行政院國家科學委員會專題研究計畫　成果報告

同時具連續與離散解釋變數之穩健迴歸診斷分析

計畫類別：個別型計畫
計畫編號：NSC93-2118-M-004-007-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：國立政治大學統計學系

計畫主持人：鄭宗記

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 94 年 10 月 28 日

Robust Diagnostics for Linear Regression With Both Continuous and Binary Regressors

Tsung-Chi Cheng*

Abstract

Robust regression diagnostics is discussed for the case when the linear regression model contains both continuous and categorical regressors. A hybrid robust procedure is proposed, which is adapted from the M-S estimator of Maronna and Yohai (2000), and it is easier to be implemented in available statistical packages. Simulated data and real data analyses illustrate the performance of the resulting approaches.

Keywords: L_1 estimate; least trimmed squares estimator; M-S estimator; robust diagnostics; robust distance.

1 Introduction

Robust estimations and diagnostics for linear regression models have been widely discussed in the literature (see Atkinson (1985); Rousseeuw and Leroy (1987); Atkinson and Riani (2000); and references therein). Swamping (inliers appear as outlying) and masking (outliers appear as inlying) effects due to multiple outliers can be revealed and avoided by robust diagnostics. Previous studies consider the case where both response and regressors are continuous, but in practice it very often happens that data are mixed with both continuous and categorical regressor variables. A problem of singularity may occur when directly applying those robust estimators to a model of this kind.

A couple of papers solve the difficulty by separating the continuous and discrete regressors. Hubert and Rousseeuw (1997) propose a method, called RDL_1, that first downweights

*Department of Statistics, National Chengchi University, 64 Chih-Nan Road, Section 2, Taipei 11623, Taiwan. E-mail: chengt@nccu.edu.tw.

the leverage points in the space of the continuous regressors and then follows a weighted least absolute value (LAV, and denoted as L_1) fit for both continuous and categorical regressors. Maronna and Yohai (2000) consider two types of robust estimates for problems of this kind, and one of which basically alternate M and S estimates to obtain the robust estimates of regression coefficients. Further discussions about these estimators will be given in a later section.

The RDL_1 procedure of Hubert and Rousseeuw (1997) is a very intuitive approach and is easy to implement. However, it may suffer from the swamping effect due to its weights for the L_1 procedure being obtained by only considering the continuous design matrix. We therefore adapt the M-S estimate of Maronna and Yohai (2000) to obtain a new weight, which takes into account good and bad leverage points. A hybrid robust estimation procedure is then proposed, which essentially makes use of both ideas of Hubert and Rousseeuw (1997) and Maronna and Yohai (2000).

The outline of this paper is as follows. Section 2 summarizes the robust estimators proposed by Hubert and Rousseeuw (1997) and Maronna and Yohai (2000). A simulated data set demonstrates the possible swamping effect by RDL_1. A hybrid robust estimation procedure is therefore proposed in Section 3, together with illustrating the simulated data set and a real data example. Conclusions are drawn in Section 4.

2 Robust regression with both continuous and categorical predictors

In the classical linear regression model, we consider

$$y_i = \delta_0 + \boldsymbol{z}_i^T \boldsymbol{\delta} + \epsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

where explanatory variables $\boldsymbol{z}_i \in \boldsymbol{R}^p$ and they are usually quantitative, $\boldsymbol{\delta}$ is a $p \times 1$ vector of regression coefficients, and δ_0 denotes the intercept term. In addition, 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 least squares estimator (LSE) is the same as the maximum likelihood estimator (MLE). However, both are known to be sensitive to outliers. Robust methods are required for contaminated data. For most proposed algorithms to obtain robust estimates, a subset of s observations needs to be drawn first to compute the corresponding objective function, and this will be repeated many times. The value of s is $(p+1)$, such as PROGRESS in Rousseeuw and Leroy (1987) and most others (e.g. Hawkins (1994); Rousseeuw and Van Driessen (1999a)). The size of the subset starts from $(p+1)$ and then goes through to n in the forward search algorithm of Atkinson (1994).

Both quantitative and qualitative variables existing as regressors occur often in practice, and it is conventional to encode such categorical regressors by binary dummy variables. If there are m categorical variables with $c_1, \cdots c_m$ levels and $p_1 = \sum_{k=1}^{m}(c_k - 1)$ which indicates a categorical variable with c levels done with $(c-1)$ dummy variables, then we have

$$y_i = \beta_0 + \boldsymbol{x}_{1i}^T \boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}^T \boldsymbol{\beta}_2 + \epsilon_i, \tag{2}$$

where the regressors $\boldsymbol{x}_{1i} \in \boldsymbol{R}^{p_1}$ are categorical variables and $\boldsymbol{x}_{2i} \in \boldsymbol{R}^{p_2}$ are continuous variables. We denote that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are matrices with rows $\boldsymbol{x}_{i1}^T$ and $\boldsymbol{x}_{i2}^T$, respectively. Here, $\boldsymbol{X}_1$ is an $n \times p_1$ matrix with element either 0 or 1. For convenience and simplicity, we rewrite model (2) in the usual form as model (1)

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \cdots, n, \tag{3}$$

where $\boldsymbol{x}_i^T = (1 \ \boldsymbol{x}_{1i}^T \ \boldsymbol{x}_{2i}^T) \in \boldsymbol{R}^{p_1+p_2+1}$, and $\boldsymbol{\beta} = (\beta_0 \ \boldsymbol{\beta}_1^T \ \boldsymbol{\beta}_2^T)^T$ is a $(p_1 + p_2 + 1) \times 1$ parameter vector.

Researchers (eg., Draper and Smith (1998, Chapter 14)) treat categorical regressors as the usual continuous design matrix and then the ordinary LSE is applied. It seems natural to extend the robust regression methods to model (2) when contaminated data exist. However, this leads to a problem of singular matrices when treating those dummy variables as continuous regressors in the robust estimation (Hubert and Rousseeuw, 1997). As discussed above for the implementation of robust estimates, a hyperplane through those s points is obtained, and the corresponding objective function can be computed. This procedure is repeated to obtain the best fit. In the case of model (2) or (3), a large majority of the subsets

of $(p_1 + p_2 + 1)$ observations will be of less than full rank, and thus the hyperplane cannot be computed. The same problem occurs at the earlier stages (for those smaller values of s) in the forward search algorithm of Atkinson (1994).

2.1 RDL_1 estimator

Hubert and Rousseeuw (1997) propose a so-called RDL_1 estimator for model (2). The RDL_1 consists of three stages: identifying leverage points, downweighting the leverage points when estimating the parameters, and estimating the residual scale. At the beginning, they apply the minimum volume ellipsoid (MVE) estimator (see Rousseeuw and Leroy (1987)) to compute the robust distances for the continuous predictors

$$RD(\boldsymbol{x}_{2i}) = \sqrt{(\boldsymbol{x}_{2i} - \boldsymbol{t})\boldsymbol{C}^{-1}(\boldsymbol{x}_{2i} - \boldsymbol{t})^T}, \quad i = 1, \cdots, n, \tag{4}$$

where $\boldsymbol{t}$ and $\boldsymbol{C}$ are the location and scatter estimates of the $\boldsymbol{X}_2$ matrix, respectively. These distances (4) are used to identify the leverage points for the space of continuous regressors and to be the weights for estimating the regression coefficients by a weighted L_1 procedure at the second stage.

The parameters $\boldsymbol{\beta}$ of the model (2) are estimated by

$$\min_{\beta} \sum_{i=1}^{n} w_i |r_i(\boldsymbol{\beta})|, \tag{5}$$

where $r_i(\boldsymbol{\beta}) = r_i(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = y_i - (\beta_0 + \boldsymbol{x}_{1i}^T\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}^T\boldsymbol{\beta}_2)$, and

$$w_i = \min\left\{1, \frac{p}{(RD(\boldsymbol{x}_{2i}))^2}\right\} \tag{6}$$

for $i = 1, 2, \cdots, n$. The final stage calculates the estimate of the scale of the residuals, which is

$$\hat{\sigma} = 1.4826 \text{median}_i |r_i|, \tag{7}$$

where the constant 1.4826 leads to a consistent estimator under normality assumption. Hubert and Rousseeuw (1997) define the standardized residual as

$$t_i = \frac{r_i}{\hat{\sigma}}. \tag{8}$$

An observation is flagged as an outlier if its absolute value of (8) exceeds 2.5. The breakdown property of RDL_1 is referred to Hubert and Rousseeuw (1997), where an S-PLUS code to implement the RDL_1 is also provided.

Maronna and Yohai (2000) point out that there must be at least $p_1 + p_2$ null residuals by a well-known property of the weighted L_1 estimate. This may lead to the underestimation of the error variability. Instead of (7), they therefore suggest to use

$$\hat{\sigma}^* = s^*/0.675, \tag{9}$$

where s^* is the median of absolute non-null residuals. The standardized residual (8) is replaced by $\hat{\sigma}^*$ as follows,

$$t_i^* = \frac{r_i}{\hat{\sigma}^*}. \tag{10}$$

We use Maronna and Yohai's suggestions (9) and (10) in the following discussion.

2.2 M-S Estimator

Maronna and Yohai (2000) propose two types of estimates for model (2). The first one is a weighted L_1 estimate. The other consists of an M estimate for $\boldsymbol{\beta}_1$ and an S estimate for $\boldsymbol{\beta}_2$, because the former is not robust enough and the latter is too expensive for computation when using either one in the model. Their simulation results show that the weighted L_1 estimate would be better when the dimension of continuous regressors is smaller, whereas the M-S estimate is better when the dimension is equal or greater than 4, especially for high contamination.

Several versions to implement the alternating M and S estimates are discussed by Maronna and Yohai. We briefly summarize here the concept of the M-S estimator as follows. In the usual way, for example model (1), a regression M estimate for a complete dataset $(\boldsymbol{X}, \boldsymbol{y})$ is defined as

$$M(\boldsymbol{X}, \boldsymbol{y}) = \arg\min_{\delta} \sum_{i=1}^{n} \rho(y_i - \delta_0 - \boldsymbol{x}_i^T \boldsymbol{\delta}),$$

where ρ is an even convex function. If $\boldsymbol{X}$ corresponds to an ANOVA design, then M estimates attain the highest breakdown point. For each $\boldsymbol{\beta}_2$, define

$$\boldsymbol{\beta}_1^*(\boldsymbol{\beta}_2) = M(\boldsymbol{X}_1, \boldsymbol{y} - \boldsymbol{X}_2\boldsymbol{\beta}_2),$$

which is based on the idea that $\boldsymbol{\beta}_2$ is assumed to be known in model (2). Let $S(\boldsymbol{r})$ be a robust scale estimate of the residuals $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2$. An M-S estimate $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ for the linear model (2) is defined by

$$\hat{\boldsymbol{\beta}}_2 = \arg\min_{\beta_2} S(\boldsymbol{r}(\boldsymbol{\beta}_1^*(\boldsymbol{\beta}_2), \boldsymbol{\beta}_2)), \tag{11}$$

with $\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}^*(\hat{\boldsymbol{\beta}}_2)$.

Instead of (11), Maronna and Yohai (2000) suggest a simpler method by fitting $\boldsymbol{\beta}_1$ after removing the effect of $\boldsymbol{X}_1$ from $\boldsymbol{X}_2$ and $\boldsymbol{y}$. Let

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\gamma}}, \quad \tilde{\boldsymbol{X}}_2 = \boldsymbol{X}_2 - \boldsymbol{X}_1\hat{\boldsymbol{\Gamma}}, \tag{12}$$

where $\hat{\boldsymbol{\gamma}} = M(\boldsymbol{X}_1, \boldsymbol{y})$ and $\hat{\boldsymbol{\Gamma}}$ is a $p_1 \times p_2$ matrix, of which the jth column is $M(\boldsymbol{X}_1, \boldsymbol{x}_{2(j)})$, where $\boldsymbol{x}_{1(j)}$ is the jth column of $\boldsymbol{X}_1$. Define

$$\tilde{\boldsymbol{\beta}}_2 = \arg\min_{\beta_2} S(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}_2\boldsymbol{\beta}_2), \tag{13}$$

where $S(\cdot)$ is a robust scaled estimate of the residuals. Maronna and Yohai (2000) discuss that this will be a scale M estimate. Hence, $\boldsymbol{e}_0 = \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}_2\tilde{\boldsymbol{\beta}}_2$ is the residual vector.

When the columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are linearly independent, the M-S estimate $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ can be obtained as follows

$$\hat{\boldsymbol{\beta}}_2 = \tilde{\boldsymbol{\beta}}_2, \qquad \hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\beta}}_2, \tag{14}$$

which yields the same residuals

$$\boldsymbol{e}_0 = \boldsymbol{e}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = \boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2.$$

Note that the notions in this section follow Maronna and Yohai (2000), in which $\boldsymbol{X}_1$ is formed as an ANOVA design. The intercept term β_0 is then adjusted. A subsampling scheme to obtain an approximation solution of (11) is also provided in their paper. Also note that the estimates (14) are called $M1$-S estimates by them.

3 A hybrid robust estimation procedure

Inspired by the previous estimates of Hubert and Rousseeuw (1997) and Maronna and Yohai (2000), we propose a simpler robust estimate for model (2). Both ideas of these two papers are employed in the hybrid estimate, which is based on some known robust estimates. The advantages of the new estimate are: (i) it inherits the good properties of the $M1$-S estimate, (ii) it is easier to implement in most statistical packages, and (iii) it will be able to avoid the possible swamping effect by RDL_1.

In the following subsections, we first use simulated data to show that RDL_1 may produce the swamping effect for the identification of outliers. The hybrid estimate will be presented after then.

3.1 Simulated data

Before going further, we first discuss the diagnostic plot proposed by Rousseeuw and van Zomeren (1990), from which the different types of outliers can be investigated. A diagnostic plot will plot the Studentised residuals (from the least median of squares (LMS) estimate) against the robust distances (from MVE) of the $\boldsymbol{X}$ matrix for model (1). The same plot has been discussed in Atkinson and Cheng (2000) and Rousseeuw and Van Driessen (1999b). The latter paper names this plot as a D-D plot. Both suggest to use the robust residuals by the least trimmed squares (LTS) estimate and robust distances based on the minimum covariance determinant (MCD) estimate.

Observations can be classified into good data, good leverage points, vertical outliers, and bad leverage points by a diagnostic plot. Employing this idea, we generate a set of observations including these four types for model (2) with $p_1 = 1$, $p_2 = 2$, and $n = 30$. The data are generated by the following model

$$y_i = 8 + x_{1i} + x_{2i} + x_{3i} + \epsilon_i, \qquad i = 1, 2, ...30, \tag{15}$$

where both x_{1i} and x_{1i} follow a standard normal distribution, x_{3i} is a binomial distribution with a success rate 0.5, and ϵ_i is a normal distribution with mean zero and standard deviation

0.5. Once these 30 observations have been generated, cases 25 and 26 are then set to be vertical outliers by doubling their y values and keeping the others. Cases 27 and 28 are bad leverage points by adding 8 to their x_1 values and keeping the others as well. Cases 29 and 30 are good leverage points by adding 8 to both their x_1 and x_2 values and reproducing the corresponding y values as model (15). Table 1 presents the resulting simulated data.

Table 1. A simulated data set

Case	x_1	x_2	x_3	y	Case	x_1	x_2	x_3	y
1	0.92	0.06	0	9.07	16	-0.57	1.23	0	8.53
2	-0.09	1.36	0	8.85	17	0.93	-0.42	1	9.62
3	1.20	1.22	1	11.26	18	-0.56	1.76	1	9.83
4	0.34	-1.31	0	7.88	19	1.59	-1.11	0	8.29
5	0.36	-0.83	0	7.42	20	-1.13	0.08	0	7.71
6	-1.48	0.31	0	6.95	21	-0.88	1.63	0	9.36
7	2.02	0.37	1	11.53	22	0.23	-0.17	0	8.15
8	-0.44	-1.84	0	6.14	23	-0.92	-0.34	0	6.58
9	-0.76	0.55	1	9.36	24	-0.90	0.10	0	7.28
10	1.79	1.30	1	12.57	**25**	0.08	-2.04	1	13.95
11	-0.15	0.47	0	7.86	**26**	1.69	-0.91	1	19.34
12	0.84	0.53	1	10.09	**27**	8.85	0.34	0	9.59
13	-1.56	1.72	0	7.79	**28**	6.60	-0.14	1	7.94
14	0.27	-0.67	1	9.04	**29**	9.60	7.09	0	24.78
15	-1.25	-1.27	0	5.76	**30**	8.76	8.70	1	26.03

We now apply the RDL_1 estimate to these simulated data. However, instead of MVE, as discussed by Rousseeuw and Van Driessen (1999b), we apply MCD for the estimates of the location and scatter of the continuous variables $\boldsymbol{x}_2$ to obtain weights (6). Both MVE and MCD estimators provide a high breakdown robust estimation of the multivariate location and scatter (Rousseeuw and Leroy 1987). Moreover, Butler *et al.* (1993) show that the MCD estimator has better theoretical properties than the MVE. Woodruff and Rocke (1994) also provide empirical results which show that the MCD is preferred over the MVE in their applications.

Part (a) of Figure 1 shows the standardized residuals (10). Cases 25, 26, 27, 28, 29, and 30 are revealed as outliers. This is due to the fact that the weight (6) is calculated only by the

continuous design matrix without considering the model fitting. Therefore, cases 27, 28, 29, and 30 are outlying from $\boldsymbol{X}$ space and will be given relatively small weights as shown in part (b) of Figure 1. This makes cases 29 and 30 become bad leverage points in the diagnostic plot of Figure 1 (c). The cutoff values are indicated ± 2.5 and $\sqrt{\chi^2_{p_2,0.975}}$ by horizontal and vertical lines. Note that the original RDL_1 procedure of Hubert and Rousseeuw yields a similar result for this simulated data set.

Figure 1 (d) shows the LS residuals without fitting cases 25, 26, 27, and 28. Cases 29 and 30 are obviously located near the regression surface.

===Figure 1 is here===

3.2 A hybrid estimator

To overcome the possible swamping effect due to the weights when applying the weighted L_1 estimate, we consider an alternative to calculate these weights by adapting a similar idea of alternating M and S estimates. Due to the popularity of the LTS in the robust regression (Zaman *et al.* 2001), and that LTS can also be viewed as an S-estimator (Rousseeuw and Leroy, 1987, p. 144), we apply LTS instead of the S-estimate in estimate (13), which is

$$\min \sum_{i=1}^{q} (e_{(i)}(\boldsymbol{\beta}_{2,q}))^2,$$

where $e^2_{(i)} = (e_{(i)}(\boldsymbol{\beta}_{2,q}))^2$ is the ith ordered residual of e^2_i, $i = 1, \cdots, n$. If $\tilde{\boldsymbol{\beta}}_{2,q}$ denotes the LTS solution of regression $\tilde{\boldsymbol{y}}$ on $\tilde{\boldsymbol{X}}_2$ evaluated at q, then residual i is defined by

$$e_i(\tilde{\boldsymbol{\beta}}_{2,q}) = \tilde{y}_i - \tilde{\boldsymbol{x}}_2^T \tilde{\boldsymbol{\beta}}_{2,q}, \qquad , i = 1, \cdots, n.,$$

where $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{X}}_2$ are obtained by using M-estimators of (12).

The choice of value q has been discussed by Atkinson and Cheng (1999) and Zaman *et al.* (2001). We use $q = [0.75n]$ as a default in the following discussions. We then define weights by

$$\tilde{w}_i = \min \left\{ 1, \frac{s^2_{LTS}}{e^2_i} \right\}, \qquad , i = 1, \cdots, n, \tag{16}$$

where s_{LTS} denotes the scale estimate by LTS. Because these weights are computed by fitting a robust LTS, both good and bad leverage points can be accommodated. Once weights (16) have been obtained, the weighted L_1 estimate (5) is applied. Finally, scale estimate (9) is employed.

The resulting procedure also consists of three stages as RDL_1. At the first stage, we apply M-LTS to obtain robust residuals, which can accommodate both good and bad leverage points. The weights (16) are used to downweight those regression outliers including bad leverage points and vertical outliers. Parameter $\boldsymbol{\beta}$ of model (3) is then estimated by a weighted L_1 procedure at the second stage. In the final step the scale of the residuals (10) is obtained. The plot of the resulting standardized residuals versus the robust distances of (4) based on MCD is informative to present the type of data. Each observation can be classified into one of four categories based on its location in the diagnostic plot. These four categories are good data, bad leverage points, good leverage points, and vertical outliers.

The whole procedure is somehow similar to the $M1$-Sw estimate of Maronna and Yohai (2000). Therefore, it inherits the properties of the M-S estimator discussed in Maronna and Yohai's paper. Furthermore, the proposed procedure is very easy to be programmed in most statistical packages. For example, M, LTS, MCD, and L_1 are all built-in functions in S-PLUS, which are rreg, ltsreg, mcd.cov, and l1fit, respectively.

3.3 Simulated data (continued)

We next apply the robust hybrid method to the simulated data in Subsection 3.1. Parts (a), (b), and (c) of Figure 2 show the resulting robust standardized residuals, weights, and the diagnostic plot, respectively. From part (a), the proposed procedure successfully identifies cases 25, 26, 27, and 28 to be outliers. It gives weight 1 for both cases 29 and 30 as shown in part (b). The corresponding diagnostic plot also classifies all points into the right categories as the original configuration of these data being generated.

===Figure 2 is here===

3.4 Wagner's data

This data set is illustrated in Hubert and Rousseeuw (1997) and Maronna and Yohai (2000), but different conclusions have been yielded in both papers. The data set consists of 4 economical variables in 21 regions around Hannover over three time periods. The model thus contains four continuous and two categorical regressors, region and period, and a total of 63 observations. Hubert and Rousseeuw (1997) report that there are 15 points with $|t_i| > 2.5$. By using the MCD for the weights to RDL_1, Figure 3 (a) shows the standardized residual (10). It shows that the only largest $|t_i^*|$ value greater than 2.5 is point 50 and point 29 is the second largest, but not significant as suggested by Maronna and Yohai. The diagnostic plot of Figure 3 (b) points out that case 50 is a bad leverage point whereas case 29 is a good leverage point. While applying the M-S estimates, Maronna and Yohai conclude that cases 37 and 58 possess some kind of interaction and one of them is an outlier. It is not clear which one is outlying, because two of their estimates identify either one, but not both. Case 8 is another outlier. Other larger standardized residuals include cases 34, 43, 56, and 60.

Parts (c) of Figure 3 presents the standardized residual and diagnostic plots by the hybrid robust estimate. This result supports Maronna and Yohai's conclusions. Observations 29 and 50 are actually good leverage points revealed by the diagnostic plot of part (d) in Figure 3.

===Figure 3 is here===

To confirm the results, we treat these data as one whole data set without considering the periods and regions, and also separate the data into three cross-sectional regression models according to the periods. Parts (a), (b), (c), and (d) of Figure 4 present the diagnostic plot obtained by fitting LTS on the whole data, period 1, period 2, and period 3, respectively. Case 58 is an vertical outlier in period 3, but is located near the fitted regression surface when the whole period is considered. Case 50 is obviously a good leverage point no matter that it has been fitted in either one period or the whole period. Case 29 is a bad leverage point when it is only considered in period 2, but moves to a good leverage point in all-data

fitting. Case 8 appears as a bad leverage point in period 1, but is located on the boundary between bad and good leverage points if only continuous regressors are fitted. Cases 34 and 37 are vertical outliers. In sum, these can support the previous conclusions. Case 50 suffers from swamping effect by using RDL_1.

===Figure 4 is here===

We also agree with Maronna and Yohai's suspicion about case 58 and its possible relation with case 37. In Figure 4 (d), case 37 appears as negative vertical outliers, but becomes quite central in the bulk of the the whole period data in plot (a). Finally, we confirm that observation 8 is a bad leverage point in both plots (a) and (b) of Figure 4.

4 Conclusions

Robust regression diagnostics has been discussed for the linear regression model with both continuous and categorical regressors in the present paper. The problem of instability for the robust regression method can be alleviated when more data are fitted in the model (Atkinson and Cheng 1999). They may lead to several possible directions for the feasible solutions. Those potential outliers and leverage points may shift far away from each other when the subsets of observations used in the fitting are slightly different. This situation becomes more challenging for the case when the data are mixed with continuous and categorical variables. The hyperplane of the fitted regression model may be easily altered for this kind of data. The data examples give a cautionary note on the use of RDL_1. Although the proposed hybrid robust procedure is mixed with M, LTS, MCD, and L_1 methods, it is easily implemented.

References

Atkinson, A. C. (1985) *Plots, Transformations and Regression*, Oxford: Oxford University Press.

Atkinson, A. C. (1994) "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, **89**, 1329-1339.

Atkinson, A. C. and Cheng, T.-C. (1999) "Computing Least Trimmed Squares Regression with the Forward Search," *Statistics and Computing*, **9**, 251-263.

Atkinson, A. C. and Cheng, T.-C. (2000), "On Robust Linear Regression with Incomplete Data," *Computational Statistics and Data Analysis*, **33**, 361-380.

Atkinson, A. C. and Riani, M. (2000), *Robust Diagnostic and Regression Analysis*, New York: Springer-Verlag.

Butler, R. W., Davies, P. L. and Jhun, M. (1993) "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics*, **21**, 1385-1400.

Draper, N. R. and Smith, H. (1998) *Applied Regression Analysis*, 3rd ed., New York: John Wiley

Hawkins, D. M. (1994) "The Feasible Solution Algorithm for Least Trimmed Squares Regression," *Computational Statistics & Data Analysis*, **17**, 185-196.

Hubert, M. and Rousseeuw, P. J. (1997), "Robust Regression With Both Continuous and Binary Regressors," *Journal of Statistical Planning and Inference*, **57**, 153-163.

Maronna, R. A. and Yohai, V. J. (2000), "Robust Regression With Both Continuous and Categorical Predictors," *Journal of Statistical Planning and Inference*, **89**, 197-214.

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

Rousseeuw, P. J. and Van Zomeren, B. C. (1990) "Unmasking Multivariate Outliers and Leverage Points" (with discussion), *Journal of the American Statistical Association*, **85**, 633-651.

Rousseeuw, P. J. and Van Driessen, K. (1999a) "Computing LTS Regression for Large Data Sets," Technical Report, University of Antwerp.

Rousseeuw, P.J. and Van Driessen, K. (1999b) "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, **41**, 212-223.

Woodruff, D. L. and Rocke, D. M. (1994) "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association*, **89**, 888-896.

Zaman, A., Rousseeuw, P. J., and Orhan, M. (2001), "Econometric Applications of High-Breakdown Robust Regression Techniques," *Econometrics Letters*, **71**, 1-8.

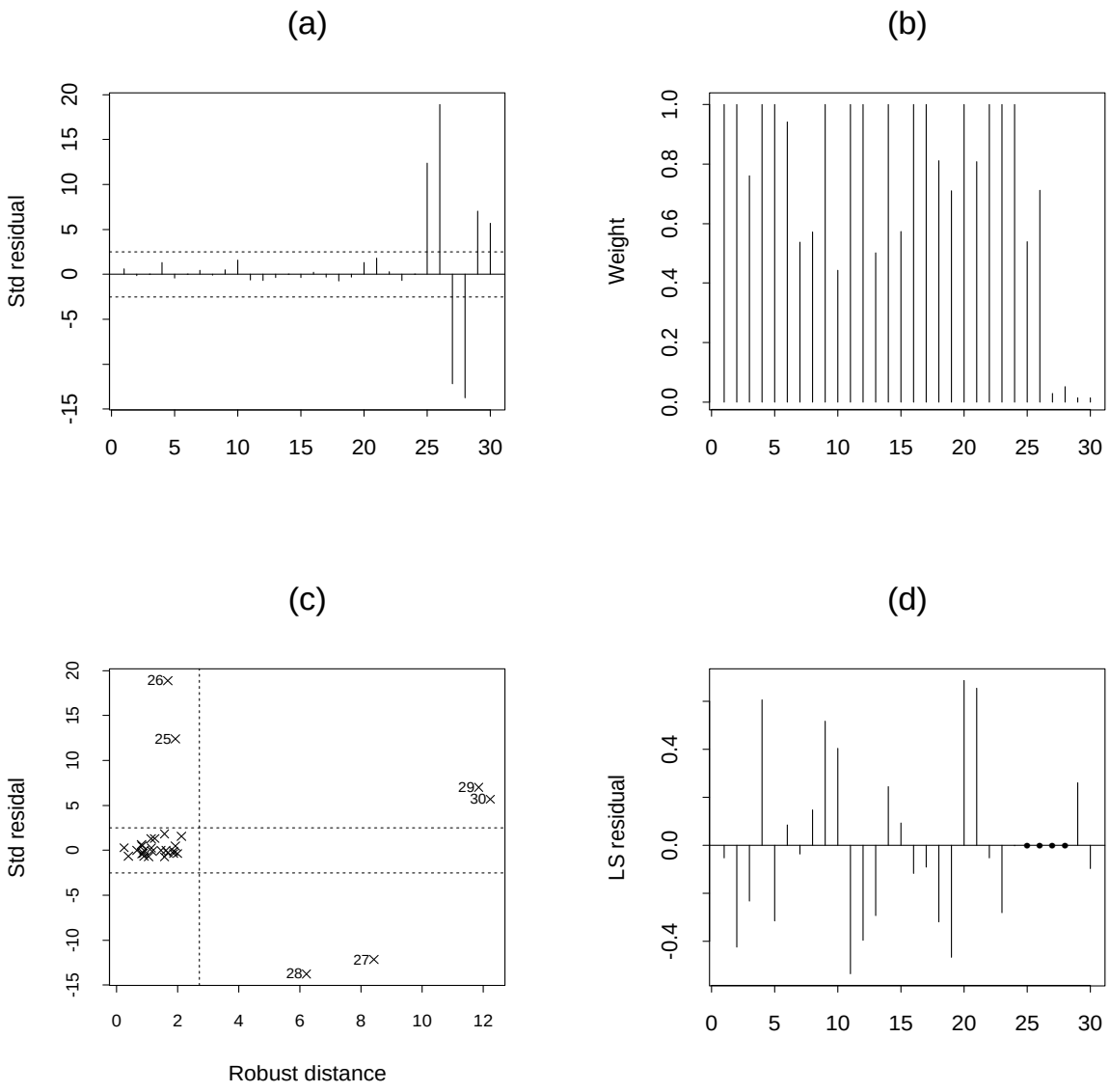

Figure 1: Simulated data analyses by using the RDL_1: (a) the plot of the standardized residuals; (b) the plot of weights; (c) the diagnostic plot; and (d) the least squares residuals without cases 25, 26, 27, and 28.

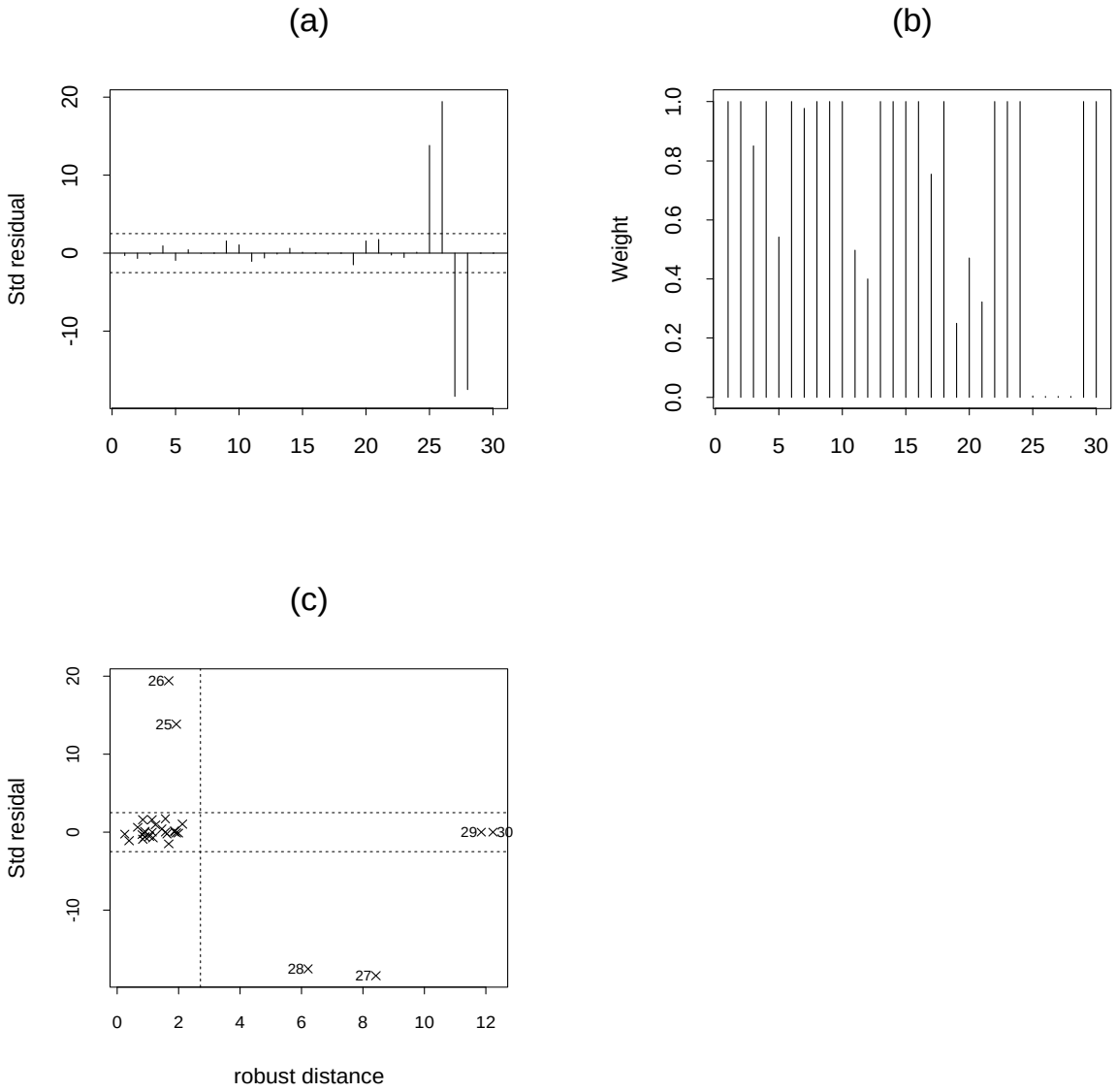

Figure 2: Simulated data analyses by using the hybrid robust procedure: (a) the plot of the standardized residuals; (b) the plot of weights; and (c) the diagnostic plot.

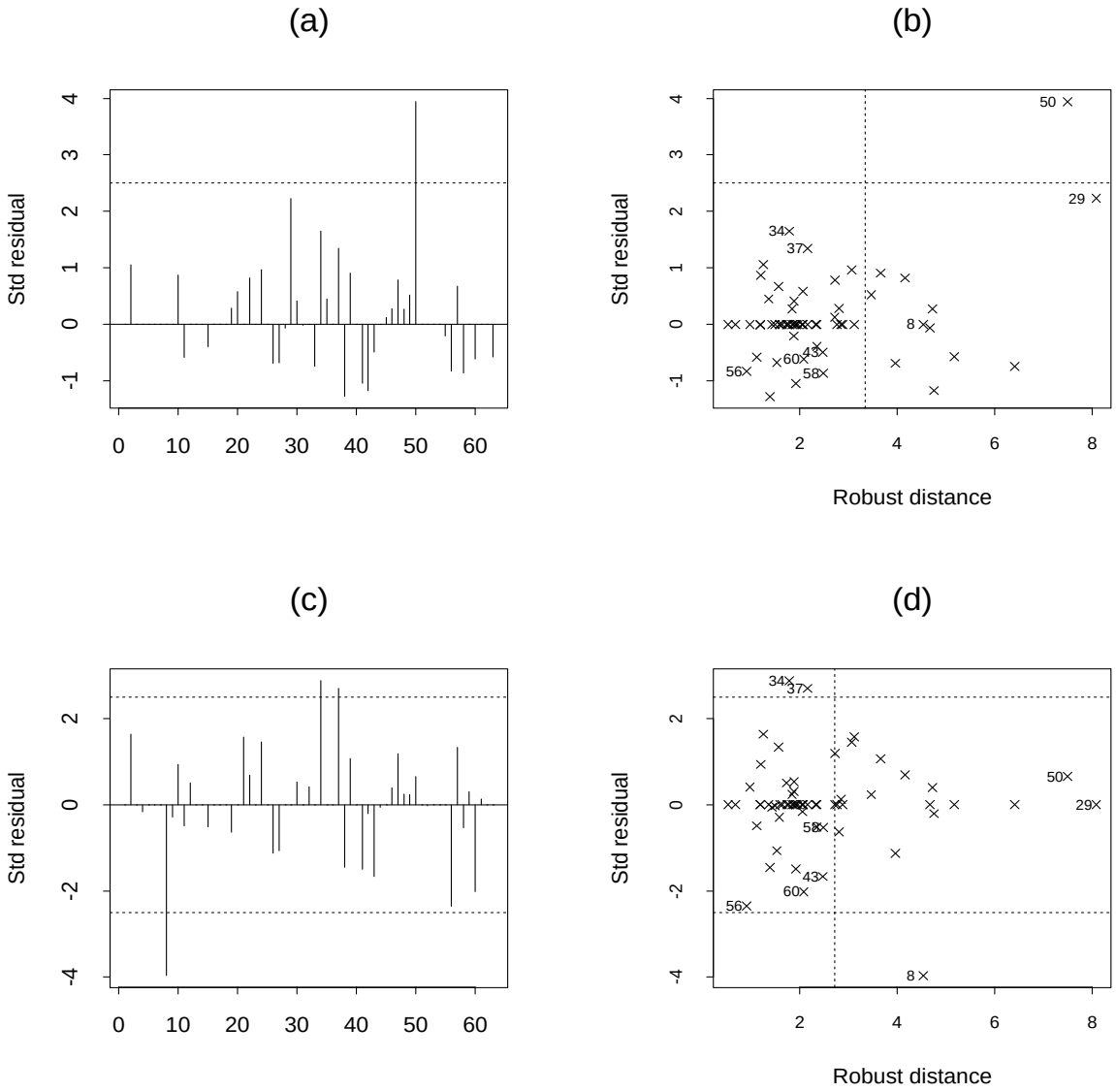

Figure 3: Wagner's data: (a) the standardized residuals and (b) the diagnostic plot by RDL1; (c) the standardized residuals and (d) the diagnostic plot by the hybrid robust procedure.

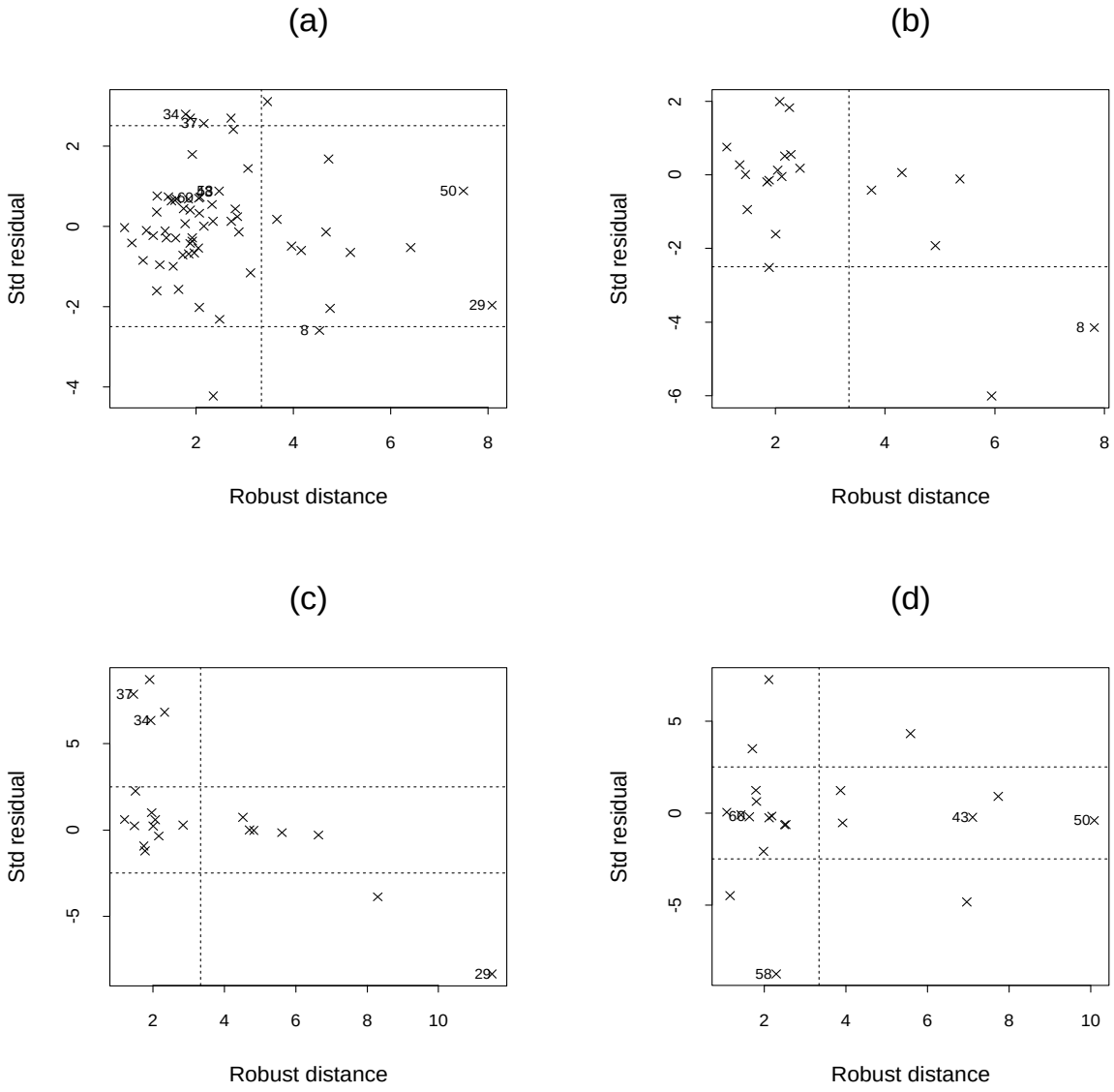

Figure 4: Diagnostic plots for Wagner's data by LTS regression: (a) Whole data; (b) Period 1; (c) Period 2; and (d) Period 3.