ELSEVIER

Journal of Statistical Planning and
Inference 57 (1997) 153–163

journal of
statistical planning
and inference

Robust regression with both continuous and binary regressors

Mia Hubert, Peter J. Rousseeuw *

Department of Mathematics and Computer Science, U.I.A., Universiteitsplein 1,
B-2610 Antwerp, Belgium

Received 1 February 1995; revised 1 May 1995

Abstract

We present a robust regression method for situations where there are continuous as well as binary regressors. The latter are often the result of encoding one or more categorical variables. In the first step we downweight leverage points by computing robust distances in the space of the continuous regressors. Then we perform a weighted least absolute values fit as a function of the continuous as well as the binary regressors. Finally, the error scale is estimated robustly. We pay particular attention to the two-way model, in which the proposed estimator is compared with an algorithm that treats the continuous and the categorical variables alternately. An S-PLUS function for the proposed estimator is given, and used to analyze a recent data set from economics.

AMS classification: 62F35, 62J05

Keywords: Analysis of covariance; Median Polish; Minimum volume ellipsoid estimator; Outlier detection; Robust distance; Weighted least absolute values

1. Introduction

In the classical linear regression model

$$y_i = \theta_0 + \sum_{j=1}^{p} \theta_j x_{ij} + e_i, \quad \text{where } e_i \sim N(0,\sigma^2),\ i = 1,\dots,n, \tag{1.1}$$

the explanatory variables x_{ij} are often quantitative. We now consider a model in which qualitative variables are also included. This situation often occurs in social and economical sciences, where the explanatory variables may include gender, ethnic background, professional occupation, marital status and so on.

* Corresponding author. Tel.: 32 3 820 24 12; fax: 32 3 820 24 21; e-mail: rousse@wins. uia. ac.be.

0378-3758/97/$17.00 © 1997 Elsevier Science B.V. All rights reserved
PII S0378-3758(96)00041-9

The usual convention is to encode such categorical regressors by binary dummy variables. If we have m categorical variables with $c_1,\ldots,c_m$ levels, we can write

$$y_i = \theta_0 + \sum_{j=1}^{p} \theta_j x_{ij} + \sum_{l=1}^{q} \gamma_l I_{il} + e_i, \tag{1.2}$$

where each I_{il} is either 0 or 1, and where $q = \sum_{k=1}^{m}(c_k - 1)$ since coding a categorical variable with c levels is done with $c - 1$ dummy variables. Actually, (1.2) is more general because the binary variables do not have to be the result of encoding categorical variables, hence other situations with dummy variables are also covered.

To clarify the ideas, let us consider a situation with two categorical variables ($m = 2$) as is often encountered in practice. Model (1.2) can then be written as

$$y_i = \theta_0 + \sum_{j=1}^{p} \theta_j x_{ij} + \sum_{k=1}^{c_1-1} \alpha_k I_{ik} + \sum_{l=1}^{c_2-1} \beta_l J_{il} + e_i. \tag{1.3}$$

The observations thus correspond to cell entries in a two-way table with c_1 rows and c_2 columns. The slope parameters are constant over all cells, but the intercept is not.

For an arbitrary value of m we obtain an $m-$way table. We will allow the number of observations to vary between cells. Also empty cells may occur, although an entirely empty cross-section is not permitted. (A cross-section consists of all the observations belonging to a fixed level of any categorical variable.) In the latter case the empty level should be removed, since its coefficient is not estimable from the data.

In general, our aim is to construct a robust estimator of the parameters θ_j and γ_l in (1.2). The least squares method (*LS*) fits the model (1.2) in a nonrobust way, by applying the standard calculations with $p + q$ regressors. It treats the dummies in the same way as the continuous regressors (see, e.g., Draper and Smith, 1981; Hardy, 1993). However, it is well known that the *LS* method is very sensitive to outliers.

The least absolute values (L_1) method as implemented by Armstrong and Frome (1977) is robust against outliers in the $y-$direction, but does not protect against points of which $(x_{i1},\ldots,x_{ip})$ is outlying. Such observations will be called *leverage points.*

A frequently used method of robust regression is $M-$estimation (also *LS* and L_1 belong to this class). This approach can also be applied to the model (1.2), as done by Birch and Myers (1982) for the case of one categorical variable. One then has to solve a system of $p + q + 1$ implicit equations, e.g., using an iteratively reweighted least squares algorithm. But M-estimators are still vulnerable to leverage points.

Therefore it seems natural to try to extend regression methods that can withstand a positive percentage of contamination, including leverage points. Typical examples are the least median of squares (LMS) estimator and the least trimmed squares (LTS) estimator (Rousseeuw, 1984), and the class of S-estimators (Rousseeuw and Yohai, 1984). However, we cannot simply run these estimators on (1.2) by treating the dummy variables in the same way as the continuous regressors, since this would lead to a problem of singular matrices. The typical algorithm for LMS regression in the model (1.1) starts by drawing a subset of $p + 1$ observations. Then the hyperplane through these $p+1$ points is obtained, and the corresponding objective function computed. This procedure is repeated often, and the best fit is kept. But in the case of $p+q$ regressors of which q are binary variables, a large majority of the $(p+q+1)$–subsets will be of less than full rank, hence the hyperplanes cannot be computed. The algorithm of Stromberg (1993) faces the same problem.

In the next section we will describe the proposed estimator. Section 3 studies the special case when there are two categorical variables in the model. A real example is worked out in Section 4. Finally, the Appendix provides S-PLUS code for the proposed algorithm.

2. Description of the estimator RDL$_1$

In this section we describe a new method for the general model (1.2). It consists of three stages. In the first stage we identify leverage points, and in the second stage these are downweighted when estimating the parameters. The final step estimates the residual scale.

In the first stage we look for leverage points, i.e., outliers in the set $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n\}$ where the components of $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ are the continuous regressors. Therefore X is a data set in p dimensions. To these data we apply the minimum volume ellipsoid estimator (MVE) introduced in (Rousseeuw, 1985). It consists of a robust location estimator $T(X)$ defined as the center of the smallest ellipsoid containing half of X, as well as a scatter matrix $C(X)$ given by the shape of that ellipsoid. One can then compute the robust distances defined as

$$RD(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - T(X))C(X)^{-1}(\boldsymbol{x}_i - T(X))^t} \tag{2.1}$$

(Rousseeuw and Leroy, 1987, pp. 265–269). If the $\boldsymbol{x}_i$ are observational (rather than designed) with a multivariate gaussian distribution, $T(X)$ and $C(X)$ are consistent for the underlying parameters (Davies, 1992). For large n the $(RD(\boldsymbol{x}_i))^2$ would thus be roughly χ^2_p distributed. Consequently, observations for which $RD(\boldsymbol{x}_i)$ is unusually large relative to that distribution can be identified as leverage points.

Based on the robust distances $RD(\boldsymbol{x}_i)$, we compute strictly positive weights w_i by

$$w_i = \min\left\{1, \frac{p}{RD(\boldsymbol{x}_i)^2}\right\} \tag{2.2}$$

for $i = 1, \ldots, n$. (The numerator p in (2.2) is the expected value of the chi-square distribution χ^2_p mentioned above.) By using strictly positive weights w_i no observations are entirely left out, thus no extra empty cells are created.

In the second step, the parameters $(\boldsymbol{\theta}, \gamma)$ of the model (1.2) are estimated by a weighted L_1 procedure

$$\underset{\boldsymbol{\theta}, \gamma}{\text{minimize}} \sum_{i=1}^{n} w_i |r_i(\boldsymbol{\theta}, \gamma)| \tag{2.3}$$

applied to the observations $(y_i, 1, x_{i1}, \dots, x_{ip}, I_{i1}, \dots, I_{iq})$. The solution $(\hat{\theta}, \hat{\gamma})$ can, e.g., be found using the least absolute values algorithm of Barrodale and Roberts (1973), who do not make a distinction between the continuous and discrete variables. Armstrong and Frome (1977) developed a faster L_1 algorithm which treats the two types of variables separately.

In the third and last step, the scale of the residuals is estimated by

$$\hat{\sigma} = 1.4826 \operatorname*{median}_i |r_i|, \tag{2.4}$$

where the constant 1.4826 makes the estimator consistent at gaussian errors.

The entire three-stage procedure using (2.1)–(2.4) will be called the *RDL*$_1$ estimator because it uses Robust Distances and L_1 regression. The robust estimate $(\hat{\theta}, \hat{\gamma}, \hat{\sigma})$ can now be used to detect regression outliers, by flagging the observations whose absolute standardized residual $|r_i/\hat{\sigma}|$ exceeds 2.5. The finite-sample efficiency of the estimators can then be increased by applying reweighted least squares to the data set, with weights depending on $|r_i/\hat{\sigma}|$. This also makes approximate inference available.

The finite-sample breakdown value ε_n^* of an estimator (Donoho and Huber, 1983) measures the maximum percentage of observations that can be replaced while leaving the estimate bounded. By construction, the estimator *RDL*$_1$ protects against leverage points by giving them small weights, whereas vertical outliers have only a small effect on the L_1 stage. However, an exact formula of the breakdown value of the *RDL*$_1$ procedure seems hard to find. At any rate, if the binary variables form an $m-$way table where each cell contains exactly one observation, we have the upper bound

$$\varepsilon_n^* \leqslant \frac{1}{\prod_{j=1}^{m} c_j} \left[\frac{1 + \prod_{j=1}^{m-1} c_{(j)}}{2} \right] \sim \frac{1}{2c_{(m)}} \tag{2.5}$$

on the breakdown value of any regression equivariant estimator. Here, $c_{(j)}$ denotes the jth smallest level among the m categorical variables. The upper bound (2.5) can be explained as follows. In the model (1.2), denote by r the rank of the set of binary vectors $\{\boldsymbol{I}_i = (I_{i1}, \dots, I_{iq}); 1 \leqslant i \leqslant n\}$. Then the breakdown value of any regression equivariant estimator of $(\theta_0, \gamma_1, \dots, \gamma_q)$ is bounded above by (see, e.g., Mili and Coakley, 1993)

$$\frac{1}{n} \left[\frac{n - N + 1}{2} \right], \tag{2.6}$$

where n denotes the number of observations and N is the maximum number of points $\boldsymbol{I}_i$ which lie in an $(r-1)$ dimensional plane. Here, the points $\boldsymbol{I}_i$ determine an m-way table with $c_1 \dots c_m$ levels. Since the complement of any cross-section of this table determines a hyperplane, N equals

$n -$ minimum number of observations in a cross-section.

If all cells contain one observation, we have $n = \prod_{j=1}^{m} c_j$ and $N = \prod_{j=1}^{m} c_j - \prod_{j=1}^{m-1} c_{(j)}$. Inserting these expressions in (2.6) leads directly to the upper bound (2.5).

2.1. Computational aspects

Our estimator RDL_1 can easily be implemented in S-PLUS (1993), because both the MVE and the L_1 estimator are built-in functions. The MVE estimates are obtained from the function *cov.mve* which uses a genetic algorithm. A special case of this algorithm, obtained by setting popsize=1, births.n=0 and maxslen=$p+1$, corresponds to the original resampling algorithm proposed by Rousseeuw and Leroy (1987, pp. 259–260). The L_1 computation in the S-PLUS function *l1fit* is based on the algorithm of Barrodale and Roberts (1973). Note that the weighted least absolute values problem on a data set $(\boldsymbol{x}_i, y_i)$ can be reduced to the least absolute values problem on the set $(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)$ obtained by the transformation

$$\tilde{\boldsymbol{x}}_i = w_i \boldsymbol{x}_i \quad \text{and} \quad \tilde{y}_i = w_i y_i.$$

The S-PLUS code of the whole RDL_1 procedure is given in the Appendix.

Alternatively, one can make use of the Fortran library ROBETH (Marazzi, 1993), which also includes procedures for the MVE and L_1 estimators.

3. The two-way layout

In this section we will focus on the case of two categorical variables ($m = 2$). For this we shall use the notation of (1.3).

In many two-way applications there is only a single observation per cell. This observation is often already a summary of actual data values, which are not available to the statistician. We then have to estimate $p + c_1 + c_2 - 1$ parameters from only $c_1 c_2$ data points. This is quite different from the one-way layout ($m = 1$), which assumes several observations for each level of the categorical variable. In that case we can estimate the parameters by least median of squares regression using a modified resampling algorithm, as described in Hubert and Rousseeuw (1995). We already pointed out in Section 1 why such an algorithm does not work well for more than one categorical variable.

For the two-way layout we also developed an alternative estimator, denoted as POL_1. The POL_1 method is also a new proposal, but we do not recommend it because its convergence is not guaranteed. Therefore, we will use it only to compare it with the RDL_1 method.

The POL_1 algorithm is defined by an iterative procedure which makes use of the two-way structure of the intercepts. In fact, the response y_i is on the one hand explained by the quantitative variables, and on the other by two qualitative variables. The former dependence can be analyzed by means of a linear model, and the latter by means of a two-way table. The POL_1 method carries out both steps alternately.

The skeleton of the POL_1 algorithm is as follows:

1. Initialize the residuals as $r_i \leftarrow y_i$.
2. Using only the continuous regressors, apply a robust regression estimator on the $(\boldsymbol{x}_i, r_i)$, yielding $(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_p)$. Put $r_i \leftarrow r_i - \hat{\theta}_0 - \sum_j \hat{\theta}_j x_{ij}$.
3. Using only the categorical variables, apply a robust technique for estimating the effects in a two-way table, here formed by the residuals obtained in step 2. Having the estimated effects $(\hat{\theta}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_{c_1-1}, \hat{\beta}_1, \ldots, \hat{\beta}_{c_2-1})$, form new residuals $r_i \leftarrow r_i - \hat{\theta}_0 - \sum_k \hat{\alpha}_k I_{ik} - \sum_l \hat{\beta}_l J_{il}$.
4. Repeat steps 2 and 3 until convergence. The final estimates $(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_{c_1-1}, \hat{\beta}_1, \ldots, \hat{\beta}_{c_2-1})$ are taken as the sum of the estimates calculated in the iterations.

We did not yet specify the robust estimators in steps 2 and 3. For the two-way table in step 3 we can use the *median polish* procedure of Tukey (1977). This is an iterative method where the estimates are obtained by subtracting row medians from the current cell entries, then subtracting column medians, and so on. This process is repeated until all rows and columns have zero median. In practice, a few iterations are usually sufficient. The median polish method can be seen as an approximation to the least absolute values estimate for a two-way table. A detailed account was given by Hoaglin et al. (1983). The median polish method is easy to implement, and is incorporated in S-PLUS.

In step 2 of the POL_1 algorithm we carry out an L_1 regression on the observations whose robust distances $RD(\boldsymbol{x}_i)$ do not exceed $\sqrt{\chi^2_{p,0.975}}$. This is a fast robust method, described in Rousseeuw and van Zomeren (1992). Note that the $RD(\boldsymbol{x}_i)$ need not be calculated at each POL_1 iteration step, as the explanatory variables $\boldsymbol{x}_i$ remain the same. The median polish in step 3 is applied to the same observations with $RD(\boldsymbol{x}_i) \leqslant \sqrt{\chi^2_{p,0.975}}$. The abbreviation POL_1 of the overall method stems from the combination of median POlish and L_1.

The POL_1 iterations are stopped if the norm of the new slope estimate $\hat{\boldsymbol{\theta}}$ in step 2 is less than a given precision, which in our program was set to 10^{-5}. We carried out several simulations (with one observation per cell) to investigate the algorithm. It converged in most cases, but sometimes diverged. Therefore, we do not recommend the POL_1 method in practice. We will compare it with the RDL_1 in the example below.

4. Example

To illustrate the RDL_1 method we consider an economics data set (Table 1) from Wagner (1994). He investigates the rate of employment growth (variable y) as a function of the percentage of people engaged in production activities (variable PA) and higher services (variable HS), and of the growth of these percentages (variables

Table 1
Real data set with four continuous and two categorical regressors

	Period 1					Period 2					Period 3				
Region	PA	GPA	HS	GHS	y	PA	GPA	HS	GHS	y	PA	GPA	HS	GHS	y
1	46.84	−2.60	1.68	0.20	0.97	44.24	3.80	1.88	0.13	8.47	48.04	−4.03	2.01	0.40	3.72
2	35.54	−1.42	1.67	0.63	2.14	34.12	−3.33	2.30	1.04	2.76	30.79	−3.10	3.34	0.25	9.29
3	28.42	−1.48	1.71	0.12	6.13	26.94	−1.71	1.83	1.28	24.08	25.23	−4.29	3.11	0.88	25.59
4	32.54	−4.51	1.37	0.32	7.36	28.03	−0.89	1.69	0.35	13.97	27.14	−2.45	2.04	1.95	27.41
5	28.92	−0.88	2.14	−0.08	3.63	28.04	−1.47	2.06	−0.81	0.63	26.57	1.31	1.25	0.67	18.32
6	36.61	−1.39	3.00	0.45	−4.30	35.22	−2.87	3.45	0.59	−1.99	32.35	−1.25	4.04	0.23	11.20
7	34.71	−2.22	2.94	0.27	2.06	32.49	−1.89	3.21	1.88	13.10	30.60	−3.21	5.09	−0.17	21.95
8	24.32	−5.11	3.57	−0.55	−18.64	19.21	0.36	3.02	2.98	15.42	19.57	2.48	6.00	2.27	33.03
9	35.15	−0.16	3.27	0.03	5.15	34.99	−4.95	3.30	0.68	19.65	30.04	−0.79	3.98	0.55	22.02
10	34.06	−3.86	2.74	0.19	6.88	30.20	−3.02	2.93	0.48	8.45	27.18	−1.14	3.41	0.28	13.68
11	37.94	−4.61	2.07	0.38	−1.24	33.33	0.06	2.45	0.24	9.04	33.39	−0.42	2.69	−0.18	11.24
12	35.88	−2.17	1.57	−0.11	−1.31	33.71	−4.67	1.46	2.59	9.47	29.04	1.40	4.05	−0.05	15.06
13	31.28	−1.90	2.74	−0.57	1.73	29.38	−2.74	2.17	0.07	24.18	26.64	1.04	2.24	0.12	10.73
14	33.61	2.02	1.92	0.32	0.44	35.63	−0.51	2.24	0.62	9.09	35.12	−0.81	2.86	0.25	1.53
15	33.86	0.75	0.86	0.46	−15.53	34.61	−5.36	1.32	0.61	−1.89	29.25	−2.56	1.93	0.30	11.37
16	43.24	−4.41	1.82	0.52	−10.99	38.83	−6.83	2.34	0.71	14.62	32.00	1.68	3.05	0.78	−0.07
17	42.65	−2.28	1.52	−0.17	0.60	40.37	−3.94	1.35	0.45	−0.44	36.43	1.00	1.80	0.37	14.17
18	37.19	−2.75	2.39	0.40	3.71	34.44	1.37	2.79	1.27	17.84	35.81	−2.29	4.06	0.18	8.05
19	49.70	−4.86	1.16	0.09	−2.38	44.84	−7.70	1.25	0.80	10.95	37.14	1.59	2.05	0.82	15.96
20	41.96	−4.59	2.00	−0.12	−1.35	37.37	−5.87	1.88	0.80	−1.55	31.50	0.70	2.68	−0.17	9.91
21	28.86	−2.11	5.17	0.46	−1.08	26.75	−1.83	5.63	1.35	−1.66	24.92	−0.14	6.98	0.59	6.94

GPA and GHS). The response also depends on the geographical region and the time period. The data set considers 21 regions around Hannover, and three time periods: 1979–1982, 1983–1988, and 1989–1992. The model thus contains four continuous and two categorical regressors. For each cell there is only one data point available, so the total number of observations equals 63.

Table 2 lists the weights w_i defined by (2.2), and the standardized residuals obtained by the LS, POL_1 and RDL_1 estimators. The residuals were divided by the classical scale estimate $\hat{\sigma} = \sqrt{(n-p-q-1)^{-1}\sum_1^n r_i^2}$ for LS, and by (2.4) for the robust methods.

The least squares residuals do not reveal any outlier. This is also seen in Fig. 1 which plots the standardized residuals $r_i/\hat{\sigma}$, none of which exceeds 2.5 in absolute value. On the other hand, both robust methods indicate the presence of outliers. The large residuals are boldfaced in Table 2, and lie outside their tolerance band in Fig. 1.

The RDL_1 method detects several outliers, whereas the iterative algorithm POL_1 finds only three (we do not consider residuals near the tolerance band to be outliers). A disadvantage of the POL_1 method is that it can create empty cross-sections, so that certain estimates and residuals are undefined. Here, this happens for region 8 (Laatzen). This problem cannot occur with the RDL_1 method. Note that the most extreme residuals obtained by RDL_1 correspond with the same region 8, in time periods 2 and 3. In Fig. 1, we see that the residual plot of RDL_1 is more informative than that of POL_1, which is still quite similar in shape to the classical LS plot.

Table 2
Weights and standardized residuals

	Period 1				Period 2				Period 3			
region	w_i	*LS*	POL_1	RDL_1	w_i	*LS*	POL_1	RDL_1	w_i	*LS*	POL_1	RDL_1
1	0.70	0.18	0.13	0.0	0.23	0.41	0.0	**2.7**	0.63	−0.59	−0.63	−1.6
2	1.0	0.83	0.32	**3.5**	1.0	−0.69	−0.62	0.0	1.0	−0.14	0.0	0.0
3	1.0	−0.38	−1.4	0.0	0.43	0.54	2.0	**3.2**	0.53	−0.16	1.4	0.41
4	0.91	0.12	0.07	0.0	1.0	−0.08	−0.07	0.0	0.18	−0.04	1.6	0.90
5	1.0	−0.32	−0.97	0.0	0.25	−0.69	−1.8	−2.3	0.56	1.0	1.0	**2.6**
6	1.0	0.03	0.0	0.0	1.0	−0.80	−0.20	−2.2	1.0	0.77	2.0	0.90
7	1.0	0.01	−1.5	0.0	0.18	−0.80	−0.75	−0.21	0.41	0.79	1.5	1.7
8	0.21	−1.2	–	0.0	0.06	0.18	–	**7.3**	0.07	1.1	–	**13.0**
9	1.0	−0.51	−2.1	0.0	1.0	0.57	1.9	1.4	1.0	−0.06	0.0	0.0
10	1.0	0.65	1.1	**2.9**	1.0	−0.43	0.0	−0.04	1.0	−0.22	−0.22	0.0
11	1.0	−0.45	−0.81	−1.9	1.0	0.04	0.0	0.0	1.0	0.41	0.01	0.0
12	1.0	0.79	−0.09	0.0	0.10	−0.79	0.53	−2.4	0.80	0.0	0.1	0.0
13	0.50	−0.90	0.0	0.0	1.0	1.8	**5.7**	**5.4**	1.0	−0.90	−0.28	−0.67
14	0.93	0.29	0.0	0.0	1.0	0.66	2.1	1.5	1.0	−0.94	−1.1	**−2.7**
15	0.78	−1.2	**−4.3**	−1.3	0.93	−0.06	0.0	0.0	1.0	1.3	1.7	2.2
16	1.0	−0.48	0.0	0.0	0.83	2.1	**6.4**	**4.4**	0.62	−1.6	−1.7	**−2.8**
17	0.82	0.42	0.0	0.0	1.0	−0.97	−1.2	**−4.2**	1.0	0.55	0.47	0.01
18	1.0	0.39	0.71	0.0	0.29	0.61	2.1	**3.0**	1.0	−1.0	−0.71	−2.0
19	0.49	−0.07	−0.07	0.94	0.50	0.38	**2.7**	0.0	0.49	−0.31	0.0	0.0
20	0.92	0.68	0.77	1.9	1.0	−1.1	−0.78	**−3.4**	1.0	0.39	0.0	0.0
21	0.47	1.2	0.64	0.0	0.18	−0.93	−1.5	**−3.8**	0.16	−0.25	−0.64	−1.9

Based on the properties of both robust methods, as illustrated in this example, we strongly recommend the RDL_1 method in practice. It is easier to implement than POL_1, runs faster, always converges, does not create empty cells, and in our experience yields more robust results.

A minor artefact of the RDL_1 plot in Fig. 1 is the zero residuals produced by the L_1 fit. If one wants to remove this effect, a simple and effective way is to append a fourth step to the RDL_1 algorithm. This final step computes a reweighted least squares (RLS) fit, in which the weight v_i of each observation depends on its RDL_1 residual, $|r_i/\hat{\sigma}|$. For our example, the RLS fit is shown in the lower portion of Fig. 1. In general, computing an RLS fit starting from a robust initial estimate tends to increase the finite-sample efficiency (see the simulations in Rousseeuw and Leroy, 1987, pp. 208–214). Moreover, the RLS yields the usual inferential output such as t-statistics and F-statistics. Note that the corresponding p-values are approximate, since they assume that the weights v_i have correctly identified the cases generated by the model (1.2).

5. Appendix

Below is an S-PLUS implementation of the proposed estimator RDL_1. It is very short because both the MVE estimator and L_1 fitting are intrinsic in S-PLUS. (Note that the current version of the function *cov.mve* cannot handle the univariate case $p = 1$, hence

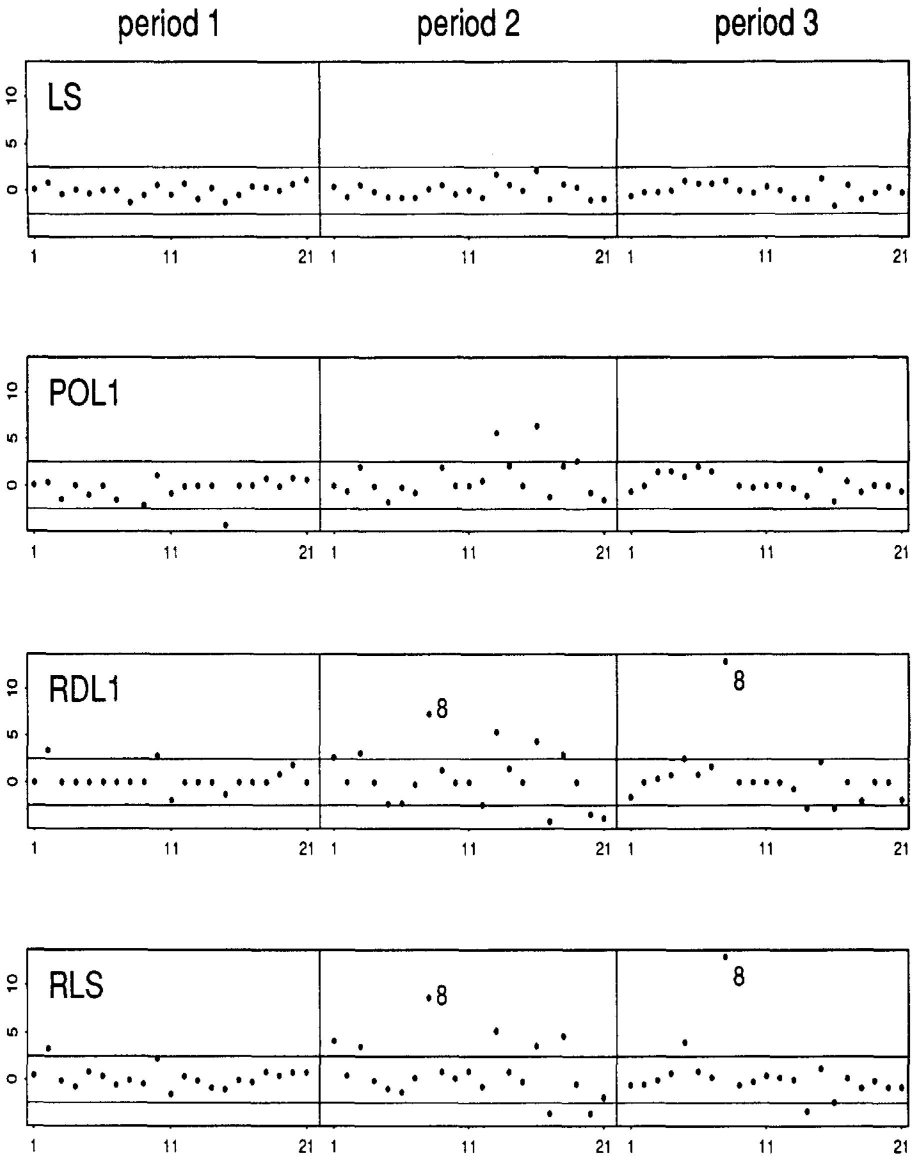

Fig. 1. Index plots of the standardized residuals produced by least squares (LS), the POL_1 method, the RDL_1 method, and reweighted least squares (RLS) based on the RDL_1 results.

the code below computes the sample median and the MAD in that case, yielding the robust distance $RD(x_i) = (x_i - \text{median}_j x_j)/(\text{mad}_j x_j)$ instead.)

```
rdl1.S <- function(x, xdum, y)
{
# The function RDL1 calculates a weighted L1-estimator
# of y on (x,xdum)
# with weights w = min(1, p/(RD^2))
#   where RD contains the robust distances
# obtained by applying the MVE estimator to x.
# input: x    : continuous regressors
#        xdum : dummy (binary) regressors
#        y    : response variable
        set.seed(4)
        x <- as.matrix(x)
        xdum <- as.matrix(xdum)
        y <- as.vector(y)
        p <- ncol(x)
        n <- nrow(x)
        xconstant <- as.matrix(rep(1,n))          # for intercept
        if(p == 1) {
                rob <- list(center = 0, cov = 1)
                rob$center <- median(x)
                rob$cov <- mad(x)^2
        }
        else rob <- cov.mve(x, print.it = F)
        robdist2 <- mahalanobis(x, rob$center, rob$cov)
        weight <- p/robdist2
        weight <- apply(cbind(1, weight), 1, min)
        estim <- l1fit(weight * cbind(xconstant,x,xdum),weight * y,
                intercept = F)
        res <- estim$res/weight
        scale <- 1.4826 * median(abs(res))
        st.res <- res/scale
        list(coef = estim$coef, weight = weight, res = res,
                    scale = scale,
               st.res = st.res)
}
```

References

Armstrong, R.D. and E.L. Frome (1977). A special purpose linear programming algorithm for obtaining least absolute value estimators in a linear model with dummy variables, *Comm. Statist. Simulation and Computation* B **6**, 383–398.

Barrodale, I. and F.D.K. Roberts (1973). An improved algorithm for discrete l_1 linear approximation, *SIAM J. Numer. Anal.* **10**, 839–848.

Birch, J.B. and R.H. Myers (1982). Robust analysis of covariance. *Biometrics* **38**, 699–713.

Davies, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* **20**, 1828–1843.

Donoho, D.L. and P.J. Huber (1983). The notion of breakdown point. In: P. Bickel, K. Doksum and J.L. Hodges, Jr., Eds. *A Festschrift for Erich Lehmann.* Wadsworth, Belmont, CA.

Draper, N.R. and H. Smith (1981). *Applied Regression Analysis.* Wiley, New York.

Hardy, M.A. (1993). *Regression with Dummy Variables.* Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-093, Sage, Newbury Park.

Hoaglin, D.C., F. Mosteller and J.W. Tukey (1983). *Understanding Robust and Exploratory Data Analysis.* Wiley, New York.

Hubert, M. and P.J. Rousseeuw (1996). Robust regression with a categorical covariable. In: H. Rieder, Ed., *Robust statistics, Data analysis, and Computer Intensive methods*, Springer, New York.

Marazzi, A. (1993). *Algorithms, Routines and S Functions for Robust Statistics: the FORTRAN library ROBETH with an Interface to S-PLUS.* Wadsworth, Belmont, CA.

Mili, L. and C.W. Coakley (1993). Robust estimation in structured linear regression. *Annals of Statistics*, to appear.

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

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In: *Mathematical Statistics and Applications.* Vol. B, W. Grossmann, G. Pflug, I. Vincze and W. Wertz, Eds., Reidel, Dordrecht, 283–297.

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

Rousseeuw, P.J. and V.K. Yohai (1984). Robust regression by means of S-estimators. In: J. Franke, W. Härdle and R.D. Martin, Eds., *Robust and Nonlinear Time Series Analysis.* Lecture Notes in Statistics, Vol. 26, Springer, New York.

Rousseeuw, P.J. and B.C. van Zomeren (1992). A comparison of some quick algorithms for robust regression, *Comput. Statist. Data Anal.* **14**, 107–116.

Statistical Sciences (1993). *S-PLUS User's Manual*, Version 3.2, StatSci, a division of MathSoft, Inc., Seattle.

Stromberg, A.J. (1993). Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression, *SIAM J. Sci. Comput.* **14**, 1289–1299.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading. MA.

Wagner, J. (1994). Regionale Beschäftigungsdynamik und höherwertige Produktionsdienste: Ergebnisse für den Grossraum Hannover (1979–1992). *Raumforschung und Raumordnung.* **52**, 146–150.