Appendix A: shinyFrontend RobStatTM for GSOC 2018 Project

S-PLUS 6 Robust Library

User's Guide Version 1.0

May 2002

Insightful Corporation
Seattle, Washington

CONTENTS

INTRODUCTION TO ROBUST LIBRARY

1

OUR OVERALL GOAL

Our overall goal is to provide a broad range of robust methods for statistical modeling with the following features:

- **Automatic Computation of both Classical and Robust Estimates** When using the graphical user interface dialog for the Robust Library, the default choice is to automatically compute both the classical and the robust estimate. A special "fit.models" function for fitting multiple models is provided to facilitate computing both classical and robust estimates at the command line.
- **Outlier Data Mining and Comparison Plots**. Diagnostic plots are provided as a fundamental data mining tool that will assist you in quickly identifying outliers, in isolation or in small clusters, and determining whether or not outliers have substantial influence on the classical estimate.
- **Trellis Graphics Diagnostic Comparison Plots**. Trellis graphics are used to display side by side diagnostic plots for comparing classical maximum likelihood estimate (MLE) model fits with robust fits.
- **Robust Statistical Inference**. Robust t-statistics, p-values, F-tests, and bias-detection tests are provided, based on robust covariances and normal distribution approximations for parameter estimates.
- **Robust vs. Classical Inference Comparison is Facilitated**. Pairwise tabular displays of the robust and classical inference results facilitate quick comparison of inference results.
- **Special Scalable Methods for Linear Model Fits and Covariance Matrix Estimation**. Special methods are provided for robust fitting of linear models with large numbers of numeric predictor variables and/or many factor variables with possibly many levels, and for robust covariance matrix estimation with large numbers of variabels and large numbers of observations.

BASIC NOTIONS OF ROBUSTNESS

Classical maximum likelihood estimates (MLE) based on assumed idealized distributions almost always lack robustness toward outliers in the sense that outliers can have a very substantial influence on maximum likelihood parameter estimates. This is true not only of Gaussian maximum likelihood estimates such as the least squares estimates of linear models and the classical covariance matrix estimates, but also of a variety of non-Gaussian maximum likelihood estimates such as the MLE's for the parameters of exponential, Weibull and gamma distributions.

The probability distribution models that generate outliers are often close to the assumed ideal distribution in the central portion of the distribution, but differ from the ideal distribution in the tails of the distribution in a seemingly small but potent manner. The major consequence of such outlier-generating distributions is that the maximum likelihood parameter estimates based on the ideal distribution can suffer from large *bias* and substantially increased *variability* (or equivalently decreased statistical *efficiency*). Furthermore, the resulting bias persists even as the sample size n increases toward infinity, while the increased variability typically tends to zero like n^{-1}. Thus control of bias is more important for larger sample sizes.

Robust estimation methods were invented to deal with the above problems, and the important properties of a good robust estimator are as follows:

- In data-oriented terms: parameter estimates and the associated robust model fit are minimally influenced by outliers, and provide a good fit to the bulk of the data.
- Diagnostic plots based on the robust fit will allow you to quickly and easily identify outliers, and determine whether or not outliers are affecting the classical MLE model fits.
- In probability-oriented terms, a robust method minimizes the bias in coefficient estimates due to outlier-generating distribution models, while at the same time achieving a high *efficiency* when the data has the assumed ideal distribution (equivalently, the variance is not much larger than that of the MLE at the assumed ideal distribution)

- The robust parameter estimates provide good approximate statistical inference based on the large sample size approximate normality of the parameter estimates.

ROBUST MODELING METHODS

The following robust modeling methods are provided in the **Robust Library**.

- **Robust Linear Regression and Model Selection**
- Robust ANOVA
- Robust Covariance and Correlation Estimation
- Robust Principal Component Analysis
- Robust Fitting of Poisson and Logistic GLIM's
- Robust Discriminant Analysis
- Robust Parameter Estimates for Asymmetric Distributions

Robust Regression for the Linear Model

Two robust linear model fitting methods are included: (1) An MM-estimate, and (2) a new adaptive estimate due to Gervini and Yohai (1999). The MM-estimate is the default choice. The new adaptive estimate has the feature that it is asymptotically *efficient* when the data is Gaussian, i.e., is as good as least-squares when the data is Gaussian, while at the same time controlling bias due to outliers in a nearly optimal manner.

computation time

Both estimators described above require a highly robust initial estimate, and for the case all the predictor variables are numeric we continue to use a sampling approach to computing an initial S-estimate. It is known that such an approach has exponential complexity of order 2^p where p is the number of predictor variables. We provide some tabled estimates, based on empirical studies, of approximate computation times of the robust linear model fit as function of *p*, the number of observations, and the computer platform. The practical limit on the number of independent variables for reasonably quick computation with present generation workstations is roughly 15. In addition, we print out estimates of the time remaining for a robust fit so that you can decide whether to wait for the result or defer the computation to a more convenient time.

fast robust regression procedure

A fast procedure for obtaining initial estimates is implemented following Pena and Yohai (1999). Although these estimates are not guaranteed to have high breakdown point, they result in enormous speed improvement for large problems. The reliability of these estimates has been confirmed by simulation.

fitting models with both numeric and categorical variables

When you have factor type variables as well as numeric variables, each factor level requires an additional linear model parameter and dummy predictor variable. In such cases you may often find yourself with many more than 15 predictor variables. On the other hand, the predictor variables used to model the factor variables only take on the values 0 and 1, and for such variables a high breakdown point initial S-estimate is not really required. A least absolute deviations (LAD) type M-estimate will suffice. Based on this observation, Maronna and Yohai (1999) designed an alternating S-estimate/M-estimate method for fitting linear models with both factor and numeric predictor variables, which we have implemented for this release. This method will allow you to handle linear models with factor variables that require many more than 15 parameters.

robust model tests and robust model selection

Robust F-tests and robust Wald tests are provided. In addition, a robust model selection criterion called RFPE is provided. RFPE is a robust version of Akaike's Final Prediction Error criterion (FPE). Also, a robust backward elimination method for model selection is provided.

SPECIAL FEATURES OF THE ROBUST LIBRARY

Plots for Outlier Detection and Comparing Fits

When both a classical and a robust fit are computed, all plots selected on the **Plots** page of the dialog are created in a Trellis display with the classical and robust results displayed sided by side. For linear regression models including ANOVA models, QQ-plots and residuals density estimates are also available as overlaid plots. For regression models, plots of standardized residuals or deviances versus robust distances, introduced by Rousseeuw and von Zomeren (1990) are provided.

Multiple Model Fits and Comparisons Paradigm

A command line creator function `fit.models` is provided for creating an object of class "`fit.models`", along with `print`, `plot` and `summary` methods for this class of objects. You can use the function `fit.models` to create an object that contains both the least-squares and robust fits. Then you can make convenient comparison of the fits with respect to inference results by using `summary`, and convenient visual comparison of the fits and visual outlier detection by using `plot`.

GUI for the NT/ Windows Version

A uniform model-fitting dialog design has been created with the following basic features. Each robust model fitting dialog uses the current dialog for the classical model fitting method, with two minimal changes to accommodate the robust method. The first is that on each dialog's **Model** page you have three fitting method choices: (1) Compute both the classical and the robust fit, (2) Compute only the classical fit, and (3) Compute only the robust fit, with choice (1) being the default. Second, an **Advanced** page containing the various parameters and tuning constants used in the numerical procedures for the robust methods has been added to each dialog. Most users will not want to bother with these options.

LOADING THE ROBUST LIBRARY

Loading the library from the NT/ Windows GUI

To load the Robust Library from the NT/Windows GUI, select **File ► Load Library...** from the S+PLUS menu bar to bring up the following dialog window.

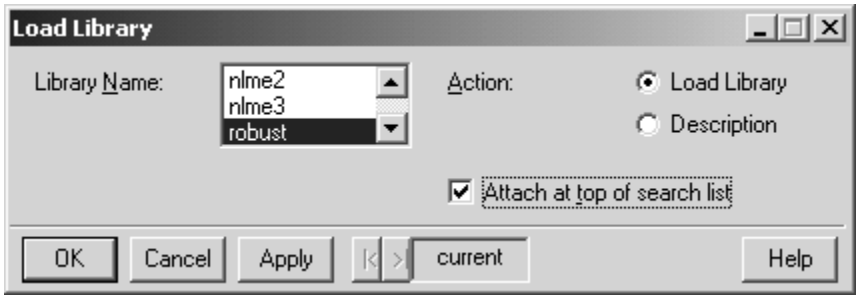

Figure 1.1: *Load Library Dialog*

Select Robust from the list of library names, and make sure to check Attach at top of search list. Click OK to load the library. After the Robust Library is loaded, a menu will be added to the S+PLUS menubar. Most of the functions provided by the Robust Library can be accessed through this menu.

Viewing the Robust Library Data Sets

If you use the Object Explorer you will want to be able to view and use the example data sets included in the **Robust Library**. To do this, right click (after loading the library) on the **Data** icon and select Advanced from the context menu to open the Database Filter dialog.

1

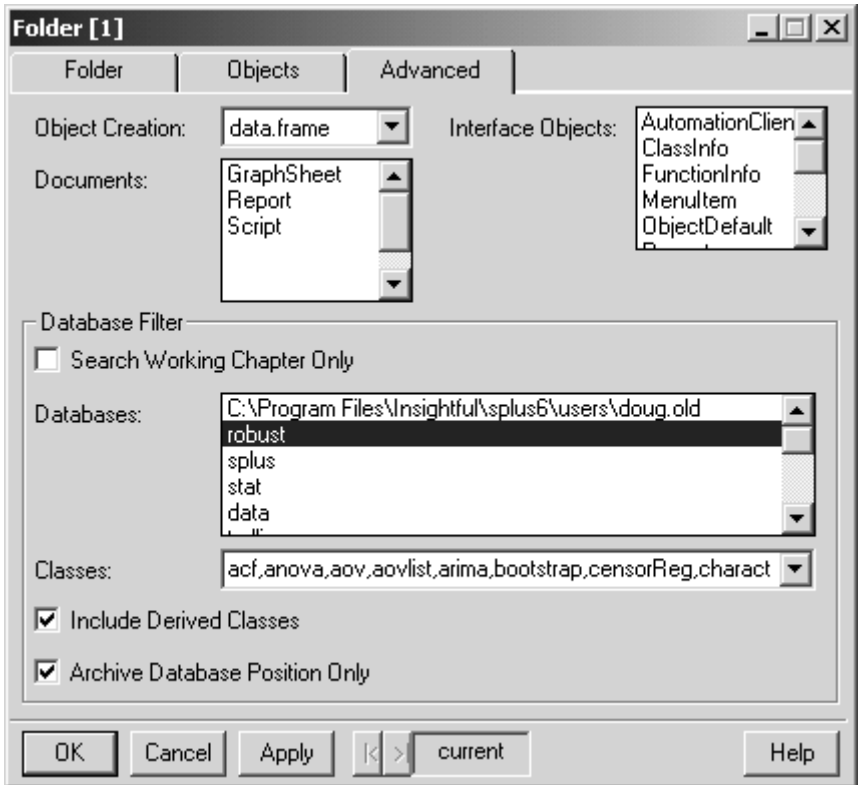

Figure 1.2: *Database Filter Dialog*

Uncheck the Search Working Chapter Only check box. Select Robust from the list of databases and enter `data.frame` Classes field (or select data.frame from the Classes drop-down list box). Then click OK. Notice that shortcuts to the **Robust Library** data sets have been added in the right pane of the Object Explorer.

Alternatively, in the Object Explorer you can expand the SearchPath object, then expand the robust object to access the data sets in the **Robust Library** directly.

Loading the Library from the Command Line

Use the following command to load the library from the commands window:

```
> library(robust, first=T)
```

This command will attach the **Robust Library** in position 2 and add the robust menu to the S-PLUS menubar.

ROBUST LINEAR REGRESSION

2

OVERVIEW OF THE METHOD: A SPECIAL M-ESTIMATE

You are fitting a general linear model of the form

$$y_i = x_i^T \beta + \varepsilon_i, \; i = 1, \; ..., \; n$$

with p-dimensional independent predictor (independent) variables x_i and coefficients β, and scalar response (dependent) variable y_i. S-PLUS computes a robust M-estimate $\hat{\beta}$ which minimizes the objective function

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T \beta}{\hat{s}}\right)$$

where $\hat{s}$ is a robust scale estimate for the residuals and ρ is a particular optimal symmetric *bounded* loss function, described in the section Theoretical Details. Alternatively $\hat{\beta}$ is a solution of the estimating equation

$$\sum_{i=1}^{n} x_i \psi\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{s}}\right) = 0$$

where $\psi = \rho'$ is a redescending (nonmonotonic) function. The shapes of the ρ and $\psi = \rho'$ functions are shown in Figure 2.17.

The above minimization problem can have more than one local minima, and correspondingly the estimating equation above can have multiple solutions. S-PLUS deals with this by computing special highly robust initial estimates $\hat{\hat{\beta}}$ and $\hat{\hat{s}}$, using the methods described in the section Theoretical Details. Then S-PLUS computes the final estimate $\hat{\beta}$ as the local minimum of the M-estimate objective function nearest to the initial estimate. We refer to an M-estimate of this type and computed in this special way as an MM-estimate, a term introduced by Yohai (1987).

COMPUTING LS AND ROBUST FITS WITH THE WINDOWS GUI

Computing Both LS and Robust Fits

You easily obtain both a least squares and robust linear model fit for the so called "stack loss" data using the Robust Linear Regression dialog in the **Robust Library**. The stack loss data is known to contain highly influential outliers, and is included in the **Robust Library** as the data frame `stack.dat`. Display the **Robust Library** data sets in the left-hand pane of the S-PLUS Object Explorer using one of the methods recommended in the Introduction chapter and select `stack.dat`. The right-hand pane of the Object Explorer displays the four variables in `stack.dat`: the dependent (response) variable `Loss`, and the three independent (predictor) variables `Air.flow`, `Water.Temp` and `Acid.Conc`. First select the response variable `Loss`, and then

)

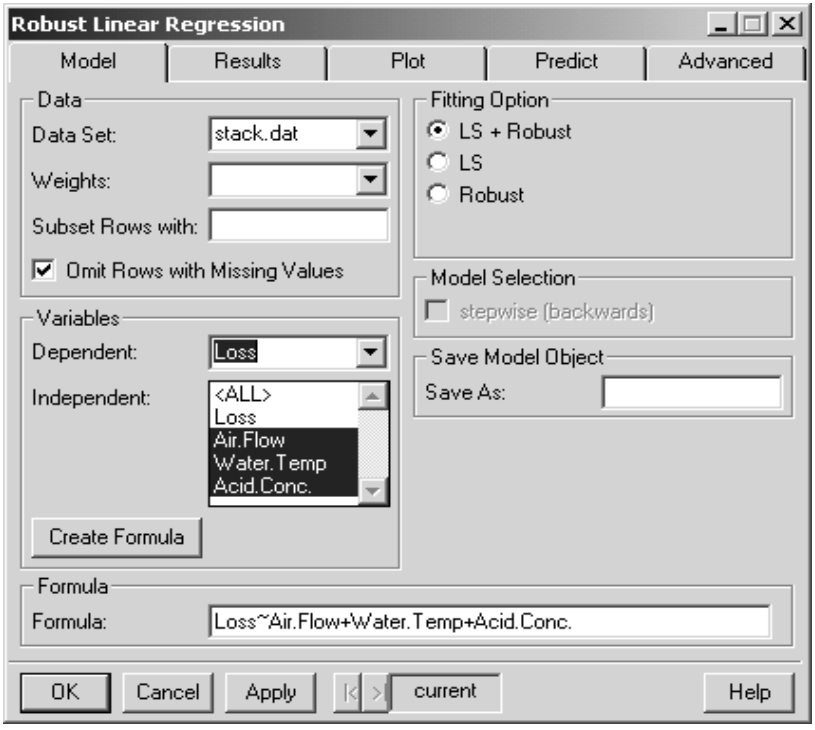

Figure 2.1: *The Linear Regression Dialog: Model Page*

select the three independent (predictor) variables (you can do this by shift clicking on `Acid.Conc`. Choose **Robust ► Linear Regression** from the menubar. The dialog shown in Figure 2.1 appears. Because you selected the response variable `Loss` first, followed by the three predictor variables, the **Formula** field is automatically filled in with correct formula `Loss ~ Air.Flow + Water.Temp + Acid.Conc.` for modeling Loss in terms of the three predictor variables.

Note that the **Model** page of this dialog looks exactly like that of the Linear Regression dialog in S-PLUS 6, except for the **Fitting Options** choices, with the default choice **LS + Robust** (both least squares and robust fits are computed) and alternate choices **LS** (least squares fit only) and **Robust** (robust fit only) and the **Advanced** tab. Click on the **Advanced** tab to access optional advanced features of the robust fitting method. These are discussed in the section Advanced Options For Robust Regression, and we suggest you wait until reading that section to experiment with the robust fitting method options.

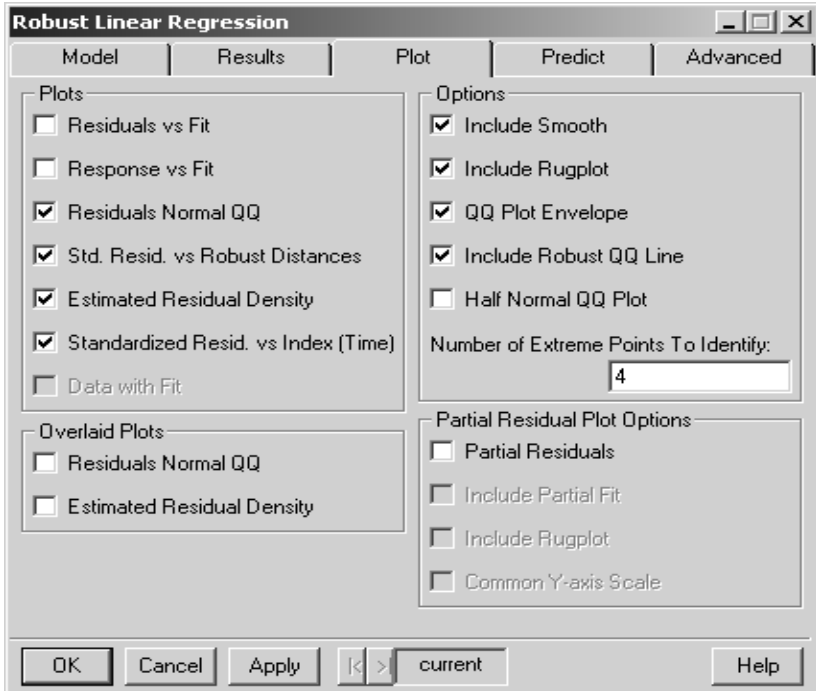

Figure 2.2: *The Linear Regression Dialog: Plot Page*

Click on the tabs labeled **Results**, **Plot** and **Predict** to look at those dialog pages. You will notice that the **Results** and **Predict** pages are identical to those of the **Linear Regression** dialog in S-PLUS 6. However, the **Plot** page shown in Figure 2.2 is different in that it has several new **Plots** region entries: *Std. Resid. vs. Robust Distances*, *Estimated Residual Density*, *Standardized Resid. vs. Index (Time)* and *Data with Fit*. The latter is greyed out when there is more than one independent variable. The **Plot** page also has a new **Overlaid Plots** region with the entries: *Residuals Normal QQ* and *Estimated Residual Density*. The latter are only available when you have chosen the default choice **LS + Robust** on the **Model** page.

We have made the default choices of plots indicated by the checked boxes. This will encourage you to quickly compare the LS and robust versions of these plots and quickly determine whether or not there are any outliers in the data, and whether or not the outliers have an impact on the least squares fit. In the **Number of Extreme Points to Identify** text box, replace the 3 by 4.

Click **OK** to compute both the LS and robust fits, along with the four diagnostic comparison plots and other standard statistical summary information. The results appear in a **Report** window and four tabbed pages of a **Graph Sheet**, respectively.

The Diagnostic Plots

Each of the **Graph Sheet** pages contains a Trellis display for the LS and robust fit, as shown below.

Normal QQ-Plots of Residuals

As seen in Figure 2.3, the normal QQ-plot for the LS fit residuals shows at most one outlier, while the one for the robust fit reveals four outliers. The outliers are those points that fall outside the 95% simulation envelopes for the normal qq-plot, shown as dotted lines. This reveals one of the most important advantages of a good robust fit relative to a least squares fit: the least squares fit is highly influenced by outliers in such a way that the outliers are not clearly revealed in the residuals, while the robust fit clearly exposes the outliers.

You also note that if you ignore the outliers, a normal distribution is a pretty good model for the residuals in both cases. However, the slope of the central linear portion of the normal QQ-plot of the residuals for the robust fit is noticeably smaller than that for the LS fit. This indicates that the normal distribution fit to the robust residuals,

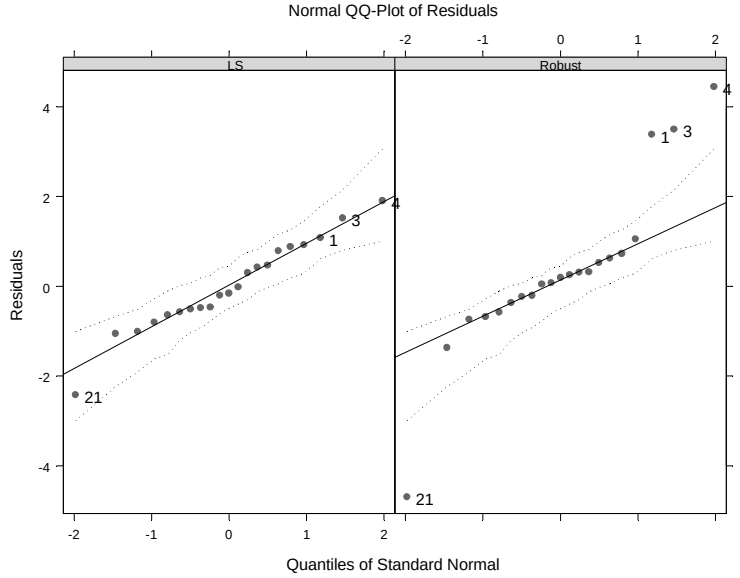

Figure 2.3: *LS and Robust Normal QQ-Plots of Residuals: stack.dat*

ignoring the outliers, has a substantially smaller standard deviation than the normal distribution fit to the LS residuals. In this sense, the robust method provides a better fit to the bulk of the data.

Probability Density Estimates of Residuals

Figure 2.4 displays the (kernel) probability density estimates for the residuals for the least squares and robust fits, and it clearly reveals the existence of outliers that adversely influence the LS fit. The story here is consistent with that provided by the normal QQ-plot comparisons: you see that density estimate of the LS residuals is much broader in the central region than that of the robust residuals, and is rather skewed and not centered on zero. The density estimate of the residuals for the robust fit is very compact and centered on zero in the central region, and exhibits two distinct bumps that indicate the presence of outliers. From this point of view, the robust fit again provides a better fit to the bulk of the data and indicates the presence of outliers.

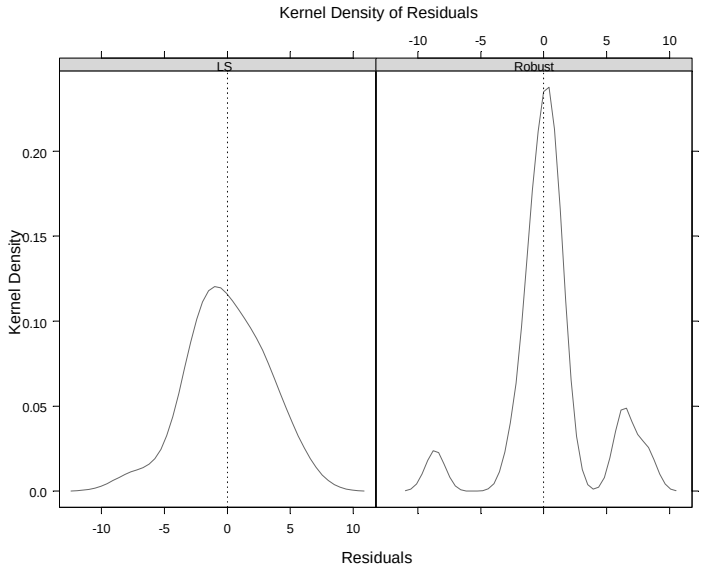

Figure 2.4: *LS and Robust Density Estimates of Residuals: stack.dat*

Standardized Residuals versus Robust Distances

A highly useful plot of scaled residuals versus robust distances of the predictor variables was invented by Rousseeuw and van Zomeren (1990). For both the LS and robust fits, the *robust distances* are the Mahalanobis distances based on a robust covariance matrix estimate for the predictor variables, as described in the section Theoretical Details. A large robust distance for a predictor variable indicates that the predictor variable has *leverage* that might exert undue influence on the fit. The scaled residuals for the LS fit are the residuals divided by the standard error of the residuals. The scaled residuals for the robust fit are the residuals divided by a robust scale estimate for the residuals, obtained as part of the robust fitting method.

The standardized residuals vs. robust distances plots for both the least squares and robust fits are shown in Figure 2.5. Following Rousseeuw and van Zomeren (1990), the horizontal dashed lines are located at +2.5 and -2.5, and the vertical line is located at the upper .975 percent point of a chi-squared distribution with p degrees of freedom, where p

= 3 in this case. Points outside the horizontal lines are regarded as residual outliers, and points to the right of the vertical line are regarded as leverage points or *x-outliers.*

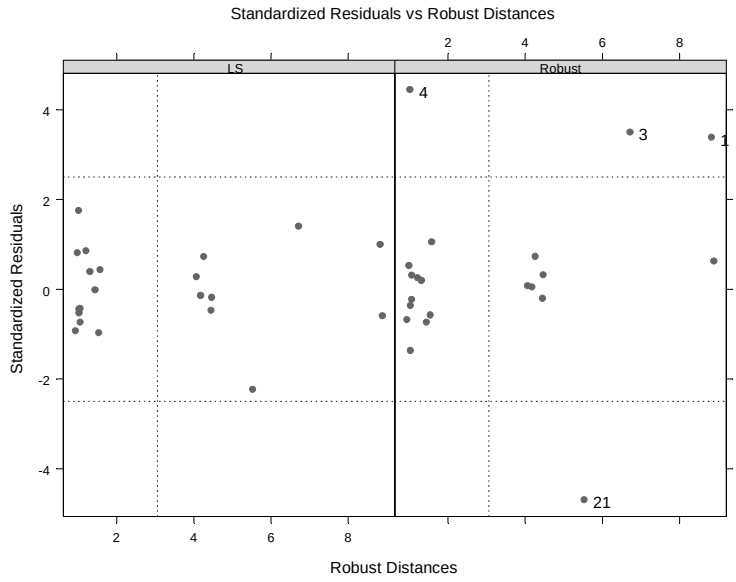

Figure 2.5: *LS and Robust Standardized Residuals vs. Robust Distances: stack.dat*

In this case the LS fit produces no residuals outliers and four x-outliers, whereas the robust fit produces four residuals outliers and four x-outliers. Three of the four x-outliers for the robust fit are also residuals outliers, while one x-outlier is not a residuals outlier. The interpretation is that three of the x-outliers have substantial influence on the LS fit, while the fourth x-outlier does not. The robust fit is not much influenced by outliers, whether they occur in the response space or the predictor space, or both.

This example illustrates the problem of outlier *masking* in least squares fits, i.e., the influence of outliers on the least squares parameter estimates distorts the parameter estimates in such a manner that the outliers can not be detected in plots of the LS residuals. The robust estimate does not suffer from this problem.

Standardized Residuals versus Index (Time)

Figure 2.6 shows the standardized residuals vs. index (time) plots for both the LS and robust fits. As in the previous plot, the standardized residuals for the LS fit are the residuals divided by the standard error of the residuals, and the standardized residuals for the robust fit are the residuals divided by a robust scale estimate for the residuals. If the response variable is a time series, then this plot is a time series plot.

From Figure 2.6, you can see that the LS fit does not reveal any outlier, while the robust fit again clearly reveals the four outliers in the data set: the first three occur at the startup of the underlying chemical process and the other one in the end when the process is shut down.

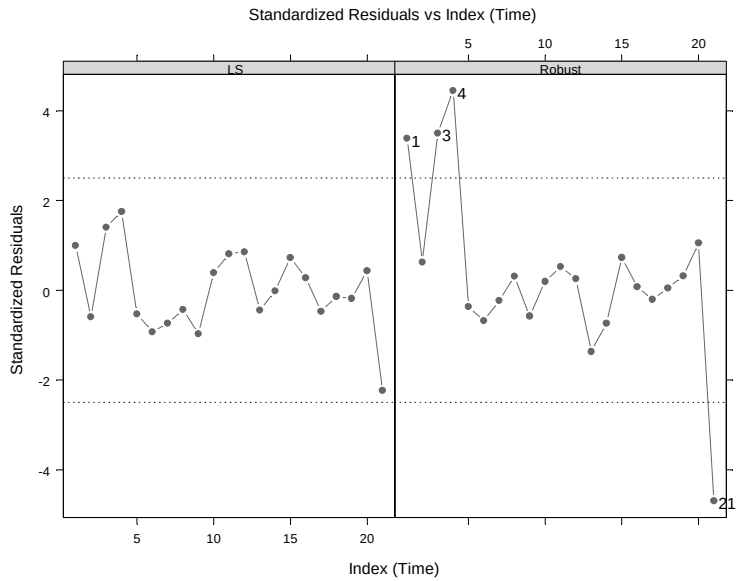

Figure 2.6: *LS and Robust Standardized Residuals vs. Index (Time): stack.dat*

The Statistics Report

The **Report** window contains the following results.

```
*** Classical and Robust Linear Regression ***

Calls:
Robust    lmRob(formula = Loss ~ Air.Flow + Water.Temp +
Acid.Conc., data = stack.dat, na.action = na.exclude)
```

```
    LS    lm(formula = Loss ~ Air.Flow + Water.Temp +
Acid.Conc., data = stack.dat, na.action = na.exclude)

Residual Statistics:
           Min      1Q  Median      3Q     Max
Robust -8.6299 -0.6713  0.3594  1.1507  8.1740
    LS -7.2377 -1.7117 -0.4551  2.3614  5.6978

Coefficients:
                      Value Std. Error  t value Pr(>|t|)
Robust (Intercept) -37.6525   5.0026    -7.5266   0.0000
    LS (Intercept) -39.9197  11.8960    -3.3557   0.0038
Robust    Air.Flow   0.7977   0.0713    11.1886   0.0000
    LS    Air.Flow   0.7156   0.1349     5.3066   0.0001
Robust  Water.Temp   0.5773   0.1755     3.2905   0.0043
    LS  Water.Temp   1.2953   0.3680     3.5196   0.0026
Robust  Acid.Conc.  -0.0671   0.0651    -1.0297   0.3176
    LS  Acid.Conc.  -0.1521   0.1563    -0.9733   0.3440

Residual Scale Estimates:
: 1.837 on 17 degrees of freedom
: 3.243 on 17 degrees of freedom

Proportion of variation in response(s) explained by
model(s):
Robust : 0.6205
    LS : 0.9136

Bias Tests for Robust Models:

Robust:
Test for Bias:
            Statistics P-value
 M-estimate       2.75    0.60
LS-estimate       2.64    0.62
```

The standard errors, the t-statistics, and the p-values of the robust coefficient estimates for the robust fit are themselves robust because they are computed using a robust covariance matrix for the

parameter estimates. The *Proportion of variation in response explained by model*, or multiple R^2, for the robust fit is a robust version of the classical least-squares R^2.

There is also a *Test for Bias* in the summary statistics provided in the **Report** window. This provides two statistical tests of the bias: the first for bias of the final M-estimate relative to a highly robust initial estimate, and the second for the bias of the LS estimate relative to the final M-estimate. In this example, the p-values for these tests are .60 and .62, indicating that for both comparisons there is little evidence of bias.

Read the section Theoretical Details to find out how these robust inference quantities are computed.