_{i \in Q} \left(e_i(\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW})\right)^2/(q-p), \tag{16}$$

where $e_i(\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW}) = z_i(\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW}) - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathscr{Q}}(\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW})$, and $i \in \mathscr{Q}$.

Once the smaller value of the objective function (16) is attained, $\hat{\lambda}_q$ is replaced by the new estimate $\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW}$ for the new selected subsample $\mathscr{S}$ of step RT1. Steps RT2 to RT4 are then repeated, and $\hat{\lambda}_q$ is updated once a smaller value of (16) is reached. The resampling scheme yields a series of values of (16), with the value defining the performance of the chosen subset. Therefore, step RT2 is used to identify the possible outliers under transformation, and RT3 gives the refined estimate of the transformation parameters.

Note that results (15) are used to carry out the *C*-step and achieve the deletion subset for computing the estimate, $\hat{\lambda}_{(\bar{\mathscr{Q}})}^{TW}$. The minimum value of (14) from all chosen subsets indicates the approximate solution of the LTS, $\hat{\boldsymbol{\theta}}_q$. When the sample size is not large, an exhaustive search is used; otherwise, several subsets are randomly drawn. Finally, for a general, but specified q, the criterion for the trimmed likelihood estimates in this resampling procedure is the same as (16) as we discussed in the previous subsection.

5. Simulation study

To see the capability of the proposed procedure in the previous section, we conduct a simulation for model (1). The data are generated in a similar manner to that of the famous Rousseeuw's data (Rousseeuw, 1984). For "good" data, each regressor, $\boldsymbol{x}_i^{\mathrm{T}}$, is generated from the uniform distribution $U(0, 6)$ and the corresponding error term $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 0.04^2)$, whereas "bad" data points are generated in the following form:

$$\begin{pmatrix} y_i \\ \boldsymbol{x}_i^{\mathrm{T}} \end{pmatrix} \sim MN\left(\begin{pmatrix} 4 \\ (14+p)\boldsymbol{J} \end{pmatrix}, \begin{pmatrix} 0.5 & \boldsymbol{0} \\ \boldsymbol{0} & 0.5\boldsymbol{I}_{p-1} \end{pmatrix}\right),$$

where $\boldsymbol{J}$ is a unit vector with dimension $(p-1)$.

Table 1
The simulation results for $p = 3$

Sample size n	Proportion of outliers (%)	λ			
		-1	-0.5	0	0.5
50	5	−0.9777	−0.4896	−0.0005	0.4881
		(0.0267)	(0.0155)	(0.0230)	(0.0193)
	10	−0.9771	−0.4870	−0.0040	0.4840
		(0.0419)	(0.0194)	(0.0256)	(0.0214)
	15	−0.9558	−0.4777	−0.0068	0.4759
		(0.0588)	(0.0245)	(0.0406)	(0.0311)
	20	−0.9415	−0.4714	−0.0155	0.4589
		(0.0696)	(0.0315)	(0.0521)	(0.0368)
100	5	−0.9816	−0.4898	−0.0027	0.4904
		(0.0195)	(0.0109)	(0.0329)	(0.0125)
	10	−0.9771	−0.4893	−0.0067	0.4818
		(0.0222)	(0.0146)	(0.0513)	(0.0134)
	15	−0.9746	−0.4825	−0.0049	0.4799
		(0.0284)	(0.0197)	(0.0645)	(0.0194)
	20	−0.9617	−0.4765	−0.0078	0.4700
		(0.0432)	(0.0267)	(0.0060)	(0.0285)

The simulation design considers the effects of dimensions of the data, sample size, and proportion of outliers in data. The dimensions, p, are 3 and 5. Two sample sizes are considered, $n = 50$ and 100. Five percent, 10%, 15%, or 20% of the observations are outlying in each simulated data set. All regression coefficients are assigned a value of 1. Once the data have been generated, a transformation parameter is given to the response variable, for which the inverse of the Box–Cox transformation (3) is applied to $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Here, we consider the values of λ to be -1, -0.5, 0, and 0.5.

To compare the performance of the robust procedure, 200 data sets are generated to implement the resampling algorithm, in which 100 subsampling schemes for each data set are carried out to obtain the optimum. The default value of q is set to be $[0.75n]$, whereas $q = [0.7n]$ is used when the outlier proportion is 20%. Tables 1 and 2 show the mean and standard deviation (in the parentheses) of the robust estimates of λ from 200 replications for dimensions 3 and 5, respectively.

From these simulation results, we have the following findings. As the proportion of outliers increases, the estimates of λ's turn away from the true parameters, and the standard deviations of the estimates become inflated when the sample size and dimension are the same. While under the same dimension, the larger the sample size is, the smaller the value will be of the standard deviation. The dimension plays an important role in the behaviors of the robust estimators as well as in our study. When $p = 5$ and $n = 50$, the values of some cells in Table 2 (e.g. 10% outliers) are not as consistent as those in Table 1 in terms of the proportion of outliers and sample size. The ratio of the sample size and dimension becomes a very important factor when the proportion of outliers is relatively high, e.g. the cell of 20% of the outliers. However, this problem can be alleviated and improved if

Table 2
The simulation results for $p = 5$

Sample size n	Proportion of outliers (%)	λ			
		-1	-0.5	0	0.5
50	5	−0.9866	−0.4933	−0.0016	0.4926
		(0.0246)	(0.0117)	(0.0219)	(0.0161)
	10	−0.9832	−0.4930	−0.0013	0.4912
		(0.0357)	(0.0443)	(0.0395)	(0.0758)
	15	−0.9401	−0.4774	−0.0058	0.4669
		(0.0764)	(0.0283)	(0.0282)	(0.0914)
	20	−0.8927	−0.4729	−0.0772	0.4420
		(0.1028)	(0.0359)	(0.0720)	(0.0629)
100	5	−0.9922	−0.4959	−0.0026	0.4958
		(0.0162)	(0.0071)	(0.0264)	(0.0084)
	10	−0.9786	−0.4908	0.0004	0.4871
		(0.0249)	(0.0119)	(0.0433)	(0.0143)
	15	−0.9572	−0.4813	−0.0056	0.4738
		(0.0474)	(0.0196)	(0.0184)	(0.0741)
	20	−0.9146	−0.4690	−0.0615	0.4506
		(0.0661)	(0.0245)	(0.0672)	(0.0489)

more subsamples are drawn. These results are quite similar to common conclusions in the literature of robustness, in which the dimension, sample size, and proportion of outliers in the data are the main factors on the influence of performance of robust estimators. The number of subsamples is an issue, which is more related to computation rather than the data.

Another effect for the LTS and MTLE is the choice of q. Atkinson and Cheng (1999) conclude that the higher the value of q is, the higher the efficiency of LTS will be, and the more stable the identification of outliers is, provided that the value of q is not large enough to include the existing outliers. A similar result can be expected in the robust estimates of transformation parameters. Moreover, smaller values of q may be needed for some simulated data. For instance, we compare $[0.65n]$ and $[0.7n]$ for the values of q when 20% of the outliers exit in the data with $n = 50$ and $p = 5$. The range of the 200 estimates of λ is smaller for $q = [0.65n]$ than that of $q = [0.7n]$. For $q = [0.65n]$, it also produces closer estimates to the true parameter and a smaller value of standard deviation of the estimates. The mean (standard deviation) of the 200 estimates of λ are -0.8927 (0.1028) as reported in Table 2 for $q = [0.7n]$, whereas the result is improved to be -0.9246 (0.0834) for $q = [0.65n]$. The smaller value of q provides a higher breakdown to prevent a relatively higher proportion of outliers when keeping other conditions the same.

Tables 1 and 2 are used to compare how some factors influence the performance of the proposed procedure. The refined output of the simulation can be obtained if more subsamples are used to find the optima, and/or different values of q. Nevertheless, the conclusion of Atkinson and Cheng (1999) is still valid whereby the larger values of q can yield higher efficient estimates and the stability of identification of outliers. In the test of the simulation

design, however, we experience that the choice of the value of q is smaller than the expected one for the transformation problem, especially when the proportion of outliers is relatively high and the ratio of the sample size and dimension is small.

6. Examples

6.1. Stack loss data

The stack loss data are "famous" in the literature of robust estimation and detection of outliers (see Atkinson, 1985, p. 129; Rousseeuw and Leroy, 1987, p. 76). There are 21 cases and 3 explanatory variables in the data, so that $p = 4$, and it is well known that observations 1, 3, 4, and 21 are extreme outliers and observation 2 is a mild outlier. The data are also used to illustrate the processes of the forward search algorithm based on LMS and LTS in Atkinson (1994) and Atkinson and Cheng (1999), respectively. Dodge (1996) traces the history of this data set to document the instances in which it has been used as an example of statistical methodology.

These data are also presented in transformation problems by several authors. The MLE of λ is 0.30 for the first-order model, and $\hat{\lambda} = 0.48$ when case 21 is excluded. The results of other models, e.g. the second-order model, refer to that of Atkinson (1985, pp. 129–136). The robust analysis of Carroll and Ruppert (1985) suggests that $\lambda = 0.5$ is reasonable. No outlier is revealed if the robust transformation parameter 0.42 is considered in Parker (1988). Atkinson and Riani (2000, Section 4.9) point out that the log transformation is not acceptable when observations 4 and 21 are deleted, but it is acceptable when all observations are included. Their fan plot of the score statistics shows that $\lambda = 0.5$ is supported by all the data.

The proposed robust method is applied to the first-order model for these data. To check the stability of the algorithm, 500 resampling schemes are used to show the performance, in which 100 subsamplings of sample size 5 are randomly drawn to obtain the optimum for each resampling scheme. When $q = [0.65n]$ and $[0.8n]$, all 500 resampling procedures yield the same estimate, $\hat{\lambda}_q = 0.4681$ and 0.4943, respectively. This is quite similar to the previous studies. If 75% of data are used for the LTS criterion, then there are two solutions in 500 subsampling schemes, $\hat{\lambda}_q = 0.9264$ for six times and $\hat{\lambda}_q = 0.4989$ for all others.

The author at first thought that the randomness of the resampling procedure and/or the number of subsamples cause the problem of a local optimum when $q = [0.75n]$. However, when 1000 subsets are randomly selected, the same two solutions are obtained again in 500 resamplings. In 16 resampling schemes, $\hat{\lambda}_q = 0.9264$, and the other 484 solutions lead to $\hat{\lambda}_q = 0.4989$; the corresponding deleted cases are {1, 3, 4, 13, 20, 21} and {2, 4, 13, 14, 20, 21}, respectively. In fact, we can see that the two solutions are not accidental, as $\hat{\lambda}_q = 0.9264$ implies that no transformation is required. The set of deleted cases includes those outliers identified when raw data are used. This may also partly explain the accuracy problem of the $\hat{\lambda}^{TW}$ diagnostic in Kim et al. (1996). We conclude that both these solutions are reasonable in terms of robustness and diagnostics when $q = [0.75n]$ is used for LTS. Nevertheless, the log transformation is more acceptable for this data set.

Table 3
The estimation results for tree data

q	$\hat{\lambda}_q$	Times in 500 samplings	Deleted cases
$[0.65n]$	0.1553	1	11, 14, 15, 16, 17, 18, 21, 23, 29, 30, 31
	0.3264	499	11, 14, 15, 16, 17, 18, 21, 23, 26, 27, 28
$[0.75n]$	0.2705	500	9, 11, 17, 21, 23, 26, 27, 28
$[0.85n]$	0.3582	476	15, 16, 18, 23, 26
	0.2115	24	15, 18, 29, 30, 31
$[0.95n]$	0.3472	500	15, 18

6.2. Minitab tree data

The Minitab tree data set is used by several authors to illustrate the problems of a regression transformation and a transform-both-sides model. This is a set of measurements on the volume, girth, and height of 31 black cherry trees. Atkinson (1985, pp. 124–129) compares some candidate models to provide a means of predicting the volume of timber in unfelled trees. He concludes that a formula based solely on girth would be preferred. If the first-order regression model is considered, which includes the response variable, volume, and two explanatory variables, girth and height, then the score statistic suggests strong evidence of a transformation on the response variable and $\hat{\lambda} = 0.3066$. Tsai and Wu (1990) conclude that the cube root transformation (the quick estimate of Cook and Wang (1983) for λ is 0.2931) to the dependent variable with the weighted regression model provides a reasonable explanation of the data.

When 500 resampling schemes are used to show the proposed algorithm, Table 3 summarizes the results produced by using 65%, 75%, 85%, and 95% of the data for the LTS criteria. If 65% of data are used for LTS, then there are two solutions resulting in 500 resamplings. One leads to $\hat{\lambda}_q = 0.1553$, and all others obtain $\hat{\lambda}_q = 0.3264$. When $q = [0.75n]$ and $[0.95n]$, all 500 resampling procedures yield the same estimates, $\hat{\lambda}_q = 0.2705$ and 0.3472, respectively. This result shows that the cube root transformation is reasonable in physical consideration.

The different estimated values of λ when different proportions of data are used for LTS may be due to the heteroscedastic error structure in the data (Tsai and Wu, 1990). We do not explore this issue further in this article.

6.3. Hill racing data

As first studied by Atkinson (1986b), this data set includes the record times for 35 hill races, together with the distances in miles and the climbs in feet. Without any transformation, Atkinson (1988) shows that cases 7, 11, 33, and 35 have largest positive residuals, in which the record time for observation 18 has been corrected. Given the initial values of $\lambda = 0$, 0.5, and 1, which appear in the null hypothesis when applying the score statistics, the quick

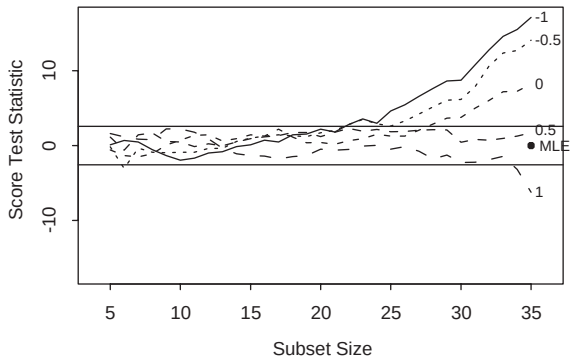

Fig. 1. Hill racing data: fan plot of the score statistics for five values of λ to the forward search algorithm.

estimates of λ vary from 0.67 to 1.10 if no case, case 7, and cases 7 and 33 are deleted from the data. Furthermore, he concludes that these data do not require a transformation based on the score test, and that cases 7 and 33 are outliers.

Applying the forward search procedure of Atkinson and Riani (2000) to these data, Fig. 1 shows the fan plot of the score test statistic, $T_p(\lambda)$, for $\lambda = -1, -0.5, 0, 0.5$, and 1. Table 4 presents the last five observations to enter the forward searches and the corresponding score test statistics and the values of the log likelihood for transformation at each stage. The transformations 0, -0.5, and 0 are clearly not acceptable. The most important feature is that $\lambda = 0.5$ is the only transformation for which the score statistic remains within the boundary through the search. Moreover, observation 7 is no longer an outlier for the square root transformation on these data. However, the largest value of the log likelihood is 158.9, which occurs at no transformation required and cases 7, 19, and 33 are excluded from the data.

Again, 500 resampling schemes are used. Table 5 shows the estimates of λ and the observations being excluded from the data when different values of q are considered. All 500 resamplings yield the same result for each q. We can see that the results are close to those of Atkinson (1988).

7. Conclusions

In this paper, we have combined robust and diagnostic approaches to deal with the problem of data transformations. This combination provides a resampling procedure which unites both a high breakdown estimate and a deletion diagnostic quantity. The performance of the proposed algorithm shows its stability and effectiveness from both simulations and real data presentations. The importance of combining both ideas of robustness and diagnostics emerges from the data analysis. The high breakdown estimate is used to resist the multiple outliers. The simulation study shows that it provides a very robust estimation procedure when an appreciable proportion of outlying points exist in the data. The deletion diagnostic quantity provides important information for data transformation in this

Table 4
Hill racing data: the last five observations to enter the forward searches for five values of λ, and the corresponding score test statistics and the values of the log likelihood for transformation at each stage

λ	Subset size				
	31	32	33	34	35
Observation					
−1	17	7	33	35	11
−0.5	17	7	33	35	11
0	17	7	33	35	11
0.5	6	14	19	35	11
1	11	6	19	33	7
Score statistic					
−1	10.81	12.77	14.59	15.54	17.14
−0.5	7.896	10.612	12.372	12.702	14.085
0	5.002	6.045	7.201	7.282	8.217
0.5	0.8157	0.7178	0.9560	1.2656	1.7300
1	−2.234	−1.887	−1.423	−3.174	−6.240
Log likelihood					
−1	86.59	82.45	80.57	79.07	78.67
−0.5	102.01	99.86	99.65	99.22	100.59
0	117.7	153.7	123.0	123.0	126.4
0.5	156.2	153.7	149.7	148.7	155.9
1	157.9	158.9	156.6	149.2	132.9

Table 5
The estimation results for hill racing data

q	$\hat{\lambda}_q$	Deleted cases
$[0.75n]$	0.8558	6, 7, 14, 15, 19, 26, 29, 30, 33
$[0.85n]$	0.9113	6, 7, 14, 19, 30, 33
$[0.90n]$	0.9277	7, 14, 19, 33
$[0.95n]$	1.0194	7, 33

paper. Furthermore, from the simulation results, the high breakdown estimate is essentially needed for the transformation problem when the percentage of outliers is relatively high in the data.

Outliers are relative to models, while transformations of the response variable lead to different models. There may be some different aspects or solutions when dealing with outlier detection and data transformations. On the other hand, as discussed by several authors, there does not exist a method that may successfully fulfill all the concerns or criteria for regression analysis. Several concerns related to the current work deserve further discussion. For examples, possible extensions can cover the high breakdown estimate with a high efficiency, or the bounded-influence estimate with a high breakdown point. Nevertheless,

the present article intends to show a possible way forward in understanding and solving part of these issues. Some interesting aspects are covered in the current study of the author (see Cheng, 2004).

Acknowledgements

The author would like to thank the Editor, and two anonymous referees for their helpful comments. This work was partly supported by a grant from National Science Council, Taiwan.

References

Atkinson, A.C., 1985. Plots, Transformations and Regression. Oxford University Press, Oxford.
Atkinson, A.C., 1986a. Diagnostic tests for transformations. Technometrics 28, 29–37.
Atkinson, A.C., 1986b. Aspects of diagnostic regression analysis (discussion of Influential observations, high leverage points, and outliers in linear regression by S. Chatterjee, A.S. Hadi). Statist. Sci. 1, 397–402.
Atkinson, A.C., 1988. Transformations unmasked. Technometrics 30, 311–318.
Atkinson, A.C., 1994. Fast very robust methods for the detection of multiple outliers. J. Amer. Statist. Assoc. 89, 1329–1339.
Atkinson, A.C., Cheng, T.-C., 1999. Computing least trimmed squares regression with the forward search. Statist. Comput. 9, 251–263.
Atkinson, A.C., Riani, M., 2000. Robust Diagnostic and Regression Analysis. Springer, New York.
Box, G.E.P., Cox, D.R., 1964. An analysis of transformations (with discussion). J. Roy. Statist. Soc. Ser. B 26, 211–252.
Carroll, R., Ruppert, D., 1985. Transformations in regression: a robust analysis. Technometrics 27, 1–12.
Carroll, R., Ruppert, D., 1987. Diagnostics and robust estimation when transforming the regression model and the response. Technometrics 29, 287–299.
Carroll, R., Ruppert, D., 1988. Transformations and Weighting in Regression. Chapman & Hall, London.
Chave, A.D., Thomson, D.J., 2003. A bounded influence regression estimator based on the statistics of the hat matrix. Appl. Statist. 52, 307–322.
Cheng, T.-C., 2004. Robust transformation for linear regression with both continuous and binary regressors (In: The presentation of JSM 2004, forthcoming)
Coakley, C.W., Hettmansperger, T.P., 1993. A bounded influence, high breakdown, efficient regression estimator. J. Amer. Statist. Assoc. 88, 872–880.
Cook, R.D., Wang, P.C., 1983. Transformations and influential cases in regression. Technometrics 25, 337–343.
Dodge, Y., 1996. The guinea pig of multiple regression. in: Rieder, H. (Ed.), Robust Statistics, Data Analysis, and Computer Intensive Methods. Springer, New York.
Fung, W.K., 1993. Unmasking outliers and leverage points: a confirmation. J. Amer. Statist. Assoc. 88, 515–519.
Hadi, A.S., Luceño, A., 1997. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. Comput. Statist. Data Anal. 25, 251–272.
Hadi, A.S., Simonoff, J.S., 1993. Procedures for identification of multiple outliers in linear models. J. Amer. Statist. Assoc. 88, 1264–1272.
Kim, C., Storer, B.E., Jeong, M., 1996. A note on Box–Cox transformation diagnostics. Technometrics 38, 178–180.
Krasker, W.S., Welsch, R.E., 1982. Efficient bounded-influence regression estimation. J. Amer. Statist. Assoc. 77, 595–604.
Lawrance, A.J., 1988. Regression transformation diagnostics using local influence. J. Amer. Statist. Assoc. 83, 1067–1072.
Müller, C.H., Neykov, N., 2003. Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. J. Statist. Plann. Inference 116, 503–519.

Parker, I., 1988. Transformations and influential observations in minimum sum of absolute errors regression. Technometrics 30, 215–220.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Amer. Statist. Assoc. 79, 871–880.

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

Rousseeuw, P. J., Van Driessen, K., 1999. Computing LTS regression for large data sets. Technical Report, University of Antwerp.

Rousseeuw, P.J., Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points (with discussion). J. Amer. Statist. Assoc. 85, 633–651.

Sakia, R.M., 1992. The Box–Cox transformation technique: a review. Statistician 41, 169–178.

Simpson, D.G.D., Ruppert, D., Carroll, R.J., 1992. On one-step GM-estimates and stability of inference in linear regression. J. Amer. Statist. Assoc. 87, 439–450.

Tsai, C.L., Wu, X., 1990. Diagnostic in transformation and weighted regression. Technometrics 32, 315–322.

Zaman, A., Rousseeuw, P.J., Orhan, M., 2001. Econometric applications of high-breakdown robust regression techniques. Econometrics Lett. 71, 1–8.