

In **The Practice of Data Analysis: Essays in Honor of John W. Tukey**, edited by D.R. Brillinger, L.T. Fernholz, and S. Morgenthaler (1997). Princeton, New Jersey: Princeton University Press, pages 193-202.

A regression analysis with categorical covariables, two-way heteroscedasticity, and hidden outliers

Mia Hubert and Peter J. Rousseeuw

*Department of Mathematics and Computer Science,
University of Antwerp, Universiteitsplein 1,
B-2610 Antwerp, Belgium.*

Mia.Hubert@uia.ua.ac.be, Peter.Rousseeuw@uia.ua.ac.be

<http://win-www.uia.ac.be/u/statis/index.html>

Abstract: We analyze a real data set described by Chatterjee and Price. The education expenditure in 50 states is explained by a linear model with three continuous regressors and two categorical variables (region and year). Estimates of the coefficients and the error scale are obtained by applying RDL_1 , a robust estimator that can handle qualitative as well as quantitative explanatory variables. Graphical displays of the residuals clearly reveal heteroscedasticity. The dispersion of the residuals is modeled as a function of the two-way structure, and estimated robustly by median polish. We then transform the data towards homoscedasticity, and obtain the final estimates. This procedure detects two large outliers that are hidden in a nonrobust analysis.

Keywords: Analysis of Covariance; Boxplot; Heteroscedasticity; Median Polish; Weighted Least Absolute Values.

1 Description of the data

In this paper we will analyze a real data set with both continuous and categorical regressors, which contains heteroscedastic errors and some hidden outliers. It is self-evident nowadays that analyzing data involves using tools and ideas introduced by John Tukey, in this case boxplots, median polish, and insights from residual plots.

The education expenditure data of Chatterjee and Price (1991, p. 119-121) consist of the per capita expenditure on education (EDUC) in the 50 states of the US, from 1965 until 1975. The states are grouped into four regions: North East (NE), North Central (NC), South (S), and West (W). Data are available for three years: 1965, 1970, and 1975. The expenditure in a state is tabulated according to the year and the region to which the state belongs. The observations thus correspond to cell entries in a two-way table with 4 rows and 3 columns.

The education expenditure is further explained by three continuous variables: the per capita personal income (P), the proportion of the population under 18 years of age (A), and the proportion of the population residing in urban areas (U).

To model this situation, we introduce indicator variables for the categorical regressors. The four regions can be coded with three indicator variables I_{i1} , I_{i2} and I_{i3} . They are defined by $I_{i1} = 1$ for observations in the first region (NE) and $I_{i1} = 0$ otherwise. Similar definitions hold for I_{i2} and I_{i3} . A fourth dummy variable I_{i4} is not necessary, since indicator variables I_{i1}, \dots, I_{i4} sum to 1 which is already used for the intercept term. The three years are coded analogously by two indicator variables I_{j1} and I_{j2} . In this paper, index i indicates the four regions, index j indicates the three different years, and index k runs through the n_{ij} observations of cell (i, j) of the two-way table. As the number of states per region did not change over time, we have $n_{ij} = n_i$. With this notation we can formulate the following linear model:

$$\begin{aligned} EDUC_{ijk} = & \theta_0 + \theta_1 P_{ijk} + \theta_2 A_{ijk} + \theta_3 U_{ijk} \\ & + \alpha_1 I_{i1} + \alpha_2 I_{i2} + \alpha_3 I_{i3} + \beta_1 I_{j1} + \beta_2 I_{j2} + e_{ijk} \end{aligned} \quad (1.1)$$

for $i = 1, 2, 3, 4$; $j = 1, 2, 3$; and $k = 1, \dots, n_i$. Initially, we assume the errors e_{ijk} to have a common unknown dispersion σ .

2 Estimation of the regression parameters

To obtain robust estimates of the regression coefficients $\{\theta_0, \theta_1, \theta_2, \theta_3, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2\}$ we need an estimator that can withstand both far leverage points and vertical outliers. Because of computational difficulties, robust regression estimators such as least median of squares (LMS) are less suitable for a model with both continuous and categorical regressors. These aspects are outlined in Hubert and Rousseeuw (1997). In the same paper, the RDL_1 estimator was proposed as a robust estimator that can handle both types of explanatory variables simultaneously. It is a weighted least absolute deviations estimator, with weights based on robust distances of the continuous regressors in the following way.

Let $\mathbf{x}_{ijk} = (P_{ijk}, A_{ijk}, U_{ijk})$ and $X = \{\mathbf{x}_{ijk}; \text{all valid combinations of } i, j \text{ and } k\}$. Then the robust distances are defined as

$$RD(\mathbf{x}_{ijk}) = \sqrt{(\mathbf{x}_{ijk} - T(X))C(X)^{-1}(\mathbf{x}_{ijk} - T(X))^t} \quad (2.1)$$

where $T(X)$ and $C(X)$ are determined by the center and the shape of the smallest ellipsoid covering half of X . Taken together, $T(X)$ and $C(X)$ constitute the minimum volume ellipsoid (MVE) estimator introduced by Rousseeuw (1985).

Based on the robust distances $RD(\mathbf{x}_{ijk})$, we compute the strictly positive weight

$$v_{ijk} = \min\left\{1, \frac{3}{RD(\mathbf{x}_{ijk})^2}\right\} \quad (2.2)$$

for each observation. Here, the '3' in the numerator is the number of continuous regressors. Finally, the parameters of the model (1.1) are estimated by a weighted L_1 procedure, which minimizes the sum of the weighted absolute residuals $\sum_{i,j,k} v_{ijk} |r_{ijk}|$.

A short S-PLUS function *rddl.s* performing these calculations is listed in Hubert and Rousseeuw (1997). For the education expenditure data, we obtain the parameter estimates shown in Table 1. For comparison it also lists the least squares estimates.

The scale of the residuals r_{ijk} can be estimated by the median of the absolute residuals, multiplied by a consistency factor for gaussian errors, namely,

$$\hat{\sigma} = 1.4826 \operatorname{median}_{i,j,k} |r_{ijk}|, \quad (2.3)$$

yielding $\hat{\sigma} = 21.59$ here. The absolute standardized residuals $|r_{ijk}/\hat{\sigma}|$ can then be compared to the cutoff value 2.5 to detect regression outliers. Figure 1 plots the standardized residuals

Table 1: RDL_1 and LS estimates of the parameters in the model (1.1)

Parameter	RDL_1 estimate	LS estimate
θ_0	-60.255	-23.162
θ_1	0.053	0.065
θ_2	0.381	0.231
θ_3	-0.025	-0.068
α_1	-26.269	-35.634
α_2	-19.262	-31.437
α_3	-28.772	-35.065
β_1	-73.887	-45.587
β_2	-21.685	-1.648
σ	21.586	30.551

obtained by the classical least squares analysis, the RDL_1 procedure, and a reweighted least squares analysis. The latter gives weight 0 to all regression outliers found by the RDL_1 estimator, and weight 1 to all other observations. Note that we could use smoother weights, as mentioned in Rousseeuw and Leroy (1987, p. 131). All analyses clearly show Alaska (AK) in 1975 to be an outlier, as noted by Chatterjee and Price (1991). The sizes of the other standardized residuals outside the tolerance band are all much smaller. (We also label Kentucky (KY) and Alabama (AL) which show up in the RDL_1 plot in 1965, since they will appear more prominently in the subsequent analysis.)

It can already be seen from Figure 1 that the residual dispersion is not homogeneous. This heteroscedasticity is clearly shown in Figure 2, which plots the RDL_1 residuals versus the fitted values \hat{y} . A rationale for plotting residuals versus \hat{y} rather than y can be found in chapter 16 of Mosteller and Tukey (1977). In the next section we will transform the data to stabilize the residual dispersion.

3 Estimating the error dispersion of each cell

Since the education expenditure data form a two-way table, we can ask whether the residual dispersion varies between cells. We first look at boxplots (Tukey 1977) of the robust residuals

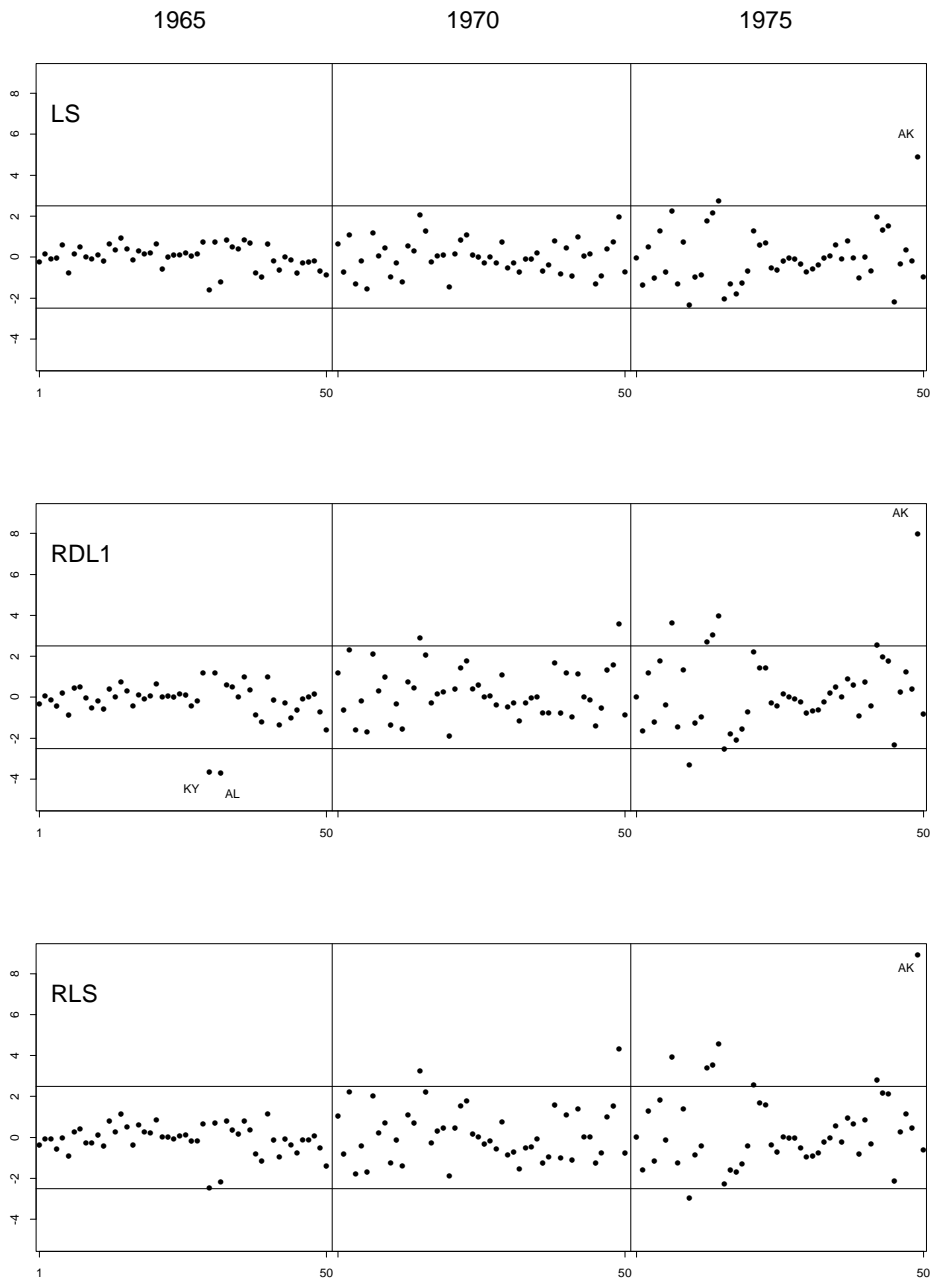


Figure 1: Standardized LS , RDL_1 , and RLS residuals

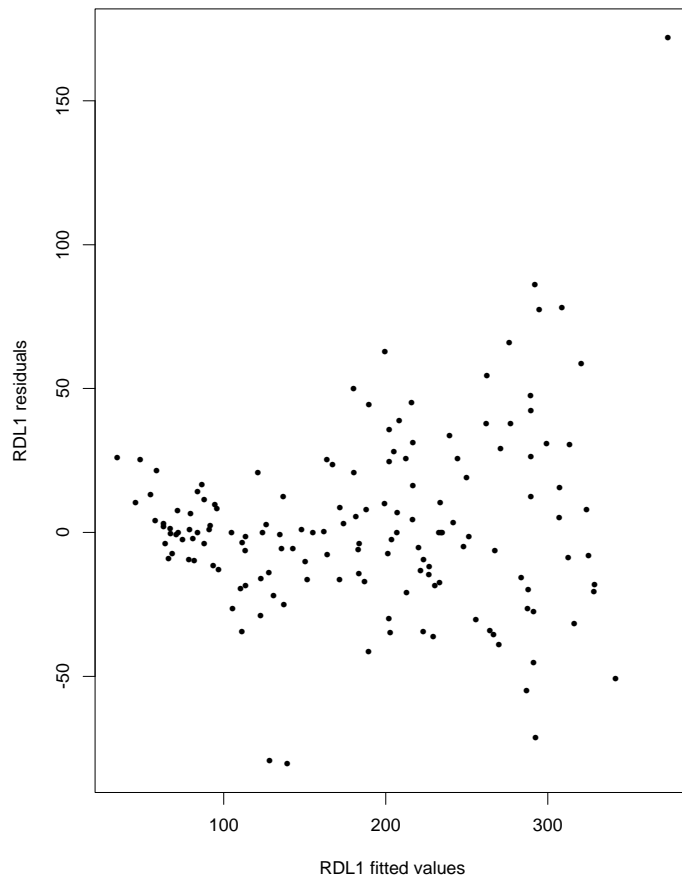


Figure 2: RDL_1 residuals versus fitted values

Table 2: Residual scale estimates per region and per year

	$\hat{\sigma}_{ij}$			s_{ij}		
	1965	1970	1975	1965	1970	1975
NE	12.42	40.91	49.18	20.35	40.91	44.43
NC	11.13	26.92	62.94	13.39	26.92	29.23
S	14.90	16.97	11.78	8.44	16.97	18.43
W	22.35	37.35	36.65	18.58	37.35	40.56

in all 12 cells, shown in Figure 3. In view of this inhomogeneity, we write σ_{ij} for the error dispersion in cell (i, j) .

An initial robust estimate $\tilde{\sigma}_{ij}$ of σ_{ij} based on the RDL_1 residuals in cell (i, j) is

$$\tilde{\sigma}_{ij} = 1.4826 \operatorname{median}_k |r_{ijk} - \operatorname{median}_k r_{ijk}|. \quad (3.1)$$

These preliminary estimates $\tilde{\sigma}_{ij}$ are used to determine a weight ω_{ijk} for each observation:

$$\omega_{ijk} = \begin{cases} 0 & \text{if } |r_{ijk}/\tilde{\sigma}_{ij}| > 2.5 \\ 1 & \text{otherwise.} \end{cases}$$

The scale estimates are then updated by the reweighting formula

$$\hat{\sigma}_{ij} = \sqrt{\frac{\sum_k \omega_{ijk} r_{ijk}^2}{\sum_k \omega_{ijk} - 4}}. \quad (3.2)$$

This $\hat{\sigma}_{ij}$ has a better finite-sample efficiency than $\tilde{\sigma}_{ij}$. Note that the '4' in the denominator of (3.2) is the number of θ -parameters in (1.1): namely, the intercept and the continuous regressors.

The left part of Table 2 lists the estimates $\hat{\sigma}_{ij}$ as a two-way table. They confirm what can already be seen from the boxplots in Figure 3: the third region (S) and the first year (1965) have a relatively small residual dispersion.

This statement leads us to consider the two-way structure of the residual dispersion σ_{ij} . It is natural to postulate a multiplicative model:

$$\sigma_{ij} = \sigma \sigma_i \sigma_j \quad (3.3)$$

which is equivalent to the additive model

$$\log(\sigma_{ij}) = \log(\sigma) + \log(\sigma_i) + \log(\sigma_j).$$

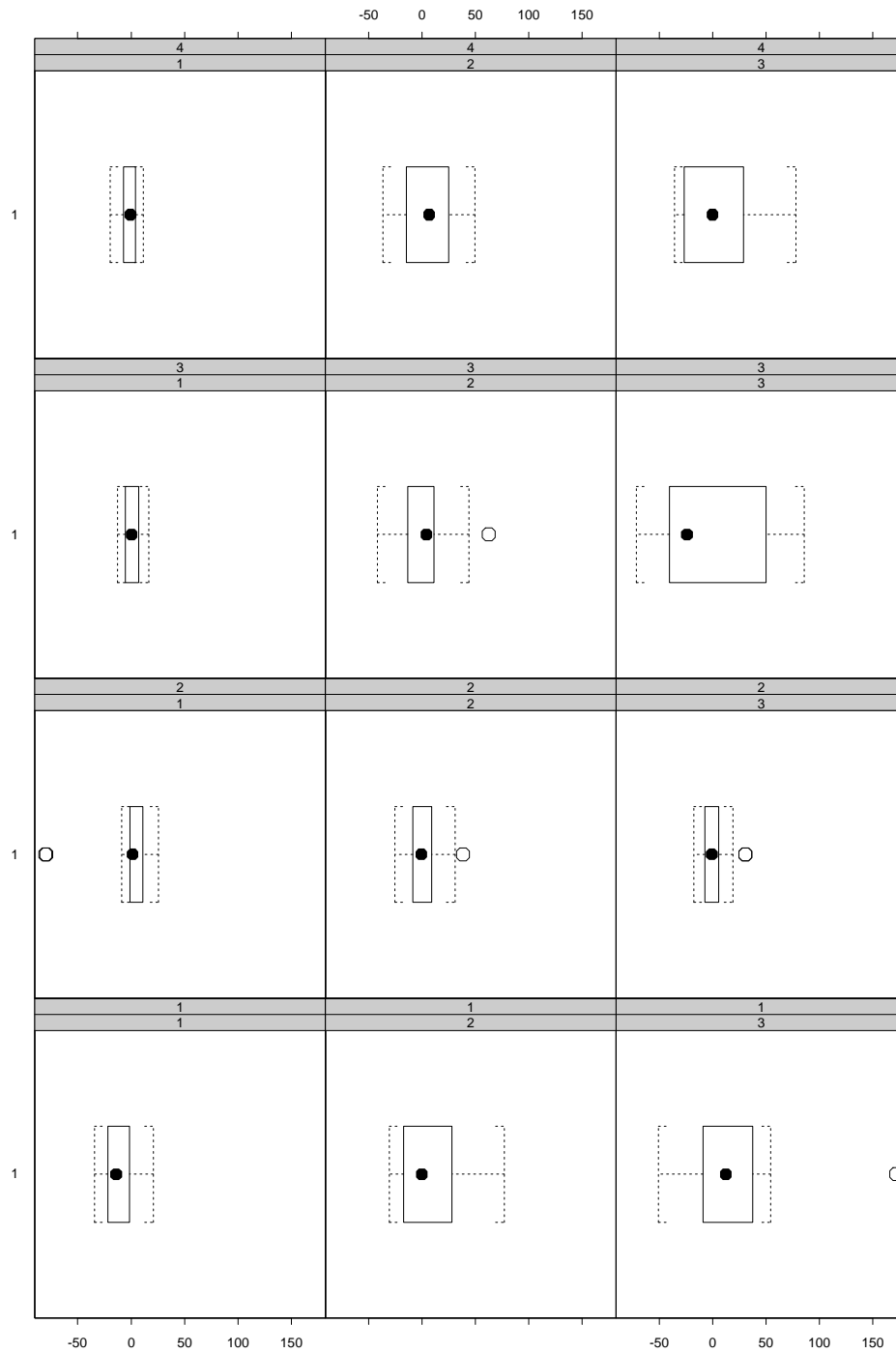


Figure 3: Boxplots of RDL_1 residuals by region (rows) and by year (columns)

We fit the latter model to the available $\log(\hat{\sigma}_{ij})$ by applying the median polish procedure of Tukey (1977); see also chapter 6 of Hoaglin, Mosteller and Tukey (1983). The effects estimates are obtained iteratively, by successively subtracting row and column medians from the current cell entries until all rows and columns have zero median. This procedure is very robust against outliers.

After applying the median polish technique to $\log(\hat{\sigma}_{ij})$ we transform the estimated cell entries $\log(\widehat{\sigma}_{ij}) = \log(\widehat{\sigma}) + \log\widehat{\sigma}_i + \log\widehat{\sigma}_j$ by taking exponentials, resulting in the estimates $s_{ij} = ss_i s_j$ shown in Table 2. The overall scale estimate $s = \exp(\widehat{\log\sigma})$ was 31.71. Note that median polish has detected one outlier: the large dispersion of NC in 1975.

4 Homogenizing the error dispersion

We can turn (1.1) into a regression model with constant error dispersion if we divide each observation by its appropriate scale estimate. Formally, we put $w_{ijk} = s/s_{ij}$ and multiply each term in (1.1) by w_{ijk} and obtain

$$\begin{aligned} w_{ijk}EDUC_{ijk} &= \theta_0 w_{ijk} + \theta_1 w_{ijk}P_{ijk} + \theta_2 w_{ijk}A_{ijk} + \theta_3 w_{ijk}U_{ijk} \\ &\quad + \sum_{l=1}^3 \alpha_l w_{ijk}I_{il} + \sum_{l=1}^2 \beta_l w_{ijk}I_{jl} + \tilde{e}_{ijk} \end{aligned} \quad (4.1)$$

for all i, j , and k . Here, the $\tilde{e}_{ijk} = w_{ijk}e_{ijk}$ have a common dispersion σ . Note that the parameters in (4.1) have another interpretation to that in (1.1), but for simplicity we will not rename them. For instance, (4.1) does no longer have a constant intercept term.

We again apply the RDL_1 estimator to the transformed variables, followed by a reweighted RLS step, as in Section 2 for the original data. Table 3 shows the resulting RLS parameter estimates. The corresponding standard errors and p -values are only approximations, since the weights w_{ijk} depend on the data.

The second part of Table 3 lists the analogous classical estimates. We started from the least squares (LS) fit on the original data, estimated the residual dispersion in each cell by the LS scale estimate, followed by a two-way analysis by row and column means. Finally, we performed an LS fit on the transformed data.

The plots of the standardized residuals in Figure 4 clearly indicate the difference between the robust and the nonrobust analysis. Whereas the least squares fit only reveals Alaska as an outlier, the robust procedure also finds two more extreme observations. The most

Table 3: *RLS* and *LS* estimates in model (4.1)

Parameter	RLS			LS		
	Estimate	Stand. Error	<i>p</i> -value	Estimate	Stand. Error	<i>p</i> -value
θ_0	-125.598	39.267	0.002	43.072	25.873	0.098
θ_1	0.063	0.005	0.000	0.057	0.005	0.000
θ_2	0.433	0.079	0.000	0.063	0.034	0.068
θ_3	-0.032	0.016	0.048	-0.041	0.015	0.006
α_1	-19.284	7.632	0.013	-36.835	6.577	0.000
α_2	-12.865	5.533	0.022	-24.263	6.243	0.000
α_3	-15.332	5.201	0.004	-27.795	5.350	0.000
β_1	-51.504	14.000	0.000	-46.186	12.676	0.000
β_2	-7.660	8.991	0.396	6.758	8.752	0.441
σ	30.822			25.041		

outlying points are now Kentucky and Alabama, with a low education expenditure in 1965 relative to the pattern set by the other states.

5 Conclusions

The education expenditure data of Chatterjee and Price can be described by a linear model with three continuous and two categorical regressors. To these data we have applied *RDL*₁, a robust regression method designed for such mixed variables.

Residual plots and within-cell boxplots of the residuals indicated heteroscedasticity as well as the presence of outliers. This again demonstrates the power of these graphical tools for model building. Tukey's median polish technique was then used to fit the two-way structure of the residual dispersion. After stabilizing the variance in this way, the *RDL*₁ method detected two hidden outliers.

References

Chatterjee, S., and Price, B. (1991), *Regression Analysis by Example, 2nd Edition*, New

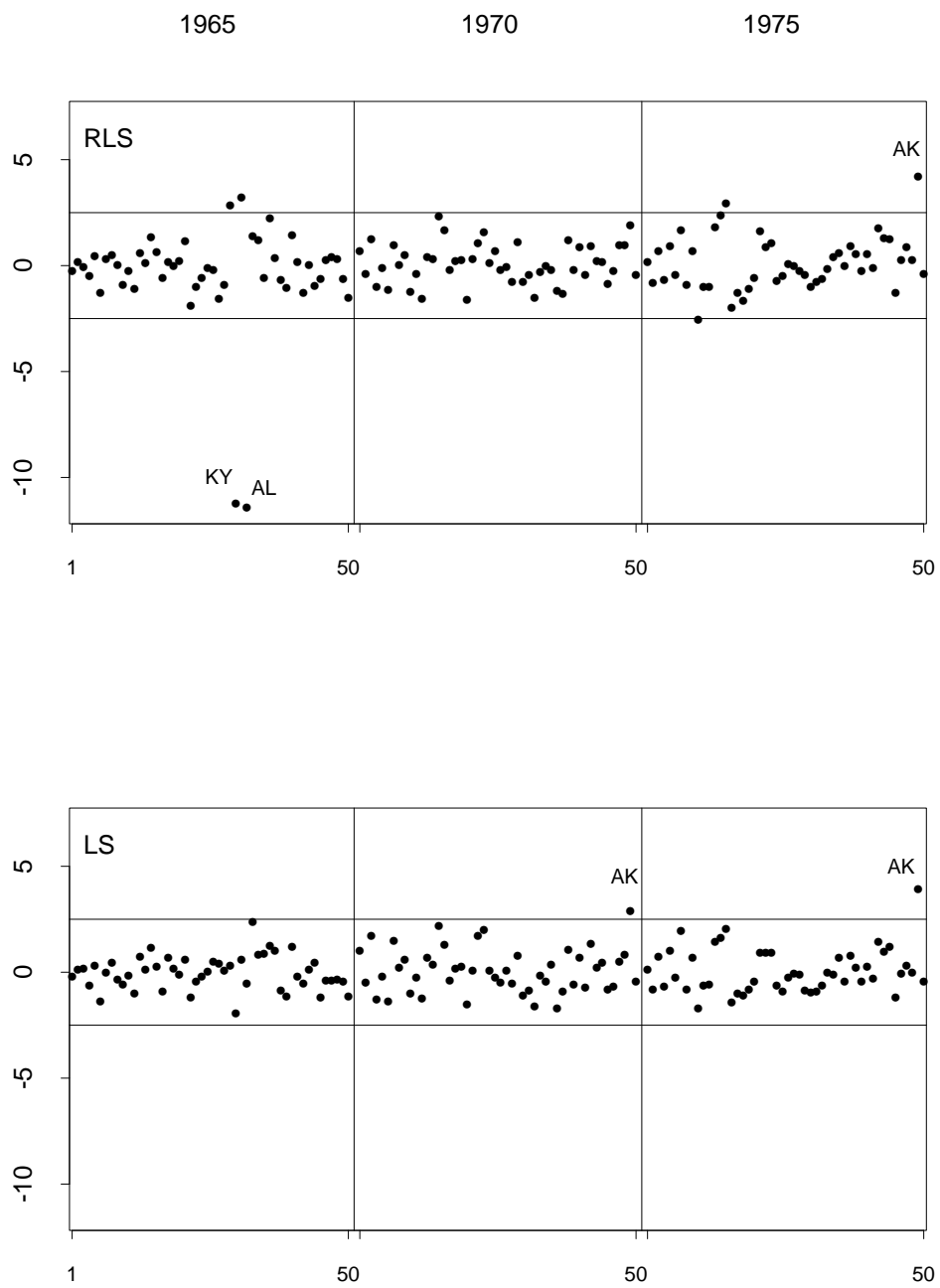


Figure 4: Standardized residuals of the *RLS* and *LS* fits to the model (4.1)

York: John Wiley.

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

Hubert, M., and Rousseeuw, P.J. (1997), Robust regression with both continuous and binary regressors, *Journal of Statistical Planning and Inference*, **57**, 153-163.

Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression*, Reading: Addison-Wesley.

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

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

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