Robust Alternatives to Least Squares

by

Werner Stahel

Research Report No. 78

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Robust Alternatives to Least Squares

Werner Stahel
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

Abstract

Least Squares estimation in regression models is well known to be optimal if the random errors are normally distributed. In the presence of longer-tailed errors or outliers, there are better alternatives, called robust estimators.

Two robustness measures are based on the idea of gross errors – arbitrary observations added to a given sample. The gross error sensitivity measures the maximal effect of a single observation, and the breakdown point characterizes the fraction of gross errors that leads to arbitrarily large effects on the estimate.

M-estimators are a flexible class of estimators suitable for obtaining robust procedures with desired properties. In the regression context, they show only limited and often unsatisfactory robustness. On the other hand, desired robustness properties can only be achieved at the cost of heavy computation.

These concepts are explained on the basis of two examples from environmental sciences and molecular spectroscopy. In the latter, robust methods were essential for finding a useful solution.

1 Introduction

a **Regression and Least Squares.** Fitting a model to data is the common theme of a large part of statistical tasks in science and technology, and certainly a basic topic in metrology. The methods most often used are based on the *Least Squares* (LS) principle. They play this fundamental role because the principle is simple, they are relatively easy and cheap to compute, and they perform optimally under the assumption of normal distributions of the random quantities. However, if the normality assumption is not valid, there exist alternative methods which beat these techniques – by far in some cases.

b **Robust Statistics.** The field of Robust Statistics studies, in general, the effects of deviations from model assumptions and provides methods which still work well if the assumptions are met only approximately.

c **Goal.** This paper deals mainly with deviations from the assumed normal distribution of the random errors in multiple linear regression and introduces some basic concepts in robustness. A related, nice introduction has already been given by Vecchia and Splett (1994) in the same context. While there is some necessary overlap, a more extended overview is given here, and two further applications illustrate the success of robust procedures. An outlook on other topics in robustness concludes the paper.

2 Multiple Linear Regression

a **Notation.** The general model for (multiple) linear regression will be denoted as

$$Y_i = \sum_j x_i^{(j)} \beta_j + E_i = \underline{x}_i^T \underline{\beta} + E_i \ . \tag{2.a}$$

Here, Y_i is the value of the *target variable* for the ith observation, and $\underline{x}_i = [x_i^{(1)}, ..., x_i^{(m)}]^T$ is the (column) vector of respective values for the *explanatory variables* (or design variables or carriers). Usually, one explanatory variable $x^{(j)}$ will be identically 1, such that the corresponding coefficient β_j acts as the "intercept". E_i is the random error. The usual model assumes that the E_i are independently, identically and normally distributed, $E_i \sim G_r = N(0, \sigma^2)$. The parameters are thus the coefficients β_j, collected into the vector $\underline{\beta}$, and the scale σ of the errors.

b **Tunnel measurements of air pollution.** Road traffic is an important source of air pollution. In order to quantify its contributions, it is necessary to obtain "emission factors", that is, to estimate the average quantity of a given pollutant like NO_x which a vehicle emits per kilometer. In a one-way tunnel, the total pollution caused by traffic in 15 minutes, say, can be measured together with the number of vehicles, and an average emission factor Y_i can be calculated. If the vehicles are classified into different categories, an emission factor for each category can be obtained: Figure 2.c shows average emission factor values together with the proportion x_i of heavy duty vehicles for a study in a tunnel near Zurich (see Staehelin, Schläpfer, Bürgin, Steinemann, Schneider, Brunner, Bäumle, Meier, Zahner, Keiser, Keller and Stahel, 1995 for details). Since heavy duty vehicles emit more NO_x than passenger cars, Y_i increases with x_i. In fact, since the total pollution is the sum of the contributions from light and heavy duty vehicles, it is straightforward to postulate the simple regression model

$$Y_i = \alpha(1 - x_i) + \gamma x_i + E_i = \alpha + \beta x_i + E_i \ ,$$

where α is the emission factor for light duty vehicles and $\gamma = \alpha + \beta$, for heavy ones.

c A Least Squares fit is shown in Figure 2.c. Clearly, the distribution of the errors is skewed and depends on the explanatory variable x. One wonders if a different straight line might give more adequate results for the two important parameters α and γ.

d **An example from Molecular Spectroscopy.** Insight into the physical structure of atoms and molecules can be obtained by assessing the pattern of energy levels associated with their quantum mechanical states. Experiments associated with this topic form the scope of *atomic and molecular spectroscopy.* The energy levels are obtained by measuring the energies connected to the transitions of a molecule between different quantum mechanical states. The corresponding model is extremely simple and reads

$$Y_{\ell u} = -\theta_\ell + \theta'_u + E_{\ell u}, \tag{2.d}$$

where $Y_{\ell u}$ is the energy released when the molecule falls from the *upper state* u to the *lower state* ℓ, and $E_{\ell u}$ is the random measurement error. The quantities θ_ℓ and θ'_u are the unknown energy levels of the two states. Since only differences are measured, the levels can be determined only up to an additive constant. The energy level of the lowest θ_ℓ is therefore set to zero.

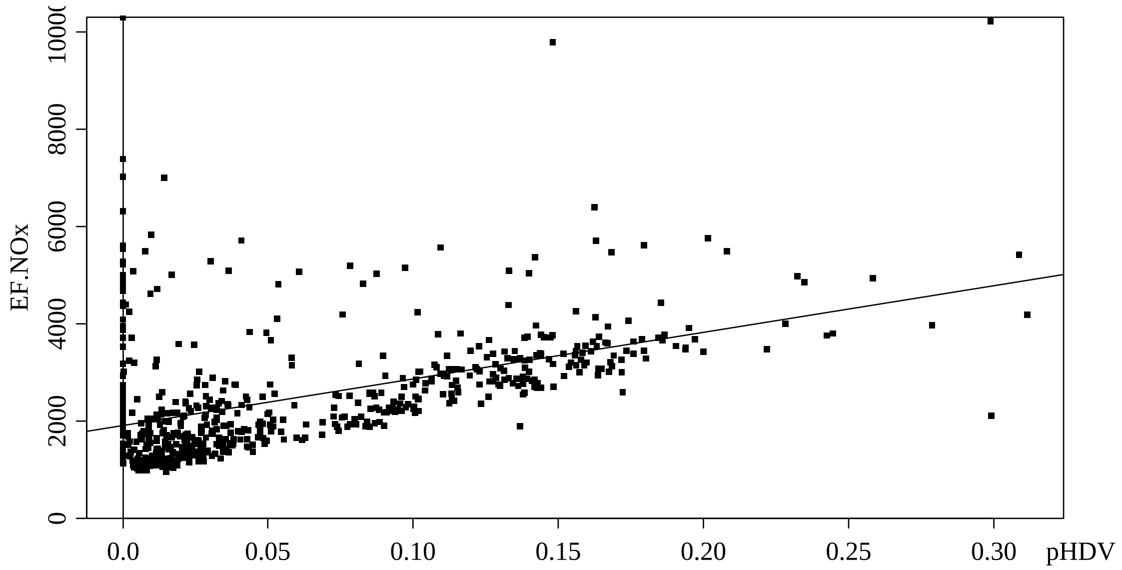

Figure 2.c: LS regression of the average emission factor against the proportion of Heavy Duty Vehicles.

e The data set which will be analyzed here consists of 659 measured transitions involving 249 quantum mechanical states of the tritium molecule T_2, which can be separated into 54 lower and 195 upper states. (See Stahel, Ruckstuhl, Senn and Dressler (1994) for a more complete description.)

f The model can be rewritten in the form of a multiple regression model, $Y_i = \sum_j x_i^{(j)} \theta_j + E_i$ using

$$x_i^{(j)} = \begin{cases} 1 & \text{if } j \text{ is the upper state for the } i\text{th observation,} \\ -1 & \text{if } j \text{ is the lower state for the } i\text{th observation,} \\ 0 & \text{otherwise.} \end{cases}$$

Each $\underline{x}_i$ thus contains a 1, a −1, and 246 zeros (except for those transitions involving the lowest state). Such a situation is called a *sparse design.*

g The measurement errors E_i are extremely small. Since molecules cannot be forced to switch between selected states, highly sophisticated methods are used to determine a pair $[u_i, \ell_i]$ for each measurement Y_i, and this process may partly fail. Mistakes lead to outliers of arbitrary size with respect to the given model. We shall show that the LS method leads to useless results, while a robust alternative yields adequate estimates of the parameters and allows for a generally reliable identification of the falsely identified pairs of states.

h **Are random errors normal?** Even though the mechanism leading to gross errors is very special in this example, practitioners know that *outliers* do occur. Usual practice suggests to search for outliers informally or by formal tests, find the reason for the blunders, correct or remove the observations in question, and use LS on the rest.

Metrology is often concerned with high-quality data, and outliers rarely cause a real problem. On the other hand, random measurement errors are usually not normally distributed, but they show *longer tails.* This statement is difficult to verify, since some of the common goodness-of-fit tests have resticted power and samples are too small. In large samples of high-quality data, long tails become evident (see Hampel, Ronchetti, Rousseeuw and Stahel, 1986, p. 22).

i **Least Squares.** It is generally known that

• the Gauss-Markov theorem proves that the LS estimators have minimal variance also for non-normal errors (as long as the variance of the E_i exists) within the class of *linear* estimators (linear in the observations Y_i). – Robust alternatives are non-linear (as is outlier rejection followed by LS). They typically show much *smaller standard errors* (are statistically more efficient) for long-tailed error distributions. For strictly normal random errors, they are necessarily less precise than LS, but this loss can be kept small. In terms of an insurance, a small premium (efficiency loss for normal errors) is sufficient for a large protection (against efficiency loss for longer-tailed distributions).

• t- and F-tests based on LS are adequate as long as the errors are approximately normal or there are many more observations than parameters in a reasonably balanced design. – While this is true for their validity (error of the first kind), alternatives usually have better power (smaller probability of errors of the second kind) and thus lead to shorter confidence intervals for long-tailed error distributions.

There is a large variety of alternative, robust methods (see Hampel et al., 1986, Sec. 2.3b and Stahel and Weisberg, 1991). Here, we want to present only two basic classes.

3 M- and GM-estimators

a **Definition.** A first class of alternative estimators is obtained by slightly changing the criterion minimized by Least Squares: An *M-estimator* for model (2.a) is the value $\underline{\beta}$ which minimizes

$$\sum\nolimits_i \rho(r_i/\widehat{\sigma}) \ , \qquad (3.a)$$

where $\widehat{\sigma}$ is a suitable estimator of the spread of the random errors, and ρ is a suitable scalar function. In practice, $\widehat{\sigma}$ usually needs to be obtained from the residuals, which in turn depend on β. If $\rho(r) = r^2$, the LS estimator results.

b What functions ρ are suitable? If we knew that the errors followed a known long-tailed distribution with density f, we would be led to using the corresponding maximum likelihood estimator, since it is asymptotically most efficient. In fact, this is an M-estimator with ρ-function $\rho(r) = -\log f(r)$. If f is the normal density, the LS estimate is obtained. Figure 3.b shows a few useful ρ-functions and their derivatives $\psi = \rho'$.

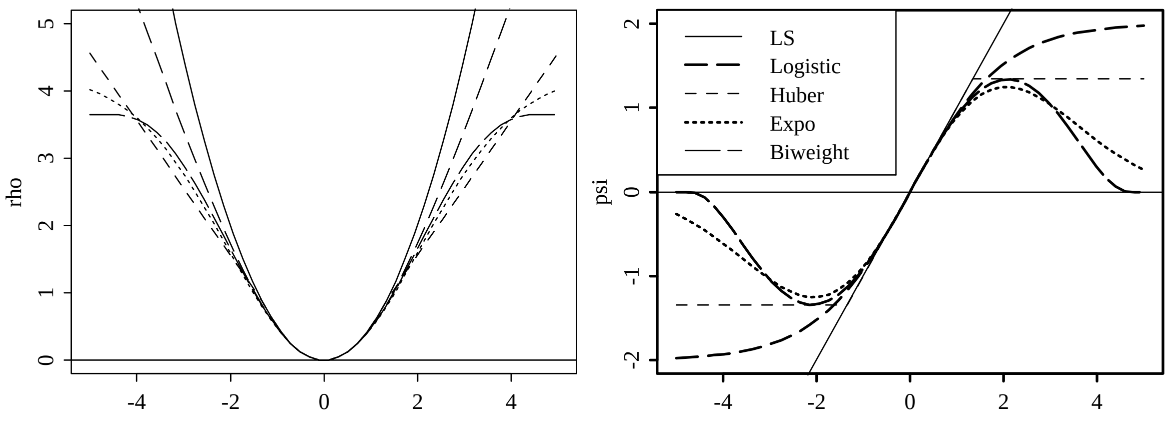

Figure 3.b: Five useful ρ-functions (left panel) and their derivative ψ (right panel)

c In order to solve the problem of minimizing (3.a), the derivatives are set equal to zero, which leads to the so-called normal equations

$$\sum_i \psi([Y_i - \underline{x}_i^T \underline{\beta}]/\widehat{\sigma})\; x_i^{(j)} = 0 \tag{3.c}$$

for all j, with $\psi = \rho'$. This helps to devise both theoretical results and algorithms. As we shall see shortly, the function ψ plays a more prominent role than ρ.

d A more general class allows the ψ-function to depend on $\underline{x}$: A *GM-estimator* is defined as the solution of

$$\sum_i \eta(\underline{x}_i, [Y_i - \underline{x}_i^T \underline{\beta}]/\widehat{\sigma})\; x_i^{(j)} = 0\ , \tag{3.d}$$

where η is usually of the form $\eta(\underline{x}, r) = \eta(\|\boldsymbol{A}\underline{x}\|, r)$ for a suitable matrix $\boldsymbol{A}$ which depends on the design $\underline{x}_1, ..., \underline{x}_n$.

e **Their Use.** In applications, the differences between LS and robust (point) estimates are often rather small. This relates to the general fact that even considerable efficiency losses are difficult to grasp on the basis of single cases.

In practice therefore, robust estimates are often used only as a *diagnostic tool*: Both the LS and a robust estimate is obtained. If they show reasonable agreement, the LS results are reported, because they are not questioned by referees and readers. If larger discrepancies appear, they can often be traced down to a few extreme observations. These outliers may then be rejected, and the LS results for the remaining observations are reported. Clearly, tests and confidence intervals based on such a streamlined sample are, in principle, invalid. The strategy is a version of formal or informal *outlier rejection* procedures. Rejection rules are usually based on a LS fit for the whole data set. It can be shown that rejection of outliers based on a robust fit instead is more reliable.

f **Results.** For the *tunnel example*, the results of a robust fit are compared to the LS results in Figure 3.f. In this study, a reason could be found for the many observations with unexpectedly high Y values: When there was little traffic, the air flow was not strong and steady enough for the calculation of meaningful average emission factors Y_i for 15 minutes. The LS estimate (and the robust one, not shown) for the observations with sufficient air velocity agrees reasonably with the robust estimate based on all the data, but is rather different from the corresponding LS estimate.

Of course, in such a case, where reasons for outliers are available and the model can easily be visualized, any reasonable analysis will find a reasonable result. Robust estimators are needed more urgently in more complex situations.

g *In the Molecular Spectroscopy example*, we applied both LS and the M-estimator based on the logistic ψ-function, using the MAD of the residuals as a scale estimate. The LS residuals show a clearly long-tailed distribution with some 4% extreme outliers (Figure 3.g). Their median absolute value (MAD) is 80, but the measurement error is known to have a standard deviation of about 0.02 units (wave numbers, cm^{-1})! As a consequence of the sparse design, there are 49 exactly zero residuals. All the other residuals are larger than 0.1! For the robust fit, only about 1/3 of the residuals are larger than this value, the MAD being 0.021 (corresponding to a standard deviation of 0.03). They can be regarded as cases of wrong identification of lower or upper quantum mechanical states. Furthermore, the estimated energy levels should show a certain structure with respect to their quantum numbers according to physical theory. The pattern of the robust estimates fits well in contrast to the LS estimates. See Stahel et al. (1994) for details.

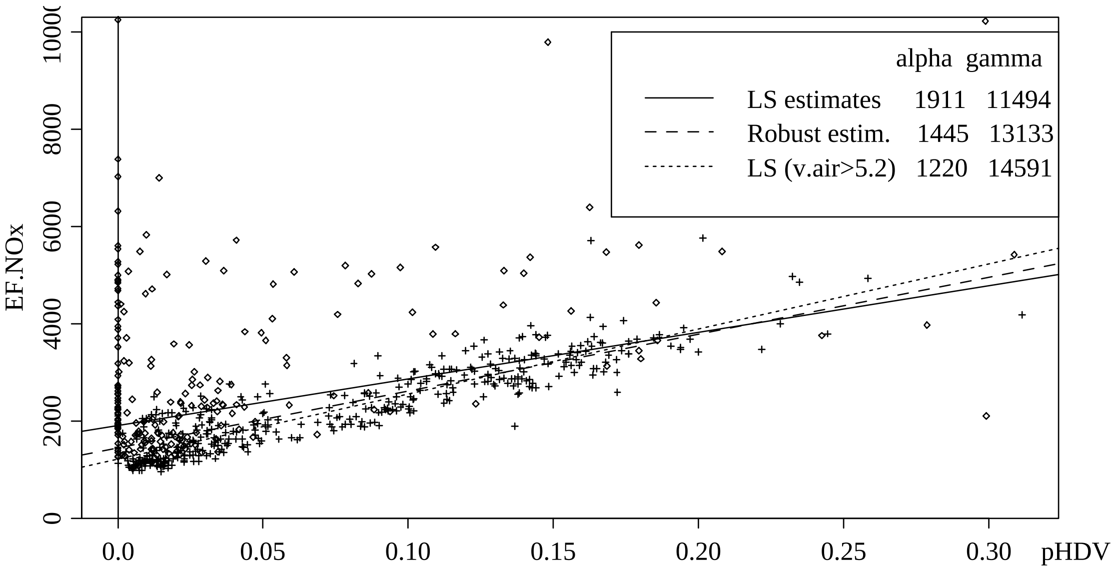

Figure 3.f: Three regression lines for the tunnel data: LS, robust, and LS applied to the data with enough air flow. The latter are marked with crosses, the rest, with diamonds

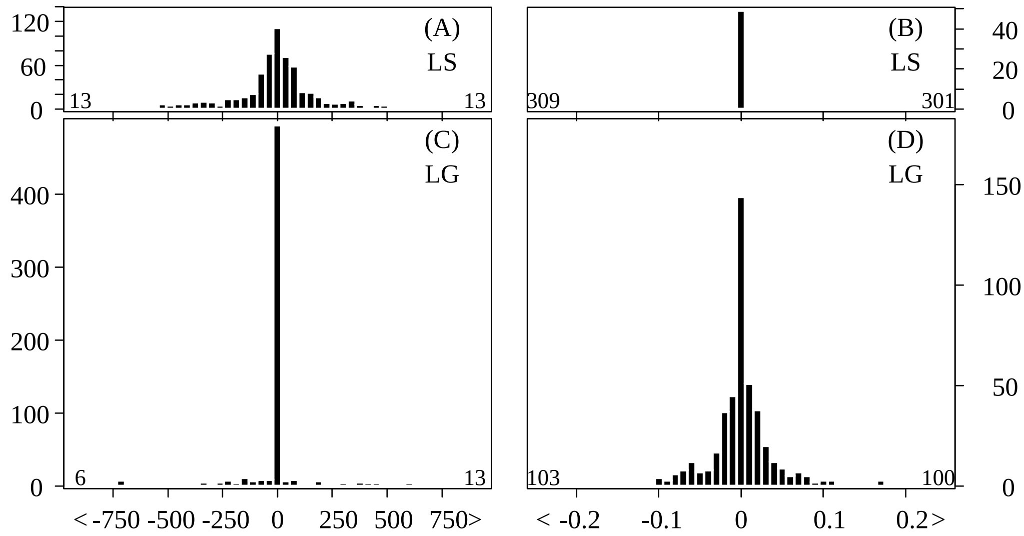

Figure 3.g: Distribution of residuals from LS (A,B) and a robust fit (C,D). The histograms show the residual in 2 different ranges (A,C and B,D). The numbers of residuals falling outside the ranges are noted next to the histograms.

4 Robustness Measures

a In order to compare the robust performance of estimators, it is useful to introduce criteria which describe their reaction to the violation of model assumptions. A procedure is *(locally) robust if small deviations from the assumptions lead to small changes in the results.* It is *globally robust* or *resistant*, if the results remain reasonable even for rather large deviations from the assumptions.

b For studying the local behavior, we calculate the change in the estimate caused by an observation at $[\underline{x}, y]$ added to the sample $[\underline{x}_1, y_1], ..., [\underline{x}_n, y_n]$ and obtain the Sensitivity Curve (or function)

$$\mathrm{SC}(\underline{x}, y; [\underline{x}_1, y_1], ..., [\underline{x}_n, y_n]) = n\left(\widehat{\underline{\beta}}([\underline{x}_1, y_1], ..., [\underline{x}_n, y_n], [\underline{x}, y]) - \widehat{\underline{\beta}}([\underline{x}_1, y_1], ..., [\underline{x}_n, y_n])\right) .$$

Figure 4.b shows the effect of adding an observation with $x = 0.34$ as a function of a variable emission factor y for the M-estimators of the slope corresponding to the ρ- and ψ-functions in Figure 3.b. They have a similar shape as the ψ-functions.

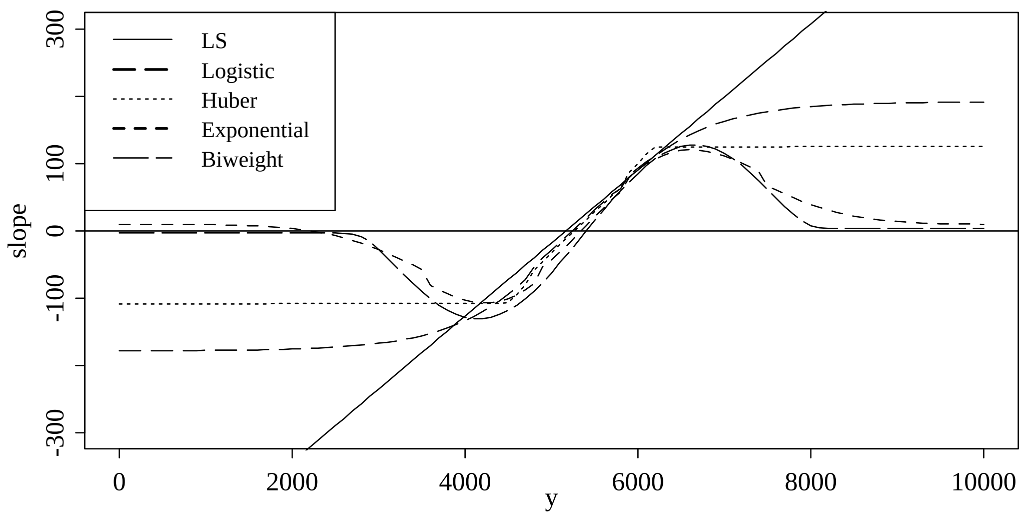

Figure 4.b: Sensitivity Curves: Estimated slope as a function of the y coordinate of an additional observation at $x = 0.34$ in the tunnel example

Since the SC is vector valued like $\widehat{\underline{\beta}}$, a suitable norm $\|.\|$ is used to quantify the effect, and the maximum norm,

$$\max \|\mathrm{SC}(.,.; [\underline{x}_1, y_1], ..., [\underline{x}_n, y_n])\|$$

is a useful measure of robustness.

For theoretical considerations, a measure that depends on the assumed model rather than a sample is more practical. The Influence Function corresponds in this sense to the Sensitivity Curve. The maximum of its norm is the *Gross Error Sensitivity*. It measures the worst effect that a "small contamination" (an additional observation) can have on the estimate, and is therefore a useful measure of local robustness.

c For M- (or GM-) estimators, the Influence Function is exactly proportional to the ψ- (or η-) function and to a suitable linear transformation of $\underline{x}$,

$$\mathrm{IF} = \boldsymbol{M}\underline{x}\ \psi(y - \underline{x}^T\underline{\beta}) \quad (= \boldsymbol{M}\underline{x}\ \eta(\underline{x}, y - \underline{x}^T\underline{\beta})\ )$$

for a suitable matrix $\boldsymbol{M}$. If both of them are bounded, the Gross Error Sensitivity of M-estimators is finite – in contrast to LS. For appropriate choices of η, GM-estimators have a bounded Influence Function for any design. They are called *Bounded-Influence Estimators*, see Hampel et al. (1986, Chap. 6).

d **Global Robustness.** In order to study a "global" aspect of robustness, we now play the game of altering a fixed number m of observations rather than adding a single one: Let $[\underline{x}_1^*, y_1^*], ..., [\underline{x}_n^*, y_n^*]$ be a sample for which $n - m$ observations are identical to observations in the given sample $[\underline{x}_1, y_1], ..., [\underline{x}_n, y_n]$, while the other m observations are arbitrary. In the tunnel example, we may ask: What happens if we change the last $m = 1, 2, 3, ...$ observations to be equal to the outlier $[x = 0.3, y = 10,215]$ in the upper right hand corner of Figure 2.c? Figure 4.d shows the LS and robust estimates of the two quantities of interest, α and $\gamma = \alpha + \beta$, as a function of the number m of faked data.

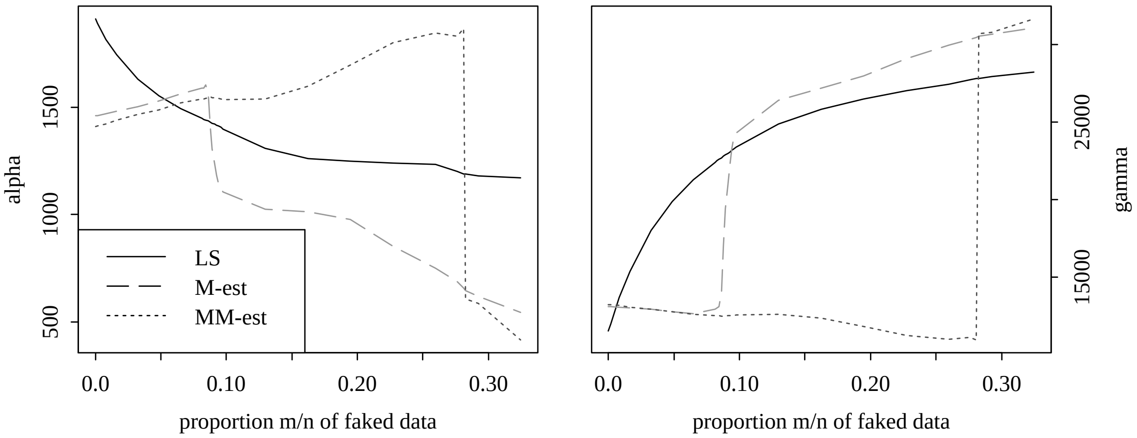

Figure 4.d: Estimates $\widehat{\alpha}$ and $\widehat{\gamma} = \widehat{\alpha} + \widehat{\beta}$ as a function of the number m of observations [pHDV=0.3,y=10,215] added

e Now examine the worst effect of such alterations on the estimator! If any component of $\underline{\widehat{\beta}}$ can be made arbitrarily large by "wicked" choice of the faked observations, we say that *the estimator breaks down.* The largest proportion $(m - 1)/n$ for which this does not occur is the *finite-sample breakdown point.* Usually, it is essentially independent of the sample size *and* of the data values of the "good" observations $[\underline{x}_1, y_1], ..., [\underline{x}_n, y_n]$ (unless they show a special configuration). As with local robustness, there is an "asymptotic" version of the concept.

f In our *Molecular Spectroscopy example*, it seems clear at first glance that the robust procedure has worked, and the estimator has not broken down. Formally, however, this is not true, since some components of the parameter vector appear wrong in the light of their clear mismatch with the physical theory mentioned above. The reason is easily understood: Some of the energy levels θ_u of upper quantum mechanical states are connected only to very few observed transition energies $Y_{\ell u}$. If half of these are modified appropriately, the estimate is misled to assume an arbitrary wrong value. This phenomenon, which is a consequence of the sparse design, is called *partial breakdown*, see Hampel (1987) and Ruckstuhl (1996).

g **The Breakdown of GM-Estimators** was found to be bounded by $1/p$ (in case of random explanatory variables $\underline{x}_i$), where p is the number of parameters β_j. That is, for 7 explanatory variables, any M-estimator breaks down if we are allowed to manipulate 1/7=14% freely. Now, this seems to be a rather high fraction of gross errors that can be tolerated. Note, however, that an observation $[\underline{x}_i, y_i]$ becomes useless in principle if only one of its components is grossly wrong. It only takes a fraction of 2% of false values to make the probability of at least 1 false in 8 values equal to 1/7 – and 2% is a rather low fraction of gross errors in practice! It has been argued that the concept of breakdown is unrealistic and pessimistic since "nature" is usually not so wicked as to select the worst spot for gross errors. Nevertheless, the situation is not satisfactory.

h The bound $1/p$ applies to GM-estimators defined as the solution of the normal equations. If there is more than one solution, they usually correspond to local minima of the corresponding criterion (3.a). Then, the result says that at least one of the solutions breaks down for contamination fractions $\geq 1/p$. There may still be an adequate solution, and the next section describes how to find it.

5 S- and MM-estimators

a **S-estimators.** A second class of estimators, besides the M-estimators, has been found to show good breakdown properties. The idea of Least Squares is to minimize the variance of the residuals. Since the variance is a non-robust measure of the spread, it is natural to replace it by a robust measure $\widehat{\sigma}(r_1, ..., r_n)$. Thus, an *S-estimator* of regression is given by

$$\underline{\widehat{\beta}} = \arg\ \min_{\underline{\beta}}\ \widehat{\sigma}(y_1 - \underline{x}_1^T \underline{\beta}, ..., y_n - \underline{x}_n^T \underline{\beta})$$

The most well-known instance is the Least Median of Squares method of Hampel and Rousseeuw (Rousseeuw, 1984), which minimizes median(r_i^2).

b The function to be minimized is not convex for robust members of this class. Therefore, calculation of the minimizing $\underline{\beta}$ is extremely difficult for more than 2- or 3-dimensional arguments $\underline{\beta}$. A stochastic algorithm is used which finds the solution with some desired probability (Rousseeuw and Leroy, 1987). If this number should be high for a high-dimensional problem, the computer time needed is exorbitant. In practice, a compromise is needed, and the high protection from breakdown often remains a hope.

A second handicap of these estimators is their statistical inefficiency: Estimators with a high breakdown point are much more variable than good M-estimators.

c **MM-estimators.** Both high breakdown point and high efficiency can be obtained by combining S- and M-estimation, leading to MM-estimators. These proceed in two steps:

Step 1: Calculate an S-estimate $\underline{\widehat{\beta}}_0$ (by the probabilistic algorithm mentioned above, followed by a local minimization).

Step 2: Find the local minimum of the criterion 3.a of an M-estimator which is "nearest" to $\underline{\widehat{\beta}}_0$ by performing an iterative algorithm for solving the normal equations 3.c, starting from $\underline{\widehat{\beta}}_0$.

A detailed proposal is given by Yohai, Zamar and Stahel (1991) based on Yohai (1987).

6 Tests and Confidence Regions

a **Asymptotic Inference.** Point estimates are of limited practical use unless they are accompanied by testing procedures and confidence intervals. To this end, the distributions of the estimates need to be known. For M-estimators, asymptotic results are well-known and quite simple to understand: They are asymptotically normal with covariance matrix $\sigma^2 \tau \boldsymbol{C}$, where $\boldsymbol{C} = (\sum_i \underline{x}_i \underline{x}_i^T)/n$ and σ is the spread (standard deviation) of the error distribution G_r as for LS, but there appears a "correction factor" $\tau = \int \psi(r/\sigma)^2 dG_r(r) / (\int \psi'(r/\sigma) dG_r(r))^2$.

b The covariance matrix of the estimates can thus be estimated by $\widehat{\boldsymbol{V}} = (\widehat{\sigma}^2/n)\widehat{\tau}\widehat{\boldsymbol{C}}$, where

$$\widehat{\tau} = \frac{\text{ave}_i[\psi(\widetilde{r}_i)^2]}{[\text{ave}_i\psi'(\widetilde{r}_i)]^2}\,, \qquad \widehat{\boldsymbol{C}} = \frac{\text{ave}_i\underline{x}_i\underline{x}_i^T w_i}{\text{ave}_i w_i}\,, \tag{6.b}$$

$\widetilde{r}_i$ are suitably standardized residuals, like $\widetilde{r}_i = (y_i - \underline{x}_i\widehat{\underline{\beta}})/\widehat{\sigma}$, and $w_i = 1$ or, alternatively, $w_i = \psi(\widetilde{r}_i)/\widetilde{r}_i$ as proposed by Yohai et al. (1991). It is clear that for all of these factors of $\widehat{\boldsymbol{V}}$ alternative, asymptotically equivalent estimators can be proposed.

c For finite samples, the asymptotic approximation may be insufficient, and the alternatives may behave rather differently. In his fundamental paper, Huber (1972) introduced a further "finite sample correction" factor. Very little is known about the merits of such alternatives, including bootstrap procedures.

d **Is Inference for LS exact?** The need for asymptotic approximations appears to be a serious drawback of robust estimators. The distribution of LS is well-known to be exact for finite samples – *if the distribution of the errors is normal* and the variances equal (or correctly specified). If this myth is left behind, there is no reason to believe that we are better off. Rather, there are arguments which suggest that *the normal distribution approximates better for robust estimates than for LS*! (Assume a stable law with shape parameter $\alpha < 2$ for E_i. Then, the LS estimate has a stable distribution, too, and its convergence rate is slower than $\sqrt{n}$, whereas robust estimates converge at the usual speed to a normal distribution.)

7 Outlook on other Robustness Problems

In this paper, we have explained and illustrated some concepts and methods of robust statistics for the model of multiple linear regression. Similar ideas can be and are applied in other areas:

a **Heteroskedasticity.** The assumption of equal variance for all the random errors (or a known pattern of variances) can also be violated, and procedures which are robust in this sense have been studied. To some extent, the robust estimators discussed above solve this problem, too, since an E_i with a large $\text{var}(E_i)$ will, with high probability, appear as an outlier, and its influence will be reduced.

b **Non-Linearity.** If the assumed linear model $\underline{x}^T\underline{\beta}$ for the conditional expectation $E(Y|\underline{x})$ is wrong, the problem is of a quite different nature. It is then unclear what the "ideal" result would be, that is, to which parameter values an estimator should converge. The use of robust estimators can lead to worse results in such a context: Consider the situation of a single explanatory variable for which the true model should include a quadratic term, but the postulated model is linear. As long as we are interested in an approximate description or in interpolation, all estimators may have their range where they lead to most adequate results. As soon as extrapolations are needed, robust estimators may be worse than LS: They allow for some points with a bad fit, and these will often be the observations with extreme $\underline{x}$'s in such a situation. The fit beyond these extreme points may even be worse.

c **Dependence Structure.** For tests and confidence intervals to possess the assumed error or coverage probabilities, independence of the random errors is essential (or, more generally, the assumed correlation structure must be correct). Deviations from this assumption are rather difficult to study in any generality, because there are too many types. Some work has been done for the case of a sequential structure in time. If auto-correlations of the random errors are plausible, there is a classical way of dealing with them: First, obtain a preliminary fit using the independence assumption. Then, derive an AR or ARMA time series model for the residuals. Finally, use the general least squares theory (or a version of it which applies to "distribution-robust" estimation) to obtain improved estimates and/or correct tests and confidence intervals.

d The most dangerous deviations from independence, however, is a *long-range dependence*, which cannot be adequately described by the ARMA models. Such correlations can have disastrous effects for longer series. What's more, they have been shown to occur frequently in real data sets (Hampel et al., 1986, Chap. 8.1) – even in series of chemical and physical measurements where there is no plausible reason for dependence. Again, the only solution seems to consist of modeling the correlation structure and using the corresponding inference procedures. Künsch, Beran and Hampel (1993) study these questions and find, reassuringly, that if contrasts between groups or slope parameters in regression models are estimated and the design variables are in some sense orthogonal to time, the levels of tests are correct.

e **Other models.** M-estimation has been extended to *non-linear* and to *generalized linear models*. Random effects and mixed models have only recently been treated. The estimation of a "scatter matrix" (covariance matrix) forms the basis for multivariate statistics and has received considerable attention in the robustness literature. The approach of Hampel et al. (1986) can be applied to any parametric model for independent observations. For time series, the concepts and methods need some extension, see Hampel et al. (1986, Chap. 8.3).

f A taxonomy of these and other research directions is given in Stahel (1991).

8 Software

a Robust methods are not yet widely available in software packages apart from robust measures of location and scale for a univariate sample.

b The system which provides the largest coverage is the package S and its extension S-Plus (Becker, Chambers and Wilks, 1988, Chambers and Hastie, 1992, Venables and Ripley, 1994, Statistical Sciences, 1988,1993). It implements a powerful language with many useful statistical, graphical and mathematical functions and allows for manipulating data and combining it with results of analyses in very flexible ways. Routines written in C or Fortran can be incorporated. It has become a language of communication among statisticians, and a large collection of contributed procedures, called statlib, is available on the internet.

The standard S/S-Plus package presently includes several robust estimation procedures for M-estimation (rreg), L_1-regression (l1fit), the two S-estimators Least Median of Squares (lmsfit) and Least Trimmed Squares (ltsfit), and finally a kind of robust M-estimation method (robust) for ordinary and generalized linear models (which, however, is not really robust since the scale is estimated non-robustly). Tests are provided only for the last procedure. They are poorly documented.

In the "statlib" collection, there is a function for robust M-estimators (rlm) which also provides tests in the library MASS by Venables and Ripley (1994) and a function (mmreg) for the MM-estimator proposed by Yohai et al. (1991) (cf. 5.c).

c A large collection of routines implementing many robust methods has been developed by Marazzi (1993) and is called ROBETH. There is an interface which allows to use all these routines under the S system. Its use requires some expertise, however.

d Standard packages do not provide robust regression methods. SAS documents how the procedure NLIN can be used to provide M-estimates, but the standard errors and tests given in the output are inadequate.

9 Concluding Remarks

a It has been demonstrated that Least Squares methods are rather unsafe, since they give non-optimal and sometimes misleading results in situations which often do occur in practice. If diagnostics are used to verify the assumptions, it is often (but not always) possible to detect serious violations, but remedies in the form of recipes have unknown statistical properties. Therefore, it is clearly safer to use robust estimators from the start.

Such a statement is often misunderstood. It may seem that the use of robust methods avoids the need for deriving an adequate model. Of course, this would be a fundamentally wrong attitude: There is no hope to obtain good estimates and correct inference for inadequate models! In contrast:

- Suitable robust fitting methods show deviations from the assumptions more clearly than LS, and therefore chances that a suitable model is found increase. (I must admit that this statement is subject to an ongoing discussion.)
- No model ever is exactly "correct" – except for simulation studies and similar artificial situations. The model should be a good approximation, and if this is assumed, robust methods are better than "optimal" classical methods.

b Why, then, are robust methods not more popular?

- Deviations, even if only the distribution is considered, come in several flavors (few or many outliers, in the E_i only or in $\underline{x}_i$, too), which lead to different corresponding procedures.
- There have been many approaches to these problems. Finding new ramifications is more rewarding in an academic environment than "consolidating" a specific approach.
- Most research has been devoted to estimation. Users want tests and confidence intervals which go along with them. Asymptotic results have been available for 25 years now. But, in contrast to other fields like generalized linear models, robustness experts have hesitated to advocate corresponding inference procedures.
- Software package writers need one or two "general purpose" estimation and inference procedures, which should be generally accepted by the "robustness community" and can be presented in a form which is understandable for common users. They need a partner with enough knowledge and energy who can turn a statistical method into a "consumer product". S/S-Plus has made a start here.
- Convincing examples should make the need for such methods clear to the potential users. Rousseeuw (1996) has collected a number of engineering applications. We hope that this paper also contributes to this end.

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The new S Language; A Programming Environment for Data Analysis and Graphics*, Wadsworth & Brooks/Cole, Pacific Grove.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole, Pacific Grove.

Hampel, F. R. (1987). Some problems in statistics, *in* Y. Prohorov and V. V. Sazonov (eds), *Proc. 1st World Congress of the Bernoulli Soc.*, Vol. 2, VNU Science Press, Utrecht, pp. 253–256.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, N. Y.

Huber, P. J. (1972). Robust statistics: A review, *Ann. Math. Statist.* **43**: 1041–1067.

Künsch, H. R., Beran, J. and Hampel, F. R. (1993). Contrasts under long-range correlations, *Ann. Statist.* **21**(2): 943–964.

Marazzi, A. (1993). *Algorithms, Routines, and S Functions for Robust Statistics*, Wadsworth, Inc., Belmont, California.

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

Rousseeuw, P. J. (1996). New applications of robust statistics, *Compstat*, Barcelona.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression & Outlier Detection*, Wiley, N. Y.

Ruckstuhl, A. F. (1996). Partial breakdown in two-factor models, *J. Statist. Planning Inference*.

Staehelin, J., Schläpfer, K., Bürgin, T., Steinemann, U., Schneider, S., Brunner, D., Bäumle, M., Meier, M., Zahner, C., Keiser, S., Keller, C. and Stahel, W. A. (1995). Emission factors from road traffic from a tunnel study (Gubrist tunnel, Switzerland). Part i: Concept and first results, *Science Total Environment* **169**: 141–147.

Stahel, W. A. (1991). Research directions in robust statistics, *in* [see next item] (ed.), *Directions in Robust Statistics and Diagnostics*.

Stahel, W. A. and Weisberg, S. (1991). *Directions in Robust Statistics and Diagnostics*, Springer, N. Y.

Stahel, W. A., Ruckstuhl, A. F., Senn, P. and Dressler, K. (1994). Robust estimation in the analysis of complex molecular spectra, *J. Am. Statist. Assoc.* **89**(427): 788–795.

Statistical Sciences (1988,1993). *S-PLUS Guide to Statistical and Mathematical Analysis*, StatSci, MathSoft, Seattle.

Vecchia, D. F. and Splett, J. D. (1994). Outlier-resistant methods for estimation and model fitting, *in* P. Ciarlini, M. G. Cox, R. Monaco and F. Pavese (eds), *Advanced Mathematical Tools in Metrology*, World Scientific, Singapore, pp. 143–154.

Venables, W. N. and Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus*, Springer, Berlin.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression, *Ann. Statist.* **15**(2): 642–656.

Yohai, V., Zamar, R. and Stahel, W. A. (1991). A procedure for robust estimation and inference in linear regression.