

Robust Fitting of Parametric Models Based on M-Estimation

Andreas Ruckstuhl *

IDP Institute of Data Analysis and Process Design
ZHAW Zurich University of Applied Sciences in Winterthur

Version 2024[†]

*Email Address: Andreas.Ruckstuhl@zhaw.ch; Internet: <http://www.idp.zhaw.ch>

[†]The author thanks Amanda Strong for her help in translating the original German version into English.

Contents

1. Basic Concepts	1
1.1. The Regression Model and the Outlier Problem	1
1.2. Measuring Robustness	3
1.3. M-Estimation	7
1.4. Inference with M-Estimators	9
2. Linear Regression	12
2.1. Regression M-Estimation	12
2.2. Example from Molecular Spectroscopy	14
2.3. General Regression M-Estimation	17
2.4. Robust Regression MM-Estimation	19
2.5. Robust Inference and Variable Selection	23
3. Generalized Linear Models	29
3.1. Unified Model Formulation	29
3.2. Robust Estimating Procedure	30
4. Multivariate Analysis	34
4.1. Robust Estimation of the Covariance Matrix	34
4.2. Principal Component Analysis	40
4.3. Linear Discriminant Analysis	42
5. Baseline Removal Using Robust Local Regression	46
5.1. A Motivating Example	46
5.2. Local Regression	46
5.3. Baseline Removal	50
6. Some Closing Comments	52
6.1. General	52
6.2. Statistics Programs	52
6.3. Literature	54
A. Appendix	55
A.1. Some Thoughts on the Location Model	55
A.2. Calculation of Regression M-Estimations	56
A.3. More Regression Estimators with High Breakdown Points	58

Objectives

The block course *Robust Statistics* in the post-graduated course (Weiterbildungslehrgang WBL) in applied statistics at the ETH Zürich should

1. introduce problems where robust procedures are advantageous,
2. explain the basic idea of robust methods for linear models, and
3. introduce how the statistical software R can contribute to the solution of concrete problems.

1 Basic Concepts

1.1 The Regression Model and the Outlier Problem

- a** In practical applications of statistical methods, one often faces the situation that the necessary assumptions are not met entirely. For example, it may be assumed that the errors are normally distributed. It is very likely that the data satisfies this assumption just approximately. But in contrast to everybody's expectation, the classical methods perform just well if this assumption is satisfied *exactly* and may otherwise perform very poorly, even with what seems to be a negligible deviation.

To be more specific, let us introduce the classical regression model where we assume that the response variable Y_i can be described by a linear combination of the predictor (also called explanatory) variables up to some errors; i.e.,

$$Y_i = \sum_{j=0}^m x_i^{(j)} \beta_j + E_i.$$

The term E_i , $i = 1, \dots, n$ denotes the random error which is usually assumed to be independent and normally distributed with expected value 0 and unknown dispersion σ . (Note that the $x^{(j)}$ in the model equation do not necessarily have to be the originally observed or measured variables; they can be transformed variables. The same goes for the response variable.)

In the residual analysis, we check that the conditions that we have described are met as well as we can. Among other things, we try all the tricks to find and then remove possible outliers.

Is this effort necessary? Actually, to some extent it isn't! In this block we will show that we can find outliers more easily with the help of robust methods. And if we are primarily interested in estimating the parameter, we can also leave in a few questionable observations without needing to be afraid that the estimate will be too severely affected.

Example b Air Quality. Pollutants in the air are measured by automatic, fixed measurement stations in 10 minute phases and are also often averaged into hourly or daily values. These measurements can be accessed on the internet. However, these measurement values can only be assessed by considering the location where they were taken, which is impossible for the non-experts. And for locations far away from the measurement stations, even the experts can assess the actual air quality only inadequately. – One might wonder whether there are any statistical approaches to improve on this situations.

Yearly mean values are determined on the basis of a physical (propagation) model and displayed on highly spatially resolved (false color) maps. More attractive would be however false color maps with high *temporal* resolution, which shows the *daily mean* as a function of the location. There are several approaches to this. One is based on the empirical knowledge that the log values for the day and yearly means are linearly

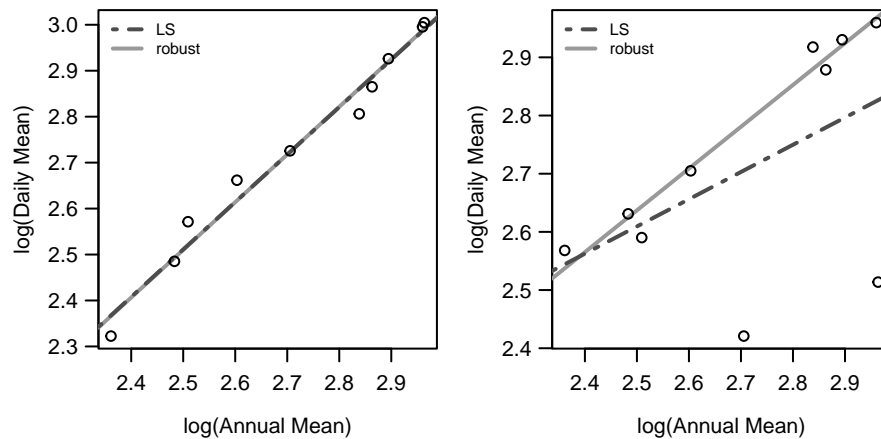


Figure 1.1.b.: Air quality example. Scatter plot of the log values for the hourly day and yearly means (concentrations in ppb) on two different days for 10 measurement stations.

related. Figure 1.1.b shows this for an example with NO_2 concentrations, which are given in ppb (parts per billion $= 1:10^9$). The slope depends on, among other things, the wide-spread weather situation.

As the right scatterplot in Figure 1.1.b shows, it can sometimes occur that a few stations don't quite fit into the empirically determined relationship. Such deviations can be explained well by local weather phenomena (Föhn – a warm wind from south, fog, inversion, ...). The weather phenomena are difficult to predict for individual stations and manually removing these outliers is too laborious for daily use. However, an ordinary least-square fit leads to a “compromise line”, which represents the relationship unsatisfactorily. We thus seek a clever way to handle the values that don't fit.

Joint work with René Locher, www.idp.zhaw.ch/OLK

- c Gross Error Model.** Even if there is only one outlying point in the right scatterplot of Figure 1.1.b, we still obtain a unsatisfactory least-square fit. Most people consider such deviations from our model (assumption of normal distribution) small and there is hardly any reasonable alternative for the normal distribution. Hence, we would like to stick with the normal distribution but assume that in a few cases a deviation from this can occur. This “contaminated normal distribution” is also known by the name “gross error model.” The goal is still to estimate the parameters $\underline{\beta}$ (and σ) from the data, as if the “contamination” didn't exist. This goal should be achieved by so-called **robust** estimation methods.
- d First informal approach to robustness,** which is still commonly used, is to first examine the data for obvious outliers, secondly to remove these and third to use optimal estimation methods for the model on the adjusted data. However, such an approach is not unproblematic, since
- even professional statisticians don't always check the data
 - it is sometimes very difficult, or even impossible, to identify the outliers, especially in high dimensional data
 - the relationship of the data, like, e.g., in a regression setting, can't be studied without first fitting a model

- it is difficult to formalize the process described above so that its properties can be studied and it can be automatized.
- inference based on applying the optimal procedure to the cleaned data is based on a distribution theory which ignores the data cleaning process and hence is theoretically not sound and might even be misleading.

However, this procedure is always better than ignoring the outliers. But keep in mind the results are always too optimistic.

- e Motivation.** In general, the choice of an appropriate estimation method depends heavily on its optimality properties. These, in turn, are based among others on the assumption regarding the distribution of the observations. However, these model distributions only exist in our minds. In reality the data only follow this distribution (model) more or less. Unfortunately, it has now been found that an estimator, which is optimal under the exact distribution, generally does not have to be approximately optimal for distributions that are close to the model distribution.

In the following, we will introduce **robust** estimates that are only approximately optimal under the exact distribution, but they still behave “well” in a neighborhood of the given distribution. From a data-analytic point of view, robust procedures should find the structure best fitting the majority of the data and identify deviating points (i.e., outliers) and substructures for further treatment.

1.2 Measuring Robustness

- a** Whether an estimator is robust can be studied with two simple measurements: **Influence function** and **Breakdown point**. Both measurements are based on the idea of studying the behavior of an estimation function under the influence of gross errors, i.e. arbitrary included data. The **sensitivity** (gross error sensitivity) is based on the influence function and measures the maximal effect of an individual observation on the estimated value. Or in other words, it measures the local stability of the estimating procedure.

On the other hand, the breakdown point measures the global reliability or safety of the estimation procedure by determining the minimal proportion of incongruous observations that suffice to give completely implausible estimates.

- b Location Model.** To illustrate what these two values measure, we restrict ourselves for now to the *simplest* linear regression model, the so-called **location model**:

$$Y_i = \beta_o + E_i, \quad i = 1, \dots, n.$$

(In the following paragraphs, we use β instead of β_o to avoid unnecessary subscripts.)

* A discussion of exactly what we want to achieve with a location model can be found in the appendix A.1.

Example c Sleep Data. The following historical example of a study of the effectiveness of two medications has become famous in statistics. For 10 subjects, the average lengthening of sleep for medication A versus medication B was measured (Cushny and Peebles, 1905). The results (in hours) were

1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4.

This data were for a long time a typical example for normally distributed data. (Sketch the fitted data!)

Now we want to estimate the mean time of how long sleep is prolonged. The most obvious estimate for β is the **arithmetic mean**: $\bar{y} = 1.58$. Another well-known estimator is the **median**: $\text{med} = 1.3$. The **10% trimmed mean** is an estimate, where we first leave out the largest 10% and the smallest 10% of values and then the arithmetic mean is calculated from the remaining values: $\bar{y}_{10\%} = 1.4$. Often, the data are examined for outliers with a **rejection rule**. The values identified as outliers are then left out and from the remaining values the arithmetic mean is again calculated. The most popular rejection rule is the so-called Grubbs' test for outliers: $(\max |y_i - \bar{y}|)/s > 2.18$ (Barnett and Lewis, 1978, Pages 167 and 377), where s is the standard deviation. For the above data with these estimators we get a value of $\bar{y}^* = 1.24$.

* The value 2.18 in Grubbs' test for outliers depends on the sample size N by $\frac{N-1}{\sqrt{N}} \sqrt{\frac{q^2}{N-2+q^2}}$ with q denoting the critical value of the t distribution with $(N-2)$ degrees of freedom and a significance level of $\alpha/(2N)$.

The last three estimators each give a significantly lower average prolongation of sleep. (Also sketch out these four estimate values!)

Definition d Empirical Influence Function. It is obvious that in the above example the differences were caused by the observation $y = 4.6$ hours. We therefore want to investigate the question how the estimated value of an estimator $\hat{\beta}\langle y_1, \dots, y_n \rangle$ changes, if we change the value y in question. An appropriate quantity for this is the so-called **empirical influence function** (or sensitivity curve)

$$SC\langle y; y_1, \dots, y_{n-1}, \hat{\beta} \rangle := \frac{\hat{\beta}\langle y_1, \dots, y_{n-1}, y \rangle - \hat{\beta}\langle y_1, \dots, y_{n-1} \rangle}{1/n}.$$

(The denominator makes this function independent of the sample size.) The empirical influence function is mainly calculated as a function of the added observation y and the estimator being used. But it also depends on the sample itself.

* In structured problems, e.g. in regression, a slightly modified definition is often used:

$$SC\langle y; y_1, \dots, y_n, \hat{\beta} \rangle := \frac{\hat{\beta}\langle y_1, \dots, y_{n-1}, y \rangle - \hat{\beta}\langle y_1, \dots, y_n \rangle}{1/n},$$

i.e. y_n is replaced by a new observation y . The version first presented is especially suitable if y_n is already an outlier. Therefore we use that in the sleep data example.

Example e Sleep Data. The empirical influence function of the four previously described estimators is shown for this example in Figure 1.2.e, if instead of the questionable observation 4.6 hours we had gotten another value y .

Comparing the four estimates shows that the arithmetic mean reacts most strongly to the observation $y = 4.6$; the median changes the least, and the 10%-trimmed mean and the arithmetic mean with the rejection rule lie in between. The "curve" for the arithmetic mean (a straight line!) increases arbitrarily if an increasingly extreme outlier ($y \rightarrow \infty$) is considered, while the other estimates either stop at a (problem-specific) value or fall back to 0 (rejection rule).

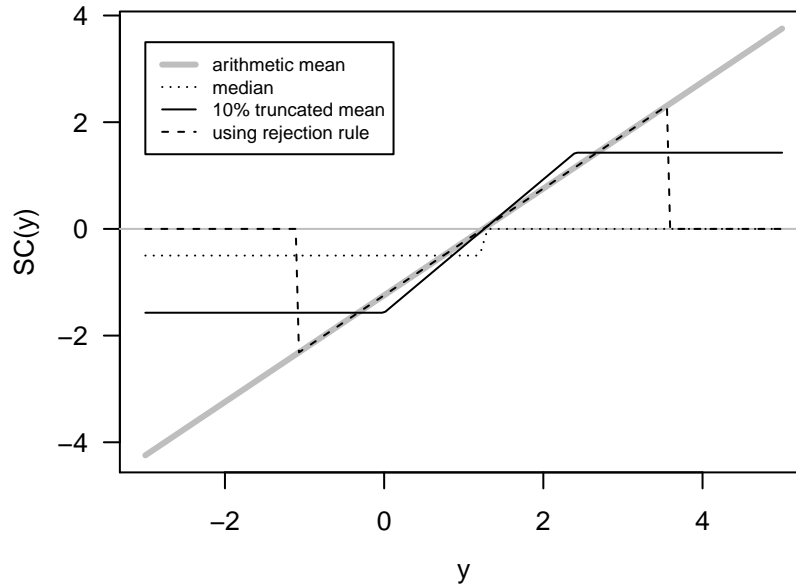


Figure 1.2.e.: Empirical influence function for the sleep data example. The “outlier” $y = 4.6$ hours is varied between $-3 < y < 5$.

Definition f Influence Function. Generally such a curve shows the **influence** of an additional observation y for a given sample. The dependency on a concrete sample y_1, \dots, y_{n-1} is annoying. We would like to have instead a function that is only characterized by the estimation method. Therefore, we consider the value that we would get if we had an infinitely large sample. Then it is possible to replace the sample with its underlying distribution \mathcal{F} (often the normal distribution) and the additional observation y with a point mass. The resulting function is called the **influence function** IF . For our purposes it suffices to note that

$$IF\langle y; \mathcal{F}, \hat{\beta} \rangle \approx SC\langle y; y_1, \dots, y_{n-1}, \hat{\beta} \rangle.$$

For most estimators, the influence function can be calculated.

* In mathematical statistics estimators are considered to be functionals of distribution functions ($\hat{\beta}\langle y_1, \dots, y_n \rangle = \hat{\beta}\langle \mathcal{F}_n \rangle$) with $\hat{\beta}\langle \mathcal{F}_n \rangle \rightarrow \hat{\beta}\langle \mathcal{F} \rangle$, where \mathcal{F}_n is the empirical distribution function of the sample. The formally correct definition of the influence function is then

$$IF\langle y; \hat{\beta}, \mathcal{F} \rangle := \lim_{\varepsilon \rightarrow 0} \frac{\hat{\beta}\langle (1-\varepsilon)\mathcal{F} + \varepsilon\Delta_y \rangle - \hat{\beta}\langle \mathcal{F} \rangle}{\varepsilon},$$

where Δ_y is the point mass at y . Thus, the influence function is a type of directional derivative for a functional in probability mass space. It can be shown that in many situations the empirical influence function SC converges to the influence function IF if $n \rightarrow \infty$.

Definition g Sensitivity. It is intuitively clear that for an estimator that has a bounded influence function IF (or SC), outliers can only have a bounded influence on the estimate value (as long as there are not too many). Thus, an important value for describing robustness of an estimator is the **gross error sensitivity**

$$\gamma^* := \max_y |IF\langle y; \hat{\beta}, \mathcal{F} \rangle|$$

(the mathematically precise expression for the maximum is supremum). Usually we are only interested in the question of whether the sensitivity is bounded or not. This question is easy to answer by the form of the (empirical) influence function (see Figure 1.2.e).

- h Conclusion.** According to the sensitivity criterion γ^*
- the arithmetic mean \bar{Y} is *not* a robust estimator for β . However
 - the median, $\bar{Y}_{10\%}$ as well as \bar{Y}^* are robust estimators for the location parameter β in the location model.

Example i Sleep Data. Until now we have only studied the influence of **one** (unusual or extremal) observation on the estimator. How does the estimator react to two outliers? If we consider the case $y_{n-1} = y_n \rightarrow \infty$, it then holds

$$\begin{aligned} \bar{y} &\rightarrow \infty; & \text{med} &= 1.3 \text{ (remains constant)} \\ \bar{y}_{10\%} &\rightarrow \infty; & \bar{y}^* &\rightarrow \infty \end{aligned}$$

It is notable that the rejection rule does not identify any of the two observations as outliers! That is, this estimation procedure doesn't deliver what it promises!

Although the last three estimators have displayed very similar behavior with respect to the influence function, in this new situation they are clearly distinct from one another. Except for the median, they behave like the arithmetic mean and are thus less robust against outliers than the median.

Definition j Breakdown Point. A simple measurement that captures this aspect of robustness is the **breakdown point** $\varepsilon_n^*(\hat{\beta}; \underline{y})$. It is the maximum ratio of outlying observations such that the estimator still returns reliable estimates. More formally, call \mathcal{X}_m the set of all data sets $\underline{y}^* = \{y_1^*, \dots, y_n^*\}$ of size n having $(n - m)$ elements in common with $\underline{y} = \{y_1, y_2, \dots, y_n\}$. Then

$$\varepsilon_n^*(\hat{\beta}; \underline{y}) = \frac{m^*}{n},$$

where

$$m^* = \max \{m \geq 0 : |\hat{\beta}(\underline{y}^*) - \hat{\beta}(\underline{y})| < \infty \text{ for all } \underline{y}^* \in \mathcal{X}_m\}.$$

If enough observations are arbitrarily changed, no common estimator can return reliable estimates, and we talk about the *breakdown* of the estimator. For common estimation procedures the breakdown occurs at latest for $m > n/2$ and, hence, the maximal breakdown point $\varepsilon_n^*(\hat{\beta}; \underline{y})$ is smaller than $1/2$.

* Why is the breakdown point for common estimation procedures at most $1/2$? Suppose, for example, that half of the data is shifted by the same systematic error. If this shift is larger than the random error, we can identify two groups in the data. However, without additional information it would be impossible to say which are the unmodified observations. Consequently, no common (i.e. translation equivariant) estimator gives a plausible value and the maximal breakdown point that such an estimator can have is thus $1/2$.

- k Conclusion** In the location model, out of the considered estimators
- only the median achieves the maximal breakdown point and is thus maximally resistant to outliers.
 - The other estimators that we considered are less resistant to outliers.

I Overall, we can conclude that

- the gross error sensitivity $\gamma^*(\hat{\beta}, \mathcal{F})$ measures the maximal effect of a small disturbance.
- and the breakdown point $\varepsilon_n^*(\hat{\beta}; y_1, \dots, y_n)$ measures the minimal size of the disturbance that leads to catastrophe.

1.3 M-Estimation

a In the previous section we have seen that the median is the best robust estimator for the parameter β out of the four presented estimators. The median, however, is neither very efficient (so not close to optimal for normally distributed error), nor can it be well generalized to the multiple regression model. Therefore, we introduce here another family of estimators for the location model: The M-estimators. They play a central role not only in fitting of location models but also in fitting of regression models.

b From a practical point of view, the M-estimator is essentially a weighted mean, wherein the weights are designed to prevent the influence of outliers on the estimator as much as possible:

$$\hat{\beta}_M = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}.$$

It seems pretty obvious that the weights must depend on the true parameter β , or if unknown, on the estimated parameter, since it is central to define an outlier.

c Influence function of M-estimators. In order to determine appropriate weights (or weight functions), it helps to consult the corresponding influence function IF . Theory tells that the influence function of $\hat{\beta}_M$ is equal to

$$IF\langle y, \hat{\beta}, \mathcal{N} \rangle = \text{const} \cdot w \cdot \tilde{r} = \text{const} \cdot \psi\langle \tilde{r} \rangle,$$

where \tilde{r} is the scaled residual $(y - \beta)/\sigma$ and $\psi\langle \tilde{r} \rangle := w \cdot \tilde{r}$. Hence, the influence function of an M-estimator is proportional to the ψ -function. The least squares estimator is a M-estimator, too, but a non-robust one. As shown in Figure 1.3.c(a) and (c), if $\psi\langle u \rangle = u$, the weights are all 1 and hence the above weighted mean reduces to an ordinary mean which corresponds to the classic estimator in case of Gaussian error.

To obtain a robust M-estimator we must choose a *bounded* ψ -function like, e.g., the ψ -function of Huber's M-estimator shown in Figure 1.3.c(b). The corresponding weight function can be determined by $w_i = \psi(\tilde{r}_i)/\tilde{r}_i$ (cf. Figure 1.3.c(d)).

* It can be shown that if a *monotone, bounded* ψ -function is used, then

- the corresponding M-Estimator is defined uniquely and
- the **breakdown point of the corresponding M-Estimator** is approximately 1/2.

Definition d Usually, the **M-Estimator** is defined by an implicit equation,

$$\sum_{i=1}^n \psi \left\langle \frac{r_i\langle \hat{\beta} \rangle}{\sigma} \right\rangle = 0 \quad \text{with } r_i\langle \beta \rangle = y_i - \beta,$$

where σ is the scale parameter. Basically, this equation corresponds to the normal equation of least-squares estimators. As we have seen and will see later, the ψ -function plays an important role in describing the properties of M-estimators.

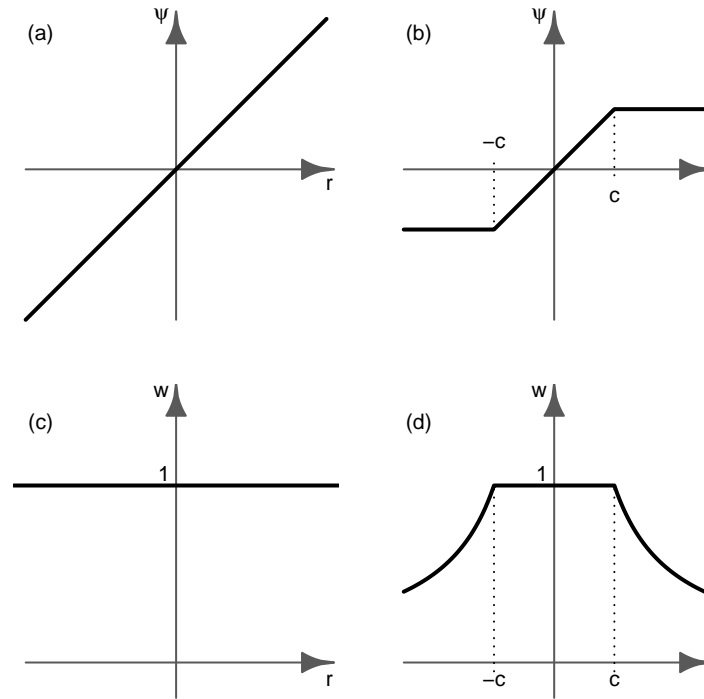


Figure 1.3.c.: ψ - and weight function for the ordinary least squares estimation (on the left), which is not robust, and for a robust M-estimator (on the right). The latter ψ -function is also known as Huber's ψ -function.

- e Computation of M-estimators.** Since the weights in 1.3.b depend on the unknown parameter β (and σ), we cannot calculate the weighted mean explicitly. But this weighted-means representation of M-estimators leads to a simple iterative algorithm for calculating the M-estimator.
1. We start with the median as an initial estimate of β and then estimate σ (see below).
 2. Calculate the weights w_i by $w_i = \psi(\tilde{r}_i)/\tilde{r}_i$
 3. Calculate a new estimate of β by 1.3.b
 4. Repeat Step 2 and 3 until the algorithm converges
- f Estimation of the scale parameter.** To make M-estimation applicable, except in the case of the arithmetic mean or the median (L_1 estimation), the points where the data lose influence have to be determined. For the Huber ψ function this is the point where there is a “kink” in the function. This can only be done in a reasonable way if we take into account the scaling of the residuals. Thus in the definition of the M-estimator (1.3.d) we have already divided by the scale parameter σ , as was also the case in Grubbs' the rejection rule in 1.2.c. The standard deviation as an estimate for σ is very unrobust, however. Thus the rejection rule breaks down for two outliers out of ten observations.

An example of a robust estimator for the scale parameter is

$$s_{\text{MAD}} = \text{med}_i \langle |y_i - \text{med}_k \langle y_k \rangle| \rangle / 0.6745$$

(**median absolute deviation**). Via the "correction" $1/0.6745$ we achieve a consistent estimator for the standard deviation σ of a normal distribution. Now we use the estimator s_{MAD} in the M-estimator instead of σ , unless we know σ from some other source.

* There are a few other ideas of how to get a robust estimator for the scale parameter σ . Often, the scale parameter is re-estimated for each iteration step. This is referred to as "Huber's Proposal 2".

g * We can now specify a better rejection rule, the so-called Huber type skipped mean:

$$\left| \frac{y_i - \text{med}_j \langle y_j \rangle}{s_{\text{MAD}}} \right| > 3.5$$

(cf. Hampel, 1985).

h Now that we have a suitable scale estimator, we still have to explicitly determine the "kink" points of the ψ function via a so-called tuning constant. Usually this is determined such that the corresponding estimator has a relative efficiency of 95% for the model distribution, e.g. the normal distribution. Thus for the Huber ψ function we get a tuning constant of $c = 1.345$.

Example i Sleep Data. If we apply the M-Estimator with Huber's ψ function and $c = 1.345$ to the sleep data example, we get the estimate value of 1.37 for the "average" prolongation of sleep β . This value is, as expected, smaller than the Maximum-Likelihood Estimator $\bar{Y} = 1.58$.

In R you might compute Huber's M-Estimator by

```
> Sleep <- c(1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4)
> library(robustbase)
> (Sleep.HM <- huberM(Sleep, k=1.345))
## (shortened)
## $mu = 1.371091
## $s = 0.59304
```

j Conclusion. Robust methods make it possible to confidently detect outliers. On the other hand, we can estimate the parameter β reliably without an exact identification of the outliers if we are only interested in the estimation of the parameter, since the outliers won't do any damage to the estimation.

Often it is worthwhile to investigate the outliers. If a transcription error can be handled easily, it should be corrected. If a foreign population is identified (e.g. sick people), not only the outliers but all observations corresponding to this population should be eliminated. And don't forget: Outliers can reveal valuable information that might lead to surprising insights.

1.4 Inference with M-Estimators

a Asymptotic Distribution of the M-Estimator. A point estimate is incomplete without specification of the corresponding confidence interval. To determine this, we have to know the distribution of the estimator. For M-estimators, (only) asymptotic results are known: They are *consistent* ("asymptotically unbiased") and *asymptotically normally distributed with variance* $\tau\sigma^2$. The correction factor

$$\tau = \frac{\int \psi^2 \langle u \rangle \phi \langle u \rangle du}{\left(\int \psi' \langle u \rangle \phi \langle u \rangle du \right)^2},$$

where $\phi\langle u \rangle$ is the density of the normal distribution \mathcal{N} , takes into consideration the down-weighting of “good” observations and is larger than 1. This expression for τ can be explicitly calculated in advance. Another possibility is to estimate it from the observations by

$$\hat{\tau} = \frac{\frac{1}{n-1} \sum_{i=1}^n \psi^2\langle \tilde{r}_i\langle \hat{\beta} \rangle \rangle}{\left(\frac{1}{n} \sum_{i=1}^n \psi'\langle \tilde{r}_i\langle \hat{\beta} \rangle \rangle \right)^2}$$

(see, e.g., Huber and Ronchetti, 2009). Here, $\tilde{r}_i\langle \hat{\beta} \rangle$ are the appropriate scaled residuals $\tilde{r}_i\langle \hat{\beta} \rangle = (x_i - \hat{\beta})/\hat{\sigma}$ and $\hat{\sigma}$ is a robust scale estimator like, for example, $\hat{\sigma} = s_{\text{MAD}}$.

Some more details about the asymptotic distribution of M-estimators is given in Section 5.4 of the Block on “Resampling, Nonparametric Tests and Asymptotics” by Werner Stahel.

* Note that when considering the variance asymptotically, it is inflated by n , analogous to the Central Limit Theorem, i.e. for practical use it must be divided by n .

- b Asymptotic Confidence Interval.** The asymptotic 95% confidence interval for the location parameter β is now given by

$$\hat{\beta} \pm 1.96 \text{ se} \langle \hat{\beta} \rangle \quad \text{mit } \text{se} \langle \hat{\beta} \rangle = \sqrt{\hat{\tau}} \frac{\hat{\sigma}}{\sqrt{n}}$$

This confidence interval is, up the correction factor $\sqrt{\hat{\tau}}$, analogous to the one that is derived from the z-test. For finite sample sizes, this asymptotic approximation can be insufficient. To better align the actual coverage probability with its nominal value (=95%), substitute the value 1.96 with the quantile value of the corresponding t distribution (i.e., we take into account that the standard deviation σ must be estimated). Even though this and further corrections were proposed a long time ago, their finite sample properties have still been little studied.

Example c Sleep Data. To compute the 95% confidence interval we continue the calculation of 1.3.i. Next, we estimate σ and τ : $s_{\text{MAD}} = 0.593$ and $\sqrt{\hat{\tau}} = \sqrt{1.5068} = 1.2275$. If we now use the t_9 quantile value, we get the following approximate 95% confidence interval from the M-Estimator: $[0.85, 1.89]$. For comparison we also give the classical 95% confidence interval $\bar{x} \pm s q_{0.975}^{t_9}$ an: $[0.70, 2.46]$; it is obviously longer than that of the M-Estimator. Some R-computation is given in R Output 1.4.c. In practise, we would use the R function `lmrob()` which we introduce later.

- d Robust Tests.** Note that with the help of the approximate confidence interval of the M-Estimator, statistical tests can be carried out. It can be shown that these are robust in the following sense: We want to avoid the erroneous conclusion that the specified alternative be considered to be true, if neither the null hypothesis nor the alternative hold, but instead some other alternative (e.g. the data are not normally distributed). Stahel (2007, Kapitel 8.6) calls such erroneous conclusions “Type 3 Error.”

Recap e Before we move to the regression model, we once again emphasize:

- \bar{X} is an optimal estimator for the location parameter β under exactly normally distributed data. Real data are, however, *never exactly* normally distributed. But even a single outlier makes \bar{X} into an ineffective estimator for the location parameter.

```

> (Sleep.HM <- huberM(Sleep, se=TRUE, k=1.345)) # See also 1.3.i
## $mu    # = 1.371091
## $s     # = 0.59304
## $it    # = 11
## $SE    # = 0.2302046
## Confidence interval
> Sleep.HM$mu + c(-1,1)*qt(0.975, length(Sleep)-1) * Sleep.HM$SE
## 0.8503317 1.8918499
## Calculating  $\tau$  separately
> robustbase::tauHuber(Sleep, mu=Sleep.HM$mu, s=Sleep.HM$s, k=1.345)
## [1] 1.506816

## Classical estimation
> t.test(Sleep)
## (shortened)
## 95 percent confidence interval: 0.7001142 2.4598858
## sample estimates: mean of x = 1.58

```

R-Output 1.4.c: Some R-Output for calculating the 95% confidence interval for the example Sleep Data. When `se=TRUE` in `huberM(...)`, the standard error is computed using the τ correction factor.

- Better (= more efficient, intuitively “more correct”) procedures exist. With M-estimators we can control the influence of gross errors very precisely.
- If a rejection rule, then Huber Skipped Mean.

We have achieved the described results by considering the influence function and breakdown point, which are two important measurements for characterizing the fundamental robustness of estimators.

2 Linear Regression

2.1 Regression M-Estimation

- a** We return to the regression model from 1.1.a. In the air quality example we have seen that two outliers suffice to give unusable least squares lines (see Figure 1.1.b).

Definition b Regression M-Estimator. Analogously to the location model (cf. 1.3.d), we will replace the ordinary least squares estimator by a weighted least squares estimator, where the weights should reduce the influence of large residuals; i.e., the influence of outliers. More explicitly, let's start at the normal equations of the weighted least squares estimator, usually written as $\mathbf{X}^T \mathbf{W} \underline{r} = \underline{0}$, where \mathbf{W} is a diagonal matrix with the weights and \underline{r} is the vector of residuals $r_i(\underline{\beta}) := y_i - \sum_{j=0}^m x_i^{(j)} \beta_j$. For our purpose, it is more useful to write this system of equations as

$$\sum_{i=1}^n w_i \cdot r_i \cdot \underline{x}_i = \underline{0}.$$

Replacing again $(w_i \cdot r_i)$ by $\psi(r_i/\sigma)$ yields the formal definition of a **regression M-estimator**

$$\sum_{i=1}^n \psi \left\langle \frac{r_i(\underline{\beta})}{\sigma} \right\rangle \underline{x}_i = \underline{0}.$$

In order to determine appropriate weights (or weight functions), it again helps to consult the corresponding influence function IF of the regression M-estimator.

- c Influence Function of a Regression M-Estimator.** As in the location model, the ψ function of the regression M-estimator again reflects the influence of residuals on the estimate; for the influence function we have

$$IF \left\langle \underline{x}, y; \hat{\underline{\beta}}, \mathcal{N} \right\rangle = \psi \left\langle \frac{r}{\sigma} \right\rangle \mathbf{M} \underline{x},$$

where the matrix \mathbf{M} is yet to be determined. In contrast to the location problem, a term $\mathbf{M} \underline{x}$ additionally comes into play, which takes into account the position of the observation in space. We'll discuss the consequences of this term in section 2.3.

To limit the influence of large residuals, the ψ -function must be bounded in the same way as with the robust M-estimator of location (cf. 1.3.c). The corresponding robustness weights are again given by $w_i = \psi(\tilde{r}_i)/\tilde{r}_i$.

* If the explanatory variables $\underline{x}_i = (x_i^{(0)}, \dots, x_i^{(m)})^T$ can be described by a distribution \mathcal{K} , then the matrix \mathbf{M} is given by

$$\mathbf{M}^{-1} := \int \psi'(r) \phi(r) dr \int \underline{x} \underline{x}^T d\mathcal{K}(\underline{x}).$$

There is also a version for known, fixed explanatory variables \underline{x}_i (see Hampel, Ronchetti, Rousseeuw and Stahel, 1986, Kap. 6.2).

- d Computation.** According to the explanations given in 2.1.b the regression M-estimator can be interpreted as a weighted least squares estimator, where the weights however depends on the unknown parameters $\underline{\beta}$ and σ . But still, this representation shows a way for applying an iterative algorithm for solving the system of equations that defines the regression M-estimator. In appendix A.2, it is described in more detail. There you also find a second algorithm, whose schematic representation clearly shows the treatment of outliers (see Figure A.2.c).
- e Asymptotic Distribution.** The regression M-estimator is consistent and asymptotically normally distributed with covariance matrix $\sigma^2 \tau \mathbf{C}^{-1}$, where $\mathbf{C} := \sum_i \underline{x}_i \underline{x}_i^T$. Up to the correction factor τ , this covariance matrix corresponds to the covariance matrix of the least squares estimation. The correction factor τ is defined analogously as in 1.4.a.

Thus the covariance matrix of the estimator $\underline{\hat{\beta}}$ can be estimated by $\widehat{\mathbf{V}} = (\hat{\sigma}^2) \hat{\tau} \hat{\mathbf{C}}^{-1}$ where

$$\hat{\mathbf{C}} = \frac{\sum_i \tilde{w}_i \underline{x}_i \underline{x}_i^T}{\frac{1}{n} \sum_i \tilde{w}_i}$$

with \tilde{w}_i either $\tilde{w}_i := 1$ or $\tilde{w}_i := \psi(\tilde{r}_i)/\tilde{r}_i$. The \tilde{r}_i are again the appropriate scaled residuals $\tilde{r}_i = r_i(\underline{\hat{\beta}})/\hat{\sigma}$. Weights of the form $\tilde{w}_i = \psi(\tilde{r}_i)/\tilde{r}_i$ are sensible in the calculation of $\hat{\mathbf{C}}$, since contaminated observations should be down-weighted too in the estimation of the covariance matrix according to their potential degree of contamination, measured by \tilde{w}_i .

* Again, corrections that take into account the finiteness of the sample can be appropriate (also see 1.4.b). They are based on recommendations from Huber (see e.g. Huber and Ronchetti, 2009, Section 7.6).

- f Estimation of the Scale Parameter.** Usually, in regression M-estimation, the scale parameter σ must be estimated simultaneously with the regression parameters. An estimator analogous to s_{MAD} can be used, which takes into account that the expected value of the error is 0. This is the standardized **median of absolute values**

$$s_{\text{MAV}} = \text{med}_i(|r_i|)/0.6745.$$

Example g Air Quality (Modified). To demonstrate the effects of the robust regression M-estimator, we use a data set similar to that in Figure 1.1.b, on the right. The modified data set contains just one outlier. In Figure 2.1.g, the data are shown superimposed by the solutions of the least squares fit and the regression M-fit. As we again see, a single outlier is enough to cause the least squares method to give an unsatisfactory result. It passes through areas where no data is present. However, the Huber type regression M-estimator describes the majority of the data very well. In the residual analysis the outlier is also very visible. This is especially important in cases where the number of parameters $p \gg 2$, because then a visualization like in Figure Figure 2.1.g is no longer possible.

* In R you can fit the regression model robustly using an M-estimator for example by

```
> require(MASS)
> AQ.Mfit <- rlm(y ~ x, method="M", data=xy)
```

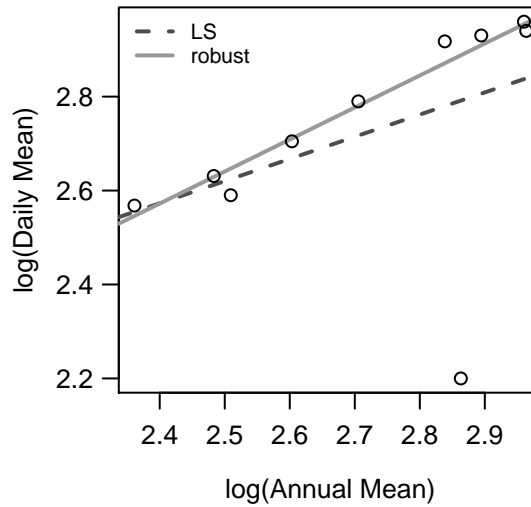


Figure 2.1.g.: Modified Air Quality Example with two fitted lines: least squares method (dashed line) and regression M-Estimator with Huber's ψ function (solid line).

2.2 Example from Molecular Spectroscopy

- a** With this high-dimensional example I would like to show how robust methods can reliably estimate parameters and successfully identify outliers. In contrast to the preceding “two-dimensional” example, the problems here are not so obvious and it is very difficult to recognize the outliers.
- b Term Value Model.** We can gain insight into the physical structure of atoms and molecules by studying their energy levels, which are linked to the quantum mechanical state of the atom or molecule. Precise direct measurements of these energy levels are not possible. However, energy differences (called transition energy) between pairs of states can be measured very precisely:

$$Y_i = \beta_{u_i} - \beta_{\ell_i} + E_i.$$

Here Y_i is the measured transition energy from the higher state u_i to the lower state ℓ_i , β_k is the energy of the molecule in state k and E_i is the measurement error. Since the variable β_k can only be determined up to an additive constant, the energy of an arbitrary base state (usually for the lowest value) is set to 0. The variable β_k therefore has its own name, term value for the state k .

- c Assignments from Dieke and Tomkins.** For part of the tritium (T_2) spectrum, Dieke and Tomkins (1951) have assigned the spectral lines (transition energies) Y_i to the corresponding pair of quantum mechanical states (u_i, ℓ_i) . For complex spectra, like that of the tritium molecule, the assignment process is very difficult. False assignments are unavoidable. Robust estimation techniques have proven to be very helpful in achieving reliable estimates for the term values and in quickly recognizing wrong assignments.
- d The Term Value Model is a Regression Model.** All transition energies are collected into the vector $\underline{Y} = (Y_1, \dots, Y_n)^T$ and all term values form the vector $\underline{\beta}$, so the model in 2.2.b can be written as

$$\underline{Y} = \mathbf{X} \underline{\beta} + \underline{E}$$

where the design matrix \mathbf{X} is given by

$$X_{ij} = \begin{cases} 1 & \text{if } j = u_i \\ -1 & \text{if } j = \ell_i \\ 0 & \text{otherwise} \end{cases}$$

In essence, this design matrix has the form that occurs for unbalanced two-way analysis of variance problems. Consequently, the parameters can be estimated via regression methods.

In this example we have 608 observations, i.e. 608 spectral lines Y_i , to which a pair of quantum mechanical states (u_i, ℓ_i) have been assigned. From this there are now 64 unknown upper states and 100 unknown lower states, so a total of 164 parameters to estimate. To estimate these parameters, the regression M-Estimator with Huber's ψ function is used. The tuning constant c is set to the value 1.345. Since the variance of the measurement error can be set to be identical to the variance of the measurement procedure, for the scale parameter $\hat{\sigma}$ a fixed value of 0.02 cm^{-1} is inserted.

Residual Analysis

- e Least Squares Solution.** The histogram (Figure 2.2.e(A)) shows a long-tailed distribution for the least squares solution, with a standard deviation of 1.9 cm^{-1} . Since the standard deviation is not a good measurement for the variability of a long-tailed distribution, we use the standardized MAV (see 2.1.f). With this we get a value of 0.158 cm^{-1} . This value is still about eight times larger than it should be. So for the measurement error a value of about 0.02 cm^{-1} is assumed.
- f Regression M-Estimation.** The residuals of the regression M-Estimators (Fig. 2.2.e(B)), however, show an extremely clear structure: 94% of the residuals (570 of 608) are smaller than 0.1 cm^{-1} . The standardized MAV of the residuals is 0.024 cm^{-1} . This agrees well with our prior knowledge of the precision of the measurement. The remaining 6% of the residuals outside of $\pm 0.1 \text{ cm}^{-1}$ are distributed over the interval from -54 cm^{-1} to 9 cm^{-1} . This clear structure is a good basis for a preliminary separation between "good" and clearly false assignments.

Analysis of the Estimated Term Values

- g Least Squares Solution.** In a further step, the estimated term values are now compared with the theoretically determined values. These theoretically determined term values are especially appropriate for comparison, since they are based on theoretical prior knowledge and not on measurements. Therefore they cannot be influenced by false assignments. The disadvantage of the theoretically determined term values is due to the fact that the calculations are only approximations and thus can exhibit systematic deviations in comparison to the true term values. Generally we want to directly evaluate the quality of the approximation from the analysis of the experiment.

Figure 2.2.g shows the differences between the estimated and the calculated term values against the calculated term values. The term values, which were obtained via least squares estimation, generally demonstrate expected systematic deviations. However they are overlaid by many scattered individual points (Fig. Figure 2.2.g(A)).

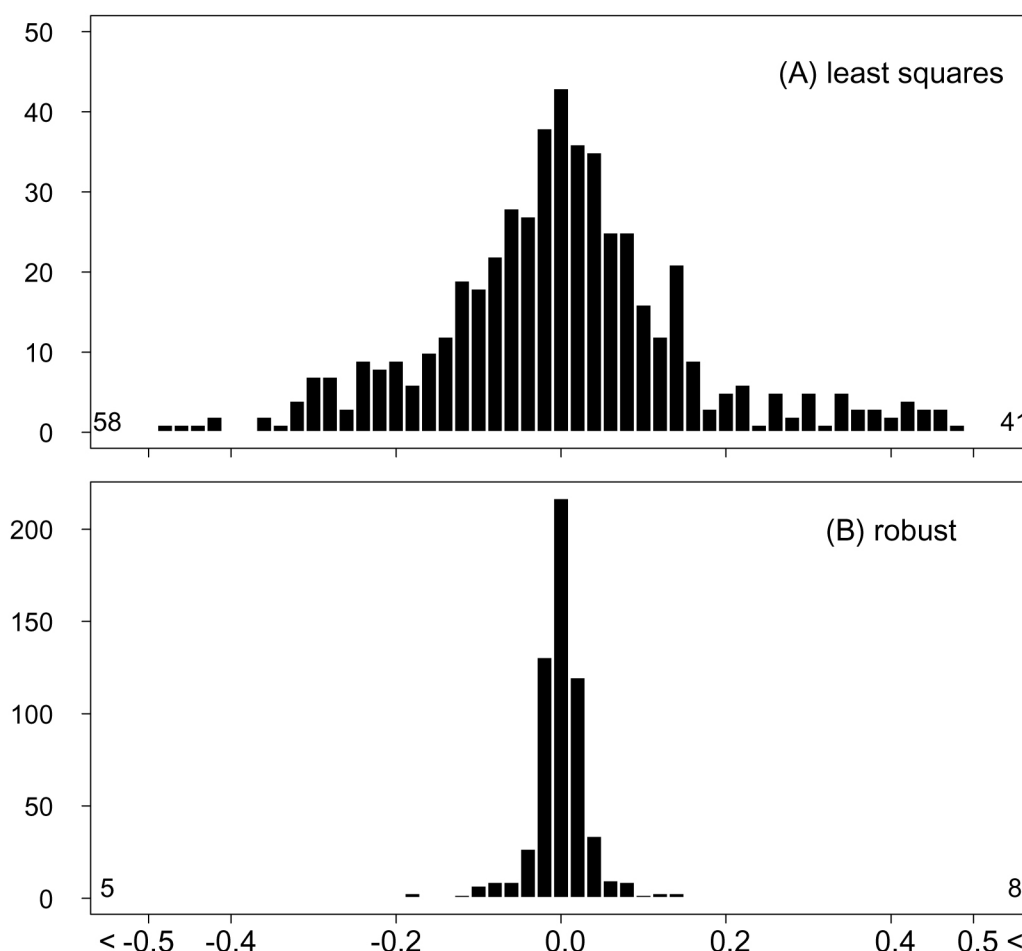


Figure 2.2.e.: Histogram of the least squares residuals (A) and the residuals of the regression M-Estimator (B). The x values are limited to $\pm 0.5 \text{ cm}^{-1}$ in order to get a better look into the middle of the empirical distribution. The numbers on the edges of the histograms give the number of residuals outside of the shown interval.

- h **Regression M-Estimation** The term values that were obtained via the regression M-Estimator now show very smooth systematic deviations (Fig. 2.2.g(B)). This graph corresponds much better to the expectations of the researcher. However, some points deviate from the smooth lines. They consist of those term value estimations which have broken down. This effect is called **local** or **partial breakdown** (see Ruckstuhl, 1997). The cause of this is that the corresponding term values occur in too many incorrect assignments. Despite this limitation, the regression M-Estimator gives very plausible results.
- i By comparing the estimated with the theoretically determined term values, still more wrong assignments can be discovered. Together with the insight from the residual analysis, dubious (incorrect) assignments are largely corrected. Further details can be found in Ruckstuhl, Stahel and Dressler (1993).

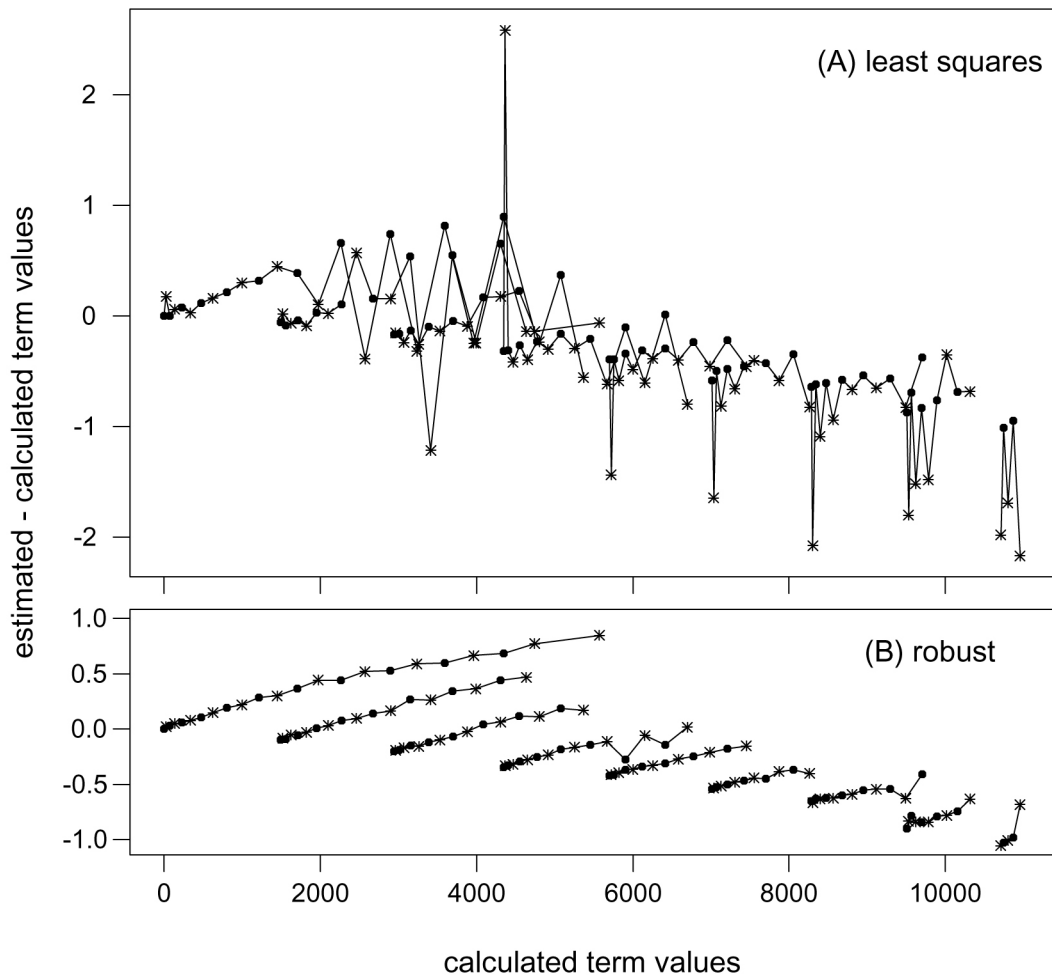


Figure 2.2.g.: Difference between the estimated and theoretically determined term values against the theoretically determined term values for the lower states. In Fig. (A) this is shown for the least squares solution and in Fig. (B) for the regression M-Estimator. Points whose states u_i have the same vibratory quantum number are connected with lines.

2.3 General Regression M-Estimation

Example a Air Quality. We now turn back to the Air Quality Example and again slightly modify the data set so that the outlier is on a leverage point. In this case the regression M-Estimation with Huber's ψ function shows the same unsatisfactory picture as the least squares estimation (Figure 2.3.a, left). This shows clearly that regression M-estimation bounds only the influence of the residuals, but not the influence function as a whole. This situation is also visible from the formula for the influence function (2.1.c) of the regression M-estimator. We therefore call this robust regression M-estimator a BIR estimator (for **b**ounding the **i**nfluence of **r**esiduals).

b General Regression M-Estimators or short GM-Estimator. Estimators bounding the total influence function can be constructed for example by modifying the regression M-estimator by incorporating the outlyingness of the observation in the design space. Two approaches are of primary interest in the literature:

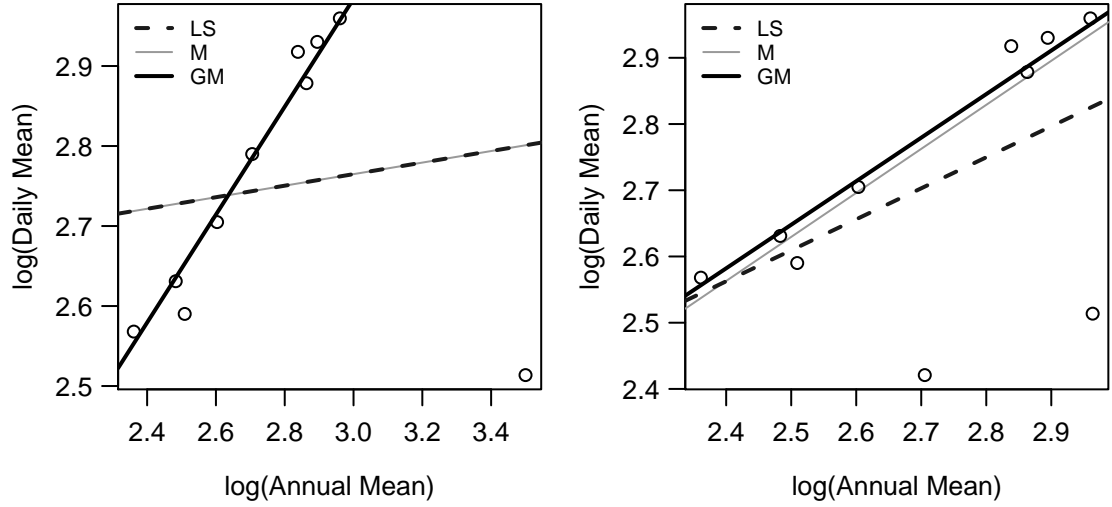


Figure 2.3.a.: Air Quality Example. Fitted lines are shown using the least squares method (dashed line), the regression M-estimator with Huber's ψ function (thin, solid line) and Schweppe type M-estimator with Huber's ψ function (thick, solid line). This is shown for a modified Air Quality example (left) and for the real example (right).

either (**Mallows**)

$$\sum_{i=1}^n w \langle d \langle \underline{x}_i \rangle \rangle \cdot \psi_c \left\langle \frac{r_i \langle \hat{\beta} \rangle}{\sigma} \right\rangle x_i^{(k)} = 0, \quad k = 1, \dots, p,$$

and (**Hampel-Krasker-Welsch, also called Schweppe**)

$$\sum_{i=1}^n w \langle d \langle \underline{x}_i \rangle \rangle \cdot \psi_c \left\langle \frac{r_i \langle \hat{\beta} \rangle}{w \langle d \langle \underline{x}_i \rangle \rangle \cdot \sigma} \right\rangle x_i^{(k)} = \sum_{i=1}^n \psi_{c, d \langle \underline{x}_i \rangle} \left\langle \frac{r_i \langle \hat{\theta} \rangle}{\sigma} \right\rangle x_i^{(k)} = 0, \quad k = 1, \dots, p,$$

where $w \langle \cdot \rangle$ is a suitable weight function and $d \langle \underline{x}_i \rangle$ is some measure of the “outlyingness” of \underline{x}_i in the design space.

Thinking of the classical definition of leverage, a straightforward approach to define the weights in Mallows' approach directly is to set $w \langle d \langle \underline{x}_i \rangle \rangle = w_{x_i} = 1 - H_{ii}$ or $w_{x_i} = \sqrt{1 - H_{ii}}$, where H_{ii} is the i th-diagonal element of the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. As we see in 2.3.e, this definition of w_{x_i} has some pitfalls. An alternative approach measures outlyingness of \underline{x}_i by

$$d \langle x_i \rangle = \frac{(x_i - \text{median} \langle x_k \rangle)}{\text{MAD} \langle x_k \rangle}$$

in one dimension or by the Mahalanobis distance in several dimensions, where the location and the covariance matrix are estimated robustly (cf. Sec. 4.1). In the presence of factor variables, this approach may fail, however. Suitable weight functions $w \langle u \rangle$ are, e.g.,

$$w \langle u \rangle = \min \left\{ 1, \frac{k}{|u|} \right\} \quad \text{Huber weights}$$

or

$$w\langle u \rangle = \left(1 - \frac{u^2}{k^2}\right)^3 I\langle |u| \leq k \rangle$$

where k is a suitable tuning constant.

Example c Air Quality. Applying the Schweppe type M-estimator to the Air Quality data shown on the right of Figure 1.1.b, we obtain the solution shown on the right of Figure 2.3.a. The weights $w\langle x_i \rangle$ have been adjusted especially to this problem to obtain satisfactory solutions with common weather phenomena.

* In R you can fit the regression model robustly using a Schweppe type M-estimator for example by

```
> require(MASS)
> x.h <- 1-hat(model.matrix(y ~ x, xy))
> AQ.GMfit <- rlm(y ~ x, data=xy, weights=x.h, wt.method="case")
```

The example from molecular spectroscopy, which was discussed in the previous chapter, does not have any leverage points because it is simply an incomplete, unbalance two-way analysis of variance.

d Breakdown Point of the Regression GM-Estimator. Unfortunately all the regression M-estimators introduced so far have a big disadvantage. Their breakdown point ε^* is maximally $1/p$, where p is the number of unknown coefficients. This means that for $p = 7$ explanatory variables, $1/7 \approx 14\%$ gross error observations suffice to cause the estimation to collapse. At first glance, this is a high proportion of possible outliers. However, if we consider that an observation $[x_i, y_i]$ is contaminated as soon as one of its components is grossly wrong, for seven explanatory variables we only need a portion of 2% grossly incorrect values to contaminate $1/7$ of the observations. This is frustratingly few, especially in light of the fact that 2% gross error is not uncommon in practice.

e Failure of Classical Residual Analysis. The approach based on GM-estimators and $w_{x_i} = 1 - H_{ii}$ mainly breaks down because it doesn't seem possible to identify an accumulation of leverage points with the projection matrix \mathbf{H} . It also follows that all diagnostic tools (residual analysis) that are based on the least squares method may not recognize the outliers in situations where the GM-estimator fails.

The reason that the diagonal elements of the projection matrix don't suffice is made clear with the following modification of the modified Air Quality Example shown on the left in Figure 2.3.a. If we place a second or even a third point at the outlying leverage point, these points are no longer exceptional with respect to leverage and, hence, to Cook's Distance in the "Residual vs Leverage" plot (see Figure 2.3.e). Naturally these three points are immediately recognized in the scatter plot. With many explanatory variables, though, the scatter plot is no longer helpful and we must be able to rely on graphics like the "Residuals vs Leverage" plot.

2.4 Robust Regression MM-Estimation

a Regressions M-Estimator with Redescending ψ Functions. Computational experiments show that the M-estimation is actually robust when a ψ -function with rejection

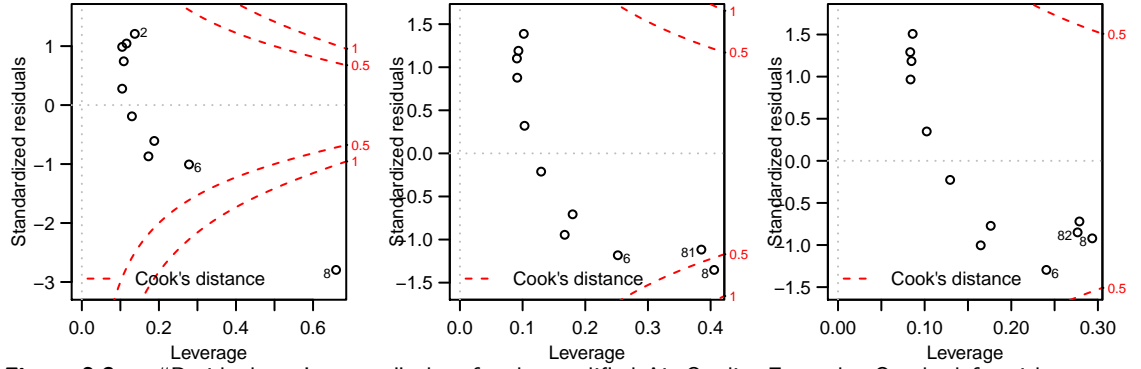


Figure 2.3.e: “Residuals vs Leverage” plots for the modified Air Quality Example. On the left, with one outlying leverage point, middle with two observations at the outlying leverage point, and on the right with three observations. Observation 8 is very influential on the left, but on the right all three observations seems harmless.

of distant outliers is chosen, since this estimator eliminates the influence of distant outliers completely. More reliable solutions are achieved when the influence of distant outliers is ignored gradually as, e.g., by the so-called redescending ψ functions shown in Figure 2.4.c. In fact, there is a solution of equation 2.1.b usually which identifies the outlier “correctly” - but there are also other solutions; the equation does not define the (regression) M-estimation uniquely. That is, we will end up with different solutions depending on the starting value for the estimating algorithm. If a suitable - robust - starting value is selected, we obtain the “correct” solution! In case of the location model, the median is such a “magic” starting value. And for the regression model? - We still need an estimate with a high breaking point to start with.

Definition b S-Estimator. In general, a regression function fits the data well, if the residuals are small. This can be formalized so that a robust scale estimate of the residuals should be as small as possible in analogy to the least squares estimator¹.

An S-estimator of $\underline{\beta}$ solves the implicit equation

$$\frac{1}{n-p} \sum_{i=1}^n \rho \left\langle \frac{y_i - \underline{x}_i^T \underline{\beta}}{s} \right\rangle = \kappa_\rho,$$

with the scale parameter s as small as possible. The parameter κ_ρ is known, but depends on the choice of the function $\rho(\cdot)$. It ensures that asymptotically the solution s is identical to the standard deviation σ for normally distributed observations. To obtain an estimator with high breakdown point of $1/2$ the function $\rho(\cdot)$ must be symmetric and bounded; i.e., the derivative $\psi = \rho'$ is then redescending.

* S in “S-estimator” refers to the fact that this estimator essentially is based on the minimization of a (robust) Scale M-Estimator given by the implicit equation above.

In the 1980s, various multiple regression estimators with high-breakdown points were proposed. Some of the more popular ones are listed in Appendix A.3.

¹Note that the sum of squared residual is proportional to the empirical variance of the residuals

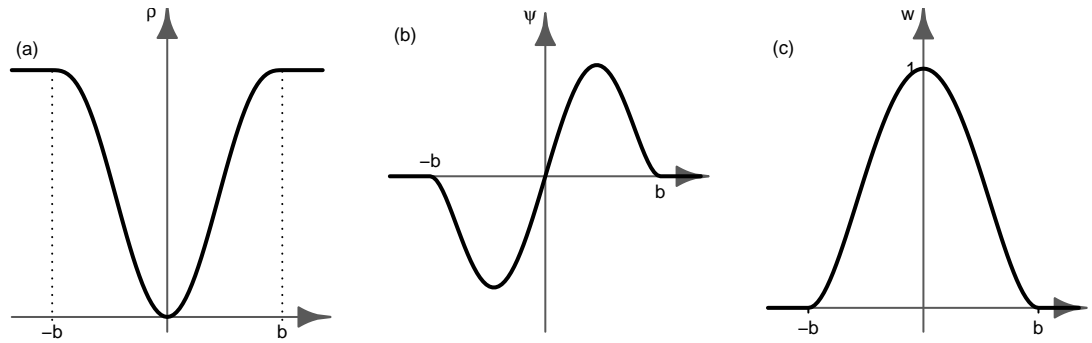


Figure 2.4.c.: ρ -, ψ - and weight function of Tukey's bisquare function (from left to right).

Definition c Tukey's Bisquare Function. In practice, a popular choice of the ρ -function is the integral of Tukey's bisquare function

$$\rho_{b_o}\langle u \rangle := \begin{cases} \frac{1}{6} - \frac{1}{6} \cdot \left(1 - \left(\frac{u}{b_o}\right)^2\right)^3 & \text{if } |u| < b_o \\ \frac{1}{6} & \text{otherwise} \end{cases}$$

with $b_o = 1.548$ and, to get a consistent estimator for normally distributed errors, $\kappa_\rho = 0.5$.

- d** The optimization problem, that must be solved to obtain the S-estimators, is very difficult. Because ρ is bounded, the function to be minimized is not convex and many local minima may emerge. Therefore, such problems are primarily solved with random resampling algorithms which find the solution just with a certain probability. This also means that when the computation is repeated with the same data, this approach may result in different estimates, which may be very awkward in practise.

Another disadvantage of S-estimators is their statistical inefficiency. Although the S-estimator is normally distributed asymptotically, its variance is more than three times larger than that of M-estimators, i.e., the relative efficiency of S-estimators is only 28.7%.

Definition e Regression MM-Estimator. However, it is possible to combine the high resistance to outliers of S-estimators with the high efficiency of regression M-estimators by taking up the idea of 2.4.a. The so-called **regression MM-estimator** (Modified M-estimator), proposed by Yohai, Stahel and Zamar (1991), is a regression M-estimator with redescending ψ function, whose initial value for $\underline{\beta}^{(o)}$ and the scale estimation s_o are taken from the above S-estimator. As a redescending ψ function, they propose Tukey's bisquare function

$$\rho'_{b_1}\langle u \rangle = \psi_{b_1}\langle u \rangle = \begin{cases} \frac{u}{b_1} \left(\left(1 - \frac{u}{b_1}\right)^2 \right)^2 & \text{if } |u| < b_1 \\ 0 & \text{otherwise} \end{cases}$$

with $b_1 = 4.687$ (see Fig. 2.4.c, in the middle). This procedure ensures (as long as $b_1 > b_o$) that the regression MM-estimator has a breakdown point of $\varepsilon^* = 1/2$ (like the initial S-estimator) and an asymptotic distribution like a regression M-estimator. With this we bring together the best statistical properties of the two estimation procedures,

although with the disadvantage that the computational effort is considerable (but not vastly larger than for the S-estimator itself). Maybe this disadvantage will become unimportant with time, since computers continually become more powerful.

* Note that MM-estimates have an unbounded influence function, although a high breakdown point. This seeming contradiction can be resolved by noting that an infinite gross-error sensitivity means only that the maximum bias of the estimates induced by altering a single observation is bounded by c/n for some constant c . In case of the MM-estimates the bias is bounded by a weaker bound c/\sqrt{n} . Hence the bias is not infinite like that of the least-squares estimates. For more details see Maronna, Martin, Yohai and Salibián-Barrera (2019, Sec. 5.5).

Example f Fund of Hedge Funds. In this example from finance, regression is applied to perform a return-based style analysis on fund of hedge funds (FoHF). A FoHF is a fund that invests in a portfolio of different hedge funds to diversify the risks associated with a single hedge fund. A hedge fund is an investment instrument that undertakes a wider range of investment and trading activities in addition to traditional long-only investment funds.

One of the difficulties in risk monitoring of FoHF is their limited transparency. Many FoHF will only disclose partial information on their underlying portfolio and the underlying investment strategy (style of FoHF), which is the crucial characterisation of FoHF, is self-declared.

A return-based style analysis searches for the combination of indices of sub-styles of hedge fund that would most closely replicate the actual performance of the FoHF over a specified time period. Such a style analysis is done basically by fitting a so-called multifactor model:

$$R_t = \alpha + \sum_{k=1}^p \beta_k I_{k,t} + E_t$$

where

R_t = return on the FoHF at time t

α = the excess return (a constant) of the FoHF

$I_{k,t}$ = the index return of sub-style k (= factor) at time t

β_k = the change in the return on the FoHF per unit change in factor k

p = the number of used sub-indices

E_t = residual (error) which cannot be explained by the factors

Figure 2.4.f shows the residuals versus time for the least squares fit (left) and the MM-fit (right).

* In R you can fit the regression model robustly using an MM-estimator for example by

```
> require(robustbase)
> FoHF.MMfit <- lmrob(FoHF ~ ., data=xy)
```

From the robust fit it is immediately clear that there are two different investment periods: one before April 2000 and one afterwards. Since the residuals are much larger in the shorter first period than in the second one, we conclude that the model does not describe the returns of the first period very well and, hence, that two different investment strategies were applied in these two periods.

Joint work with Peter Meier and his group at ZHAW (cf. Ruckstuhl and Meier, 2009).

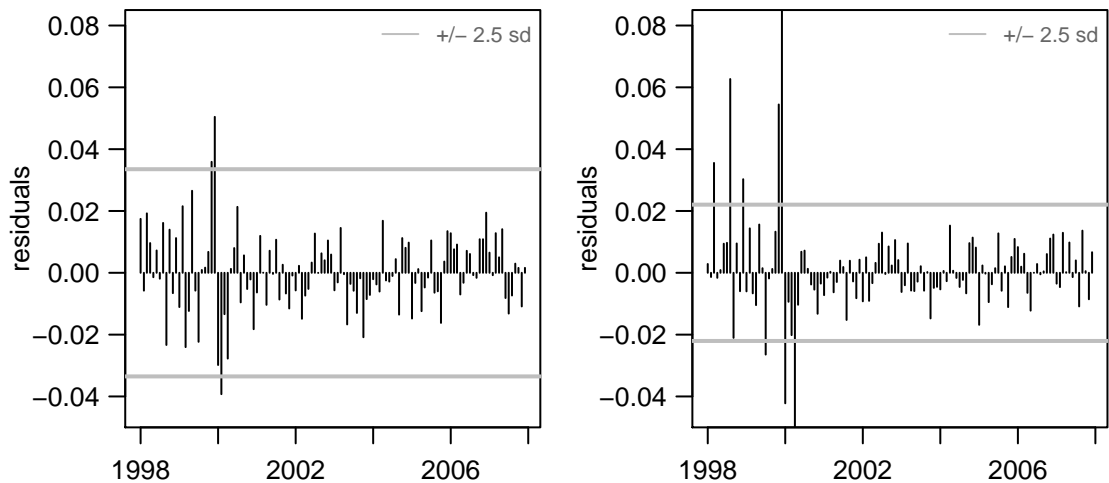


Figure 2.4.f.: Fund of Hedge Funds. Residuals versus time for the least squares fit (left) and the MM-fit (right) using `lmrob(...)` of the R package `robustbase`.

Example g Fund of Hedge Funds. The implementation of the MM-estimator in the R function `lmrob()` allows easily to run a residual analysis similar to a classical fit. The result of `plot(FoHF.rlm)` where `FoHF.rlm` is the result of `lmrob(...)` is shown in Figure 2.4.g. The major difference to the classical version is that the scatter plot of standardized residuals versus leverages is replaced by the scatter plot of standardized residuals versus robust distances of \underline{x}_i from the centre of the \underline{x} data. This robust distance corresponds to the Mahalanobis distance which is calculated using a robustly estimated covariance matrix. Some more detailed discussion on robust estimation of covariance matrices will follow in Section 4.1.

2.5 Robust Inference and Variable Selection

- a** Outliers may also influence the result of a classical test crucially. It might happen that the null hypothesis H_0 is rejected, because an interfering alternative H_I (outliers) is present. That is, the rejection of the null hypothesis is justified but accepting the actual alternative H_A is unjustified.

To understand such situations better, one can explore the effect of contamination on the level and power of tests. Heritier and Ronchetti (1994) showed that the effects of contamination on both the level and power of a test are inherited from the underlying estimator (= test statistic). That means that

the test is robust if its test statistic is based on a robust estimator.

- b Asymptotic Distribution of the MM-estimator.** In terms of the asymptotic distribution the MM-estimator is an M-estimator and hence we can apply the results of 2.1.e: The MM-estimator is asymptotically normally distributed with expected value $\underline{\beta}$. The covariance matrix is estimated analogously to 2.1.e:

$$\widehat{\mathbf{V}} = \frac{s_o^2}{n} \widehat{\boldsymbol{\tau}} \widehat{\mathbf{C}}^{-1}.$$

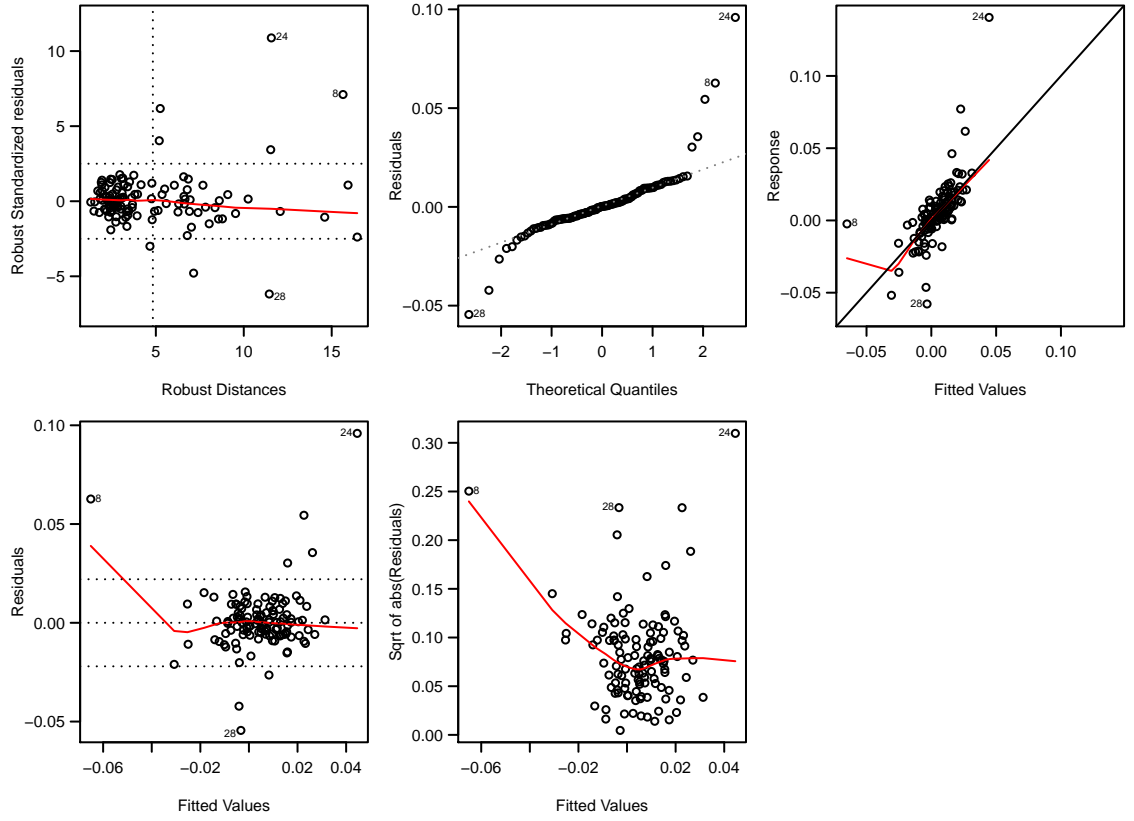


Figure 2.4.g.: Fund of Hedge Funds. The five standard plots of a residuals analysis in R when using the robust fitting procedure `lmrob()`.

* Details about estimating the **covariance matrix \mathbf{V} of the estimated parameters $\underline{\beta}$** : Let $\tilde{r}_i := (y_i - \underline{x}_i^T \underline{\hat{\beta}}^{(o)})/s_o$ and $w_i := \psi_{b_1}(\tilde{r}_i)/\tilde{r}_i$, then

$$\begin{aligned}\hat{C} &:= \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n \underline{x}_i \underline{x}_i^T w_i \\ \hat{\tau} &:= \frac{1}{n-p} \sum_{i=1}^n (\psi_{b_1}(\tilde{r}_i))^2 \left\{ \frac{1}{n} \sum_{i=1}^n \psi'_{b_1}(\tilde{r}_i) \right\}^{-2} \quad \text{and} \\ \hat{\mathbf{V}} &:= \frac{s_o^2}{n} \hat{\tau} \hat{C}^{-1}.\end{aligned}$$

Note that the estimates $\underline{\hat{\beta}}^{(o)}$ and s_o come from the initial S-estimation.

- c Based on this result we can derive approximate confidence intervals of a single parameter β_k :

$$\hat{\beta}_k \pm q_{1-\alpha/2}^{\mathcal{N}} \cdot \sqrt{\hat{V}_{kk}},$$

where $q_{1-\alpha/2}^{\mathcal{N}}$ is the $(1 - \alpha/2)$ quantile of the standard Gaussian distribution. The approximation will be better as the number of observations n is increasing.

* As common in R you can obtain these confidence intervals by

```
> confint(FoHF.MMfit)
```

	lm(FoHF ~ ., data=FoHF2)			lmrob(FoHF ~ ., data=FoHF2, setting="KS2014")		
	Estimate	se	Pr(> t)	Estimate	se	Pr(> t)
(I)	-0.0019	0.0017	0.2610	-0.0030	0.0014	0.0377
RV	0.0062	0.3306	0.9850	0.3194	0.2803	0.2564
CA	-0.0926	0.1658	0.5780	-0.0671	0.1383	0.6280
FIA	0.0757	0.1472	0.6083	-0.0204	0.1279	0.8730
EMN	0.1970	0.1558	0.2094	0.2721	0.1328	0.0430
ED	-0.3010	0.1614	0.0655	-0.4763	0.1389	0.0009
EDD	0.0687	0.1301	0.5986	0.1019	0.1112	0.3611
EDRA	0.0735	0.1882	0.6971	0.0903	0.1583	0.5689
LSE	0.4407	0.1521	0.0047	0.5813	0.1295	2.05e-05
GM	0.1723	0.0822	0.0390	-0.0159	0.0747	0.8319
EM	0.1527	0.0667	0.0245	0.1968	0.0562	0.0007
SS	0.0282	0.0414	0.4973	0.0749	0.0356	0.0378
	Residual standard error: 0.009315			Residual standard error: 0.007723		

The 95% confidence interval of β_{SS} is

$0.028 \pm 1.99 \cdot 0.041 = [-0.054, 0.110]$	$0.075 \pm 1.96 \cdot 0.036 = [0.004, 0.146]$
where 1.98 is the 0.975 Quantile of t_{89}	where 1.98 is the 0.975 Quantile of $\mathcal{N}(0, 1)$

Table 2.5.e.: Fund of Hedge Funds II. Summary outputs of both the least squares estimator and the SMDM-estimator. Significant (level of 5%) coefficients are highlighted by bold P-values.

- d** The estimated covariance matrix \hat{V} still contains some elements of the initial S-estimator: $\hat{\beta}^{(o)}$ and s_o . Because it is known that the S-estimator is very inefficient, we may suspect some side-effects on the confidence intervals. Koller and Stahel (2011, 2017) investigated this construction of confidence intervals with respect to the efficiency of the estimated confidence intervals and, as a consequence of this, came up with an additional modification. They extended the two steps “S-estimator” and “M-estimator with redescending ψ -function” of the MM-estimator with two additional steps in which a more efficient scale estimator replaces s_o and an M-estimator with a more slowly redescending ψ -function is applied. They called this estimation procedure **SMDM-estimator**.

SMDM-estimator

This estimation method has been implemented in the function `lmrob` of the R-package `robustbase` (Rousseeuw, Croux, Todorov, Ruckstuhl, Salibián-Barrera, Verbeke, Koller and Maechler, 2011, version 0.6-5 or younger). With the argument `setting="KS2011"`, the recommended parameters are set automatically. With the option `setting="KS2014"` even more arguments are changed. This setting should produce more stable estimates for designs with factor variables. More details are found in the help page of `lmrob.control` or in Koller and Stahel (2011, 2017).

Example e Fund of Hedge Funds II. We again perform a return-based style analysis as in 2.4.f but use another target fund of hedge fund. In Table 2.5.e, the corresponding summary outputs of both the least squares estimator and the SMDM-estimator are shown. According to the P-values, we come to different conclusions depending on the estimating procedure. We obtain three significant variables with the least squares method and five significant variables (without the intercept) with the robust method. The two sets of significant variables are not subsets of each other. Hence, we may attribute quite different styles to this fund of hedge funds.

- f** We now turn to testing the hypothesis that $q < p$ of the p parameters in the parameter vector $\underline{\beta}$ are 0. To present the test procedures efficiently, we introduce the following notation: It is no loss of generality to assume that all parameters that might be 0 are at the beginning of the parameter vector. The null hypothesis is then $H_0 : \underline{\beta}_1 = \underline{0}$,

where the whole parameter vector can be represented as $\underline{\beta} = (\beta_1^T, \beta_2^T)^T$. Also, $\widehat{\mathbf{V}}_{11}$ is the square submatrix that contains the first q row and columns of $\widehat{\mathbf{V}}$.

g Wald Test Statistic. The so-called Wald Test Statistic is

$$W = \underline{\beta}_1^T (\widehat{\mathbf{V}}_{11})^{-1} \underline{\beta}_1.$$

It can be shown that this test statistic is asymptotically χ^2 distributed with q degrees of freedom, where q is the dimension of $\underline{\beta}$.

Note: By using an estimation $\widehat{\mathbf{V}}$ of the covariance matrix, we have actually introduced additional variability into the test statistic. In classical theory, such an effect is compensated for, in that the χ_q^2 distribution is replaced by the $F_{q,n-p}$ distribution. Naturally, we can also do this for the above test statistic. However, such a procedure has, in this case, no formal basis, since asymptotic results are involved. It is therefore desirable to avoid small sample sizes (and then it holds that $\frac{1}{q} \cdot \chi_q^2 \approx F_{q,n-p}$).

The confidence intervals that were introduced in 2.1.e and 2.5.b are based on this Wald test statistic.

Example h Fund of Hedge Funds II. Funds of hedge funds (FoHF) may be classified by the style of their target funds into *focussed directional*, *focussed non-directional* or *diversified*.

If our considered FoHF is a *focussed directional* FoHF, then it should be invested in LSE, GM, EM, SS and hence the other parameters should be zero.

In R Output 2.5.h a classical model comparison based on the F-test (which is identical to the classical version of the Wald test) and the analogue robustified version are applied. In this example we arrive at contradictory conclusions. Hence the outliers have a crucial influence on the decision whether this fund of hedge fund is a focussed directional one or not.

i Robust Deviance. For a regression MM-estimator, we can define a robust deviance

$$D\langle \underline{y}, \underline{\hat{\beta}}_{\text{MM}} \rangle = 2 \cdot s_o^2 \cdot \sum_{i=1}^n \rho \left\langle \frac{y_i - \underline{x}_i^T \underline{\hat{\beta}}_{\text{MM}}}{s_o} \right\rangle.$$

Based on F-tests from classical regression,

$$\frac{(SS_{\text{reduced}} - SS_{\text{full}})/q}{SS_{\text{full}}/(n-p)} = \frac{(SS_{\text{reduced}} - SS_{\text{full}})/q}{\hat{\sigma}^2},$$

we get a robust analogue by replacing the sum of squares with the robust deviance (as in generalized linear models)

$$\begin{aligned} \Delta^* &= \tau^* \cdot \frac{D\langle \underline{y}, \underline{\hat{\beta}}_{\text{MM}}^r \rangle - D\langle \underline{y}, \underline{\hat{\beta}}_{\text{MM}}^f \rangle}{s_o^2} \\ &= 2 \cdot \tau^* \cdot \left(\sum_{i=1}^n \rho \left\langle \frac{y_i - \underline{x}_i^T \underline{\hat{\beta}}_{\text{MM}}^r}{s_o} \right\rangle - \sum_{i=1}^n \rho \left\langle \frac{y_i - \underline{x}_i^T \underline{\hat{\beta}}_{\text{MM}}^f}{s_o} \right\rangle \right) \\ &\text{with } \tau^* = \left(\frac{1}{n} \sum_{i=1}^n \psi'_{b_1}(\tilde{r}_i) \right) / \left(\frac{1}{n} \sum_{i=1}^n (\psi_{b_1}(\tilde{r}_i))^2 \right). \end{aligned}$$

```

> FoHF2.lm1 <- lm(FoHF ~ ., data=FoHF2)
> FoHF2.lm2 <- lm(FoHF ~ EMN + LSE + GM + EM + SS, data=FoHF2)
> anova(FoHF2.lm2, FoHF2.lm1)

Analysis of Variance Table

Model 1: FoHF ~ LSE + GM + EM + SS
Model 2: FoHF ~ RV + CA + FIA + EMN + ED + EDD + EDRA + LSE + GM + EM + SS

      Res.Df      RSS    Df  Sum of Sq      F  Pr(> F)
1         96 0.0085024
2         89 0.0077231    7  0.00077937  1.2831  0.2679

## Robust with SMDM-estimator
> FoHF2.rlm1 <- lmrob(FoHF ~ ., data=FoHF2, setting='KS2014')
> anova(FoHF2.rlm1, FoHF ~ LSE + GM + EM + SS, test='Wald')

Robust Wald Test Table

Model 1: FoHF ~ RV + CA + FIA + EMN + ED + EDD + EDRA + LSE + GM + EM + SS
Model 2: FoHF ~ LSE + GM + EM + SS
Largest model fitted by lmrob(), i.e. SMDM

      pseudoDf  Test.Stat  Df  Pr(> chisq)
1           89
2           96      24.956   7   0.0007727 ***

```

R-Output 2.5.h: Fund of Hedge Funds II. The hypothesis whether this fund of hedge fund is of a *focussed directional* type is tested using a classical and robust Wald test.

This test statistic is asymptotically χ^2 distributed with q degrees of freedom under the null hypothesis: $\Delta^* \stackrel{a}{\sim} \chi_q^2$ (Richardson and Welsh, 1996).

Example j Fund of Hedge Funds II. In Table 2.5.j this robust deviance test is applied to the hypothesis of 2.5.h. Both robust test procedures yield the same conclusion: According to the return patterns it is not a focussed directional fund of hedge fund.

```

> anova(FoHF2.rlm1, FoHF ~ EMN + LSE + GM + EM + SS, test='Deviance')

Robust Deviance Table

Model 1: FoHF ~ RV + CA + FIA + EMN + ED + EDD + EDRA + LSE + GM + EM + SS
Model 2: FoHF ~ LSE + GM + EM + SS
Largest model fitted by lmrob(), i.e. SMDM

      pseudoDf  Test.Stat  Df  Pr(> chisq)
1           89
2           96      25.089   7   0.0007318 ***

```

R-Output 2.5.j: Fund of Hedge Funds II. The hypothesis whether this fund of hedge fund is of a *focussed directional* type is tested using a robust deviance test.

k Variable Selection. The variable selection can now be carried out stepwise with the robust deviance test. As is usual with such procedures, we have the question of when we should stop omitting more variables. In practice, we omit variables until all are significant at the 5% level. This rule, however, cannot be theoretically justified.

A theoretically sound criterion is, e.g., the Akaike Information Criterion (AIC), which is based on the sum of squared residuals as a criterion for accuracy, but which is not robust to outliers. For a robust analysis, there is a similar procedure based on Akaike's "final prediction error" criterion (cf. Maronna et al., 2019, Sec. 5.2.6):

$$\text{RFPE}\langle C \rangle := \frac{1}{n} \sum_{i=1}^n \rho \left\langle \frac{r_i^C}{\hat{\sigma}_f} \right\rangle + \frac{q}{n} \hat{\tau}$$

with C the set of variables considered, q the number of variables considered, and $\hat{\sigma}_f$ the robust scale estimate based on the full set of variables. The correction factor $\hat{\tau}$ is calculated as in 1.4.a, except that $\hat{\sigma}$ is replaced by $\hat{\sigma}_f$.

Example 1 Fund of Hedge Funds II. We apply this robust variable selection criterion to the Fund of Hedge Fund dataset of 2.5.e and 2.5.h:

```
> step(FoHF2.lm1)
...shortened ...
FoHF ~ EMN + ED + LSE + GM + EM
      Df    Sum of Sq    RSS    AIC
<none>      1    0.00032529 0.0078582 -943.59
- EMN      1    0.00032529 0.0081835 -941.50
- ED       1    0.00043180 0.0082900 -940.19
- GM       1    0.00049510 0.0083533 -939.42
- EM       1    0.00059036 0.0084486 -938.28
- LSE      1    0.00123463 0.0090928 -930.85

## Robust variable selection criterion
> library(RobStatTM)
> h.cont <- lmrobdet.control(bb=0.5, efficiency=0.85, family="bisquare")
> FoHF2.rlm1 <- lmrobdetMM(FoHF ~ ., data=FoHF2, control=h.cont)
> FoHF2.rfpe <- step.lmrobdetMM(FoHF2.rlm1)
...shortened ...
Model: FoHF ~ RV + EMN + ED + LSE + EM + SS
scale: 0.007999167
      Df    RFPE
<none>      1    0.20127
RV       1    0.20384
EMN      1    0.20371
ED       1    0.22507
LSE      1    0.23117
EM       1    0.22365
SS       1    0.20724
```

This two procedures do not select the same variables, but agree on 4 variables.

Recap m

When is which estimation method used? Least squares estimations are unsatisfactory when the data set contains observations that “heavily” deviate from the used model. In regression it is therefore safer to use

- the regression MM-estimator `lmrob(...)`
- or if very reliable estimates of the confidence intervals are needed `lmrob(..., setting="KS2014")`
This setting should produce more stable estimates for designs with factor variables as the setting `setting="KS2011"`.

Because the regression MM-estimator is computationally expensive, you may prefer the regression M-estimator (`rlm(..., method="M")`) in case of a controlled design (i.e., all predictor variables are factor variables like in the example from molecular spectroscopy) or in case of simple regression like in the air quality example.

The function `lmrob` is found in the R package `robustbase` and `rlm` in the R package `MASS`. Further implementations will be mentioned at the end.

3 Generalized Linear Models

3.1 Unified Model Formulation

GLM **a** Generalized linear models (**GLM**) were formulated by John Nelder and Robert Wedderburn as a way of unifying various statistical regression models, including linear (Gaussian) regression, logistic regression, Poisson regression and Gamma regression. The generalization is based on a reformulation of the linear regression model. Instead of

$$Y_i = \theta_0 + \theta_1 \cdot x_i^{(1)} + \dots + \theta_p \cdot x_i^{(p)} + E_i, \quad i = 1, \dots, n, \quad \text{with } E_i \text{ indep. } \sim \mathcal{N}\langle 0, \sigma^2 \rangle$$

the model is formulated as

$$Y_i \text{ indep. } \sim \mathcal{N}\langle \mu_i, \sigma^2 \rangle \quad \text{with } \mu_i = \eta_i,$$

linear predictor where η_i is the so-called **linear predictor** $\eta_i = \theta_0 + \theta_1 \cdot x_i^{(1)} + \dots + \theta_p \cdot x_i^{(p)}$. The expectation μ_i may be linked to the linear predictor η_i by another function than the identity function. In general, we assume $g\langle \mu_i \rangle = \eta_i$, where $g\langle \cdot \rangle$ is called **link function**.

Using this structure of modelling the measurement, we can generalise the linear regression model to the family of GLMs. In this section we set the focus on three families of the GLMs, the discrete generalized linear models (binary / binomial regression model and the Poisson regression model), and Gamma regression. A forth member of the family, the linear Gaussian regression model, has been discussed in great detail in the previous sections.

b For the **binomial regression model** ($Y_i \text{ indep. } \sim \mathcal{B}\langle \pi_i, m_i \rangle$) we then have that

$$\mathbb{E}\left\langle \frac{Y_i}{m_i} | \underline{x}_i \right\rangle = \pi_i \quad \text{and} \quad \text{Var}\left\langle \frac{Y_i}{m_i} | \underline{x}_i \right\rangle = \frac{\pi_i \cdot (1 - \pi_i)}{m_i}.$$

Common link functions are

$$\begin{aligned} g\langle \pi_i \rangle &= \log \langle \pi_i / (1 - \pi_i) \rangle && \text{Logit Model} \\ g\langle \pi_i \rangle &= \Phi^{-1} \langle \pi_i \rangle && \text{Probit Model} \\ g\langle \pi_i \rangle &= \log \langle -\log \langle 1 - \pi_i \rangle \rangle && \text{Complementary log-log Model.} \end{aligned}$$

c For the **Poisson regression model** ($Y_i \text{ indep. } \mathcal{P}\langle \lambda_i \rangle$) we have

$$\mathbb{E}\langle Y_i | \underline{x}_i \rangle = \lambda_i \quad \text{and} \quad \text{Var}\langle Y_i | \underline{x}_i \rangle = \lambda_i.$$

The following link functions are used:

$$\begin{aligned} g\langle \lambda_i \rangle &= \log \langle \lambda_i \rangle && \text{log-linear Model} \\ g\langle \lambda_i \rangle &= \lambda_i && \text{identity} \\ g\langle \lambda_i \rangle &= \sqrt{\lambda_i} && \text{square root.} \end{aligned}$$

- d** With the **gamma regression model** (Y_i indep. $\sim \text{Gamma}(\alpha_i, \beta_i)$) we have

$$\mathbb{E}\langle Y_i | x_i \rangle = \frac{\alpha_i}{\beta_i} \quad \text{and} \quad \text{Var}\langle Y_i | x_i \rangle = \frac{\alpha_i}{\beta_i^2}.$$

Common link functions are

$$\begin{aligned} g\langle \mu_i \rangle &= \frac{1}{\mu_i} && \text{inverse} \\ g\langle \mu_i \rangle &= \log \langle \mu_i \rangle && \text{log-lineare model} \\ g\langle \mu_i \rangle &= \mu_i && \text{identity.} \end{aligned}$$

variance function
dispersion parameter

- e** Before we turn to the estimation equations, we factorize the variance of the response variable in a convenient way: $\text{Var}\langle Y_i \rangle = \phi V\langle \mu_i \rangle$. The second factor $V\langle \mu_i \rangle$ is called **variance function** and is a function of the expectation $\mu_i = \mathbb{E}\langle Y_i \rangle$ only. The first factor ϕ , called **dispersion parameter**, is a scale factor and is independent of the expectation μ_i . It is equal to 1 for the binomial regression model and the Poisson regression model, respectively. In the gamma regression model, ϕ is equal to $\frac{1}{\alpha_i}$ and hence the variance function is $V\langle \mu_i \rangle = \mu_i^2$. In the linear Gaussian regression model, ϕ is equal to σ^2 and hence the variance function is a constant function, $V\langle \mu_i \rangle = 1$.

- f Estimation Equations.** Based on the maximum likelihood principle the classical estimation equations of all generalized linear model families can be written in the unified form

$$\sum_{i=1}^n \frac{y_i - \mu_i}{V\langle \mu_i \rangle} \mu'_i x_i = 0,$$

where $\mu'_i = \partial \mu\langle \eta_i \rangle / \partial \eta_i$ is the derivative of the inverse link function.

The values $\tilde{r}_i := \frac{y_i - \mu_i}{\sqrt{V\langle \mu_i \rangle}}$, $i = 1, \dots, n$ are called Pearson residuals. Their variance depends on the leverage H_{ii} as it does in the linear regression model:

$$\text{Var}\left\langle \frac{y_i - \mu_i}{\sqrt{V\langle \mu_i \rangle}} \right\rangle \approx \phi \sqrt{1 - H_{ii}}.$$

Hence, if there is no leverage point in the explanatory variables, i.e., all H_{ii} are about equally sized, the variance of the Pearson residuals is approximately constant.

3.2 Robust Estimating Procedure

- a** In GLM, we face similar problems with the standard estimator as in linear regression problems at the presence of contaminated data. In Figure 3.2.a a simple binary regression with the logit-link is shown. As soon as the two most left observations (filled circles) are misclassified (i.e., they move from 1 to 0), the estimated probabilities π (solid line) changes (dashed line) clearly.

Because our eyes are struggling with assessing non-linear structure, the real impact of the misclassification is difficult to assess.

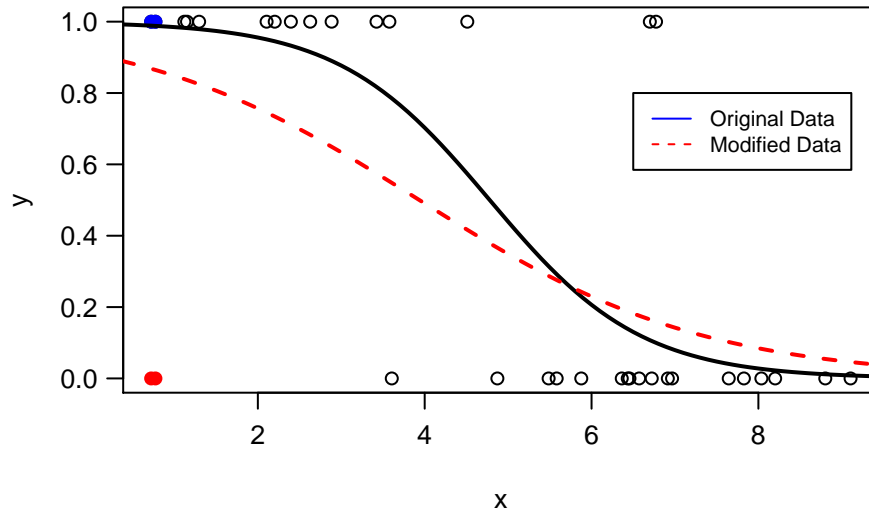


Figure 3.2.a.: Simple binary regression with the logit-link: This example shows a clear impact on the estimated probabilities π when two observations are misclassified: solid line when the two most left observations are at 1 and dashed line when the two most left observations are at 0.

- b Mallows quasi-likelihood estimator (Mqle).** This robust estimator is essentially a general regression M-estimator (see subsection 2.3), which retains the structure of the classical estimators. The Mqle is the solution of the system of equations

$$\underline{0} = \sum_{i=1}^n \left(\psi_c \langle \tilde{r}_i \rangle \frac{\mu'_i}{\sqrt{V \langle \mu_i \rangle}} \underline{x}_i w \langle \underline{x}_i \rangle - \underline{\text{fcc}} \langle \underline{\beta} \rangle \right),$$

where $\psi_c \langle \cdot \rangle$ is Huber's ψ function. The vector valued constant $\underline{\text{fcc}} \langle \underline{\beta} \rangle$ ensures the Fisher consistency of the estimator and is given by

$$\underline{\text{fcc}} \langle \underline{\beta} \rangle = \frac{1}{n} \sum_{i=1}^n \left(\sum_{k=0}^{m_i} \psi_c \langle r_i \rangle P \langle Y_i = k \rangle \right) \frac{\mu'_i}{\sqrt{V \langle \mu_i \rangle}} \underline{x}_i w \langle \underline{x}_i \rangle,$$

for binomial and Poisson regression ($m_i = \infty$). In case of the Gamma regression the inner sum is replaced by an integral.

If $w \langle \underline{x}_i \rangle = 1$ for all observations i , then the influence of a leverage point is not bounded (it is thus a BIR estimator). To limit the total influence function, we can set, e.g., $w \langle \underline{x}_i \rangle = \sqrt{1 - H_{ii}}$. However, as was discussed in section 2.3, such a choice does not lead to a high breakdown point estimator. It appears promising to use the inverse of the Mahalanobis distance of \underline{x}_i as $w \langle \underline{x}_i \rangle$, where the Mahalanobis distance is based on a robust covariance matrix estimation with a high breakdown point. Unfortunately, this procedure also does not lead to a robust estimator with a guaranteed high breakdown point.

- c Asymptotic Distribution of the Mqle.** The Mqle is asymptotically normally distributed with expected value $\underline{\beta}$ and asymptotic variance $\underline{\Omega}$, which is a rather elaborate expression. Its estimation can be calculated easily by a computer. Based on this estimation $\hat{\underline{\Omega}}$, Wald-type test can be performed and confidence intervals can be derived as described in 2.5.g.

* The asymptotic variance Ω is a product of three matrices,

$$\Omega = M^{-1} \langle \psi_c, F \rangle Q \langle \psi_c, F \rangle M^{-1} \langle \psi_c, F \rangle.$$

The matrices $Q \langle \psi_c, F \rangle$ and $M \langle \psi_c, F \rangle$ can be estimated by

$$\begin{aligned} Q \langle \psi_c, F \rangle &= \frac{1}{n} \mathbf{X}^T \mathbf{A} \mathbf{X} - \text{fcc} \langle \underline{\beta} \rangle \text{fcc} \langle \underline{\beta} \rangle^T \\ M \langle \psi_c, F \rangle &= \frac{1}{n} \mathbf{X}^T \mathbf{B} \mathbf{X} \end{aligned}$$

The matrices \mathbf{A} and \mathbf{B} are diagonal matrices with diagonal elements

$$\begin{aligned} A_{ii} &= \left(\sum_{k=0} \psi_c^2 \langle r_i \rangle P \langle Y_i = k \rangle \right) \frac{(\mu'_i)^2}{V \langle \mu_i \rangle} w^2 \langle x_i \rangle \\ B_{ii} &= \left(\sum_{k=0} \psi_c \langle r_i \rangle \frac{\partial \log \langle P \langle Y_i = k | \mu_i \rangle \rangle}{\partial \mu_i} P \langle Y_i = k \rangle \right) \frac{(\mu'_i)^2}{\sqrt{V \langle \mu_i \rangle}} w \langle x_i \rangle \end{aligned}$$

This appears relatively complicated, but can be calculated easily with the computer.

d Robust Quasi-Deviance. For the Mqle we can define a robust quasi-deviance with which a deviance type test can be constructed like in the classical case. The distribution of this test statistics is, however, more elaborate as in the classical setting. It is distributed like a linear combination of χ_1^2 distributions.

* The robust quasi-deviance is defined as

$$D \left\langle \underline{y}, \hat{\underline{\beta}}_{\text{Mqle}} \right\rangle = -2 \cdot \sum_{i=1}^n Q_m \langle y_i, \mu_i \rangle,$$

where

$$\begin{aligned} Q_m \langle y_i, \mu_i \rangle &= \int_{\tilde{s}}^{\mu_i} \psi_c \langle r_i \rangle \langle t \rangle \frac{1}{\sqrt{V \langle t \rangle}} w \langle x_i \rangle dt \\ &\quad - \frac{1}{n} \sum_{j=1}^n \int_{\tilde{t}}^{\mu_i} \mathbb{E} \left\langle \psi_c \langle r_i \rangle \langle t \rangle \frac{1}{\sqrt{V \langle t \rangle}} w \langle x_i \rangle \right\rangle dt \end{aligned}$$

and $r_i \langle \mu_i \rangle = \frac{y_i - \mu_i}{\sqrt{V \langle \mu_i \rangle}}$ are the Pearson residuals. The integration limits \tilde{s} and \tilde{t} are determined so that

$$\psi_c \langle r_i \rangle \langle \tilde{s} \rangle \frac{1}{\sqrt{V \langle \tilde{s} \rangle}} = 0$$

and also

$$\mathbb{E} \left\langle \psi_c \langle r_i \rangle \langle \tilde{t} \rangle \frac{1}{\sqrt{V \langle \tilde{t} \rangle}} \right\rangle = 0.$$

Thus we can now define a **robust measure Λ for the difference** between two nested models,

$$\Lambda = D \left\langle \underline{y}, \hat{\underline{\beta}}_{\text{Mqle}}^{\text{red}} \right\rangle - D \left\langle \underline{y}, \hat{\underline{\beta}}_{\text{Mqle}}^{\text{full}} \right\rangle = 2 \left(\sum_{i=1}^n Q_m \langle y_i, \mu_i^{\text{full}} \rangle - \sum_{i=1}^n Q_m \langle y_i, \mu_i^{\text{red}} \rangle \right).$$

The test statistic Λ is asymptotically distributed as $\sum_{k=1}^q \lambda_k Z_k^2$, where Z_k i.i.d. $\sim \mathcal{N}(0, 1)$ and λ_k are eigenvalues of a still to be specified matrix.

In a quadratic approximation of the robust quasi-deviance test statistic, however, we assume that all λ_k are equally large. Then $\Lambda/\bar{\lambda}$ is asymptotically approximated by a χ^2 distribution with q degrees of freedom.

- e The Mqle estimator is implemented in the R function `glmrob`. The function `summary` of an object that has been returned by `glmrob` summarizes the inference of the parameter estimations. The inference results are based on the asymptotic distribution results given in 2.5.g.

The comparison between two nested models can be done with the R function `anova.glmrob`, which is called by `anova` of a `glmrob` object. The different robust test methods are addressed by a character string in the argument `test`. With the argument `test="Wald"` the “Wald-type” test between two nested models is applied. A robust quasi-deviance test is used when `test="QD"` and its approximate test statistic is applied when `test="QDapprox"`.

All of these functions can be found in R package `robustbase`: Fitting is done by

```
glmrob(Y ~ ..., family=..., data=...,
       weights.on.x=c("none", "hat", "robCov", "covMcd"))
```

and testing can be done by

```
anova(Fit1, Fit2, test=c("Wald", "QD", "QDapprox"))
```

The implementation is based on several papers of Cantoni und Ronchetti which are also summarized in Heritier, Cantoni, Copt and Victoria-Feser (2009).

4 Multivariate Analysis

4.1 Robust Estimation of the Covariance Matrix

- a For continuous random variables, the normal distribution is the most common model - among other reasons, because the mathematical theory for this distribution gives nice, simple results. In multivariate statistics, the corresponding model of the multivariate normal distribution again plays a very central roll, since other models are only plausible in special applications.
- b The multivariate normal distribution is determined by its expected value $\underline{\mu}$, a vector, and the covariance matrix $\underline{\Sigma}$. Obvious estimations for these parameters are the arithmetic mean $\bar{\underline{X}}$ and the empirical covariance matrix $\hat{\underline{\Sigma}}$ (see e.g. Stahel, 2000, Section 15.2). They are optimal when the normal distribution holds exactly – the same as in the one dimensional case.

Example c Painted Turtles. Jolicoeur and Mosimann studied the relationship of size and shape for painted turtles. They measured carapace length, width and height of 24 male turtles (all information is in mm). Since these variables consist of positive numerical values, we will analyse the log-transformed variables with a multivariate normal distribution. For demonstration purposes we first look at only the two variables $\log(\text{length})$ and $\log(\text{width})$ and shift observation 24 to the point P1 (4.66, 4.4). As you see in Figure 4.1.c , the covariance estimation is very sensitive to gross error.

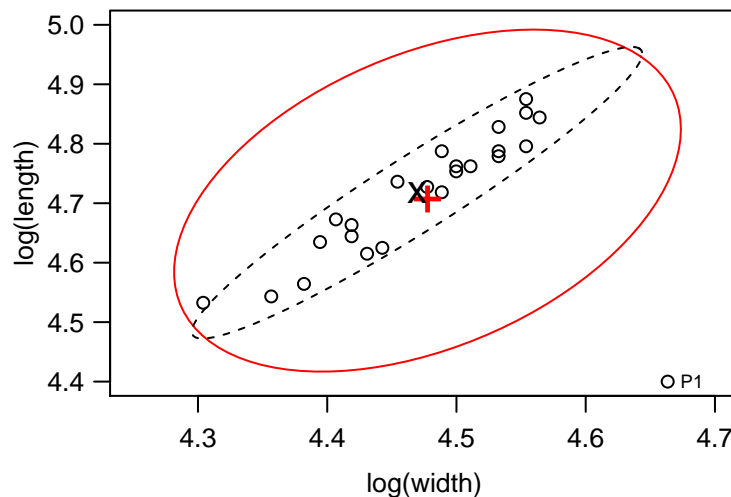


Figure 4.1.c.: Covariance matrices of the modified painted turtle data: The covariance matrices are represented by an ellipse that contains 95% of the density mass. Once with (solid, $\hat{\underline{\mu}} = +$) and once without (dashed, $\hat{\underline{\mu}} = \times$) observation P1.

This example comes from Johnson and Wichern (1998). You will find it there under the keyword “carapaces (painted turtles).”

- d How do we identify gross errors?** If only two variables are studied, we only need to look at the scatterplot and the problematic points can be seen immediately. In the painted turtle example (cf. Figure 4.1.c), one gross (lower right) and one “milder” (lower left) outlier are apparent.

In higher dimensions, it doesn’t always suffice to look at a scatterplot matrix and we must call in appropriate mathematical tools. Generally we can define an observation \underline{x}_i as an outlier with respect to a model (i.e., a probability distribution), if its probability density at \underline{x}_i is very, very small. For the normal distribution this is the case if $u_i = (\underline{x}_i - \underline{\mu})^T \underline{\Sigma}^{-1} (\underline{x}_i - \underline{\mu})$ is large. This quantity is also called squared **Mahalanobis Distance**¹ to the center $\underline{\mu}$. The corresponding random variable U_i is χ_m^2 distributed if $\underline{X}_i \sim \mathcal{N}(\underline{\mu}, \underline{\Sigma})$, where m is the number of variables, i.e., the length of the vector $\underline{\mu}$.

We can thus check the model by calculating the u_i values from the observations and comparing them with the χ_m^2 distribution, e.g. with a quantile-quantile (QQ) plot (see Figure 4.1.d for the painted turtle example). For this we would estimate the unknowns $\underline{\mu}$ and $\underline{\Sigma}$; and thus the χ_m^2 distribution holds only approximately.

* The plot in Figure 4.1.d is produced by

```
> xx <- data.matrix(turM[,c("lnWidth", "lnLen44")])
> xx.cov <- cov(xx)
> xx.D2 <- mahalanobis(x=xx, center=apply(xx, 2, mean), cov=xx.cov)
> qqplot(qchisq(ppoints(100), df=2), xx.D2, main="", las=1,
+        xlab=expression("Quantiles of the" ~chi[2]^2 ~"distribution"),
+        ylab="Squared Mahalanobis Distance")
> abline(0, 1, col="gray")
```

If the data can be modelled by a 2-dimensional Gaussian distribution, the data should scatter around the 45° line. Because of the outlier the estimation is distorted and as a consequence the non-outlying points do not scatter nicely around the 45° line.

- e Classical Procedure.** We examine the QQ plot of u_i versus the quantile of the χ_m^2 distribution and delete observations for which u_i is “too high” above the line. We then estimate the covariance matrix again. It’s possible that new outliers become visible and we have to repeat the above approach. This approach may be effective when there is a single outlier or a few wilde ones. But as in the case of location it can be useless when the number of observations n is small or several outliers may mask one another as in regression or in outlier tests (cf. 1.2.i). And even if we are successful in removing the outliers, the classical inference applied to the cleaned data is too optimistic because it ignores the data cleaning process (cf. 1.1.d).

Do robust estimation procedures help us proceed?

- f** When robustly estimating the covariance matrix we face similar challenges as in multiple regression. Approaches that proved their values in the one dimensional case (location model – scale estimation), either have poor robustness properties in the

¹For simplicity u_i is sometimes referred to as “distance”, although it should be kept in mind that it is actually a *squared* distance.

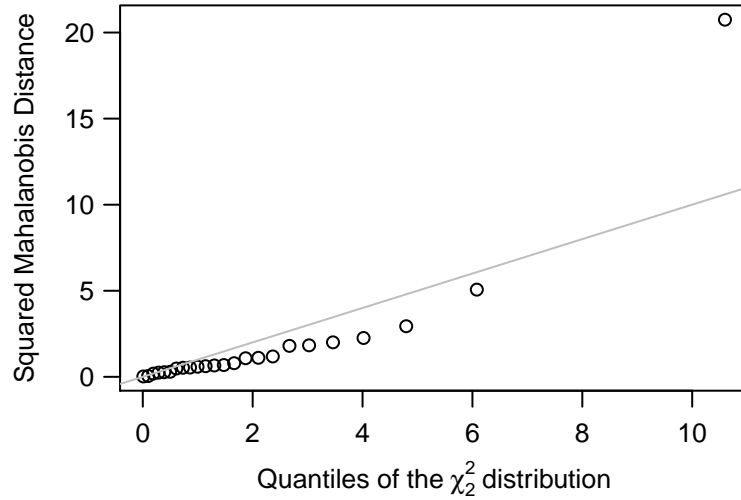


Figure 4.1.d.: QQ plot of the squared Mahalanobis distances u_i versus quantiles of the χ^2_2 distribution for the modified painted turtle data. The Mahalanobis distances are based on the classical estimation of the covariance matrix. The grey line is the 45° line.

multidimensional case (M-estimators have a low breakdown point) or they cannot be generalized to the multidimensional case at all (e.g., MAD).

- g Estimators Based on a Robust Scale.** Just with the regression estimates where we aimed at making the residuals “small”, we shall define multivariate estimates of location and dispersion that make the distance u_i “small”. To formulize such an approach, it helps to factorize the covariance matrix Σ into a scale parameter σ and the shape matrix Σ^* with $|\Sigma^*| = 1$: $\Sigma = \sigma^2 \cdot \Sigma^*$. Based on this factorization we can calculate a scaled version of the squared Mahalanobis distance,

$$d_i := d(\underline{x}_i, \underline{\mu}, \Sigma^*) := (\underline{x}_i - \underline{\mu})^T (\Sigma^*)^{-1} (\underline{x}_i - \underline{\mu}),$$

$i = 1, \dots, n$. If the observations \underline{x}_i are normally distributed with expectation $\underline{\mu}$ and covariance matrix Σ , then the variance of d_i is $2 \cdot m \cdot \sigma^2$, because $d_i/\sigma^2 \sim \chi^2_m$.

Estimators of $\underline{\mu}$ and Σ^* can be defined by minimizing a scale estimator $S\langle \rangle$, i.e.,

$$[\hat{\underline{\mu}}, \hat{\Sigma}^*] = \text{minimum (over } \underline{\mu} \text{ and } \Sigma^*) \text{ of } S\langle d(\underline{X}, \underline{\mu}, \Sigma^*) \rangle.$$

To obtain a robust estimation of $\underline{\mu}$ and Σ^* , a robust scale estimator $S\langle \rangle$ must be used.

- h MVE Estimator.** The simplest case of a scale estimator $S\langle \rangle$ is the sample median of d_i , $i = 1, \dots, n$, which are all non-negative, as it is also applied, e.g., in the MAV (see 2.1.f). The resulting location and dispersion matrix estimate is called the **Minimum Volume Ellipsoid estimator (MVE estimator)**. The name stems from the fact that among all ellipsoids containing at least half of the data points, the one given by the MVE estimator has minimum volume, i.e., the minimum $|\Sigma|$.

This estimator has a high breakdown point of $1/2$, but a consistency rate as low as the LMS (see Appendix A.3), namely only $n^{-1/3}$, and hence is very inefficient.

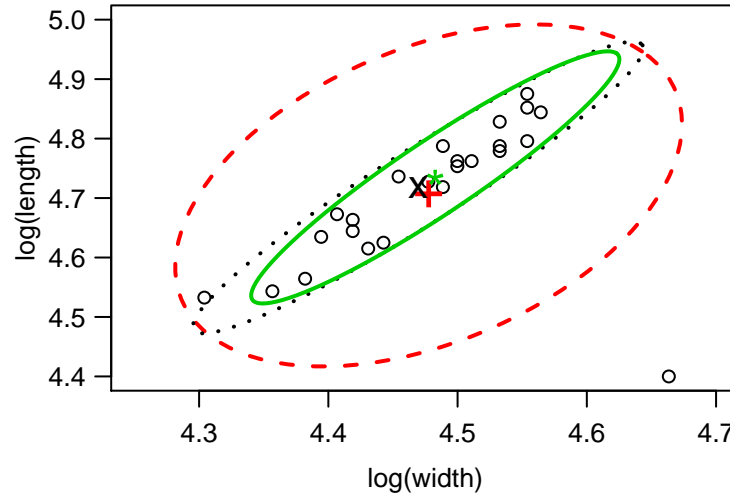


Figure 4.1.j.: Estimations of the covariance matrix from the modified painted turtle data: The classically estimated covariance matrices are represented by ellipses that contain 95% of the density mass: Once with observation P1 (dashed line, $\hat{\mu} = +$) and once without (dotted line, $\hat{\mu} = x$). The solid line ($\hat{\mu} = *$) is the robustly (MCD) estimated covariance matrix for all of the data (including P1).

- i MCD Estimator.** Another possibility is to use a trimmed scale estimator for $S \langle \rangle$ as was done to define the LTS estimator in Appendix A.3.c. Let $d_{(1)} \leq \dots \leq d_{(n)}$ be the ordered values of the squared distances d_i , $i = 1, \dots, n$. For an integer h with $0 \leq h \leq n$ define the trimmed scale of the squared distances as

$$S_{TS} \langle d_i \rangle = \sum_{i=1}^h d_{(i)}.$$

An estimator defined by 4.1.g and this trimmed scale is called a **Minimum Covariance Determinant (MCD) estimator**. The maximum break-down point of the MCD estimator is attained by taking $h = n - \lceil (n - m)/2 \rceil$ and is about $1/2$. This estimator is more efficient than the MVE estimator and should be preferred. As all high break-down estimator, the computation is very expensive and also requires a stochastic search algorithms.

Example j Painted Turtles. Based on the robust MCD estimation of the covariance matrix including all the data (i.e., including P1), two outliers were identified and “down-weighted” in the estimating procedure. Since the resulting ellipse in Figure 4.1.j is smaller than this one from the classical estimator excluding the outlier P1, the second outlier must have some negative influence on the classical estimation as well.

Example k Painted Turtles. We can again calculate Mahalanobis distances, but this time they are based on the robust estimated covariance matrix. In Figure 4.1.k the Mahalanobis distances are displayed against the observation number for the *painted turtle* example. The horizontal line represents the square root transformed 97.5% quantile of the χ^2_2 distribution. Observations that lie above this line can be considered as outliers. As shown in Figure 4.1.k yet another outlier must be considered (observation 1).

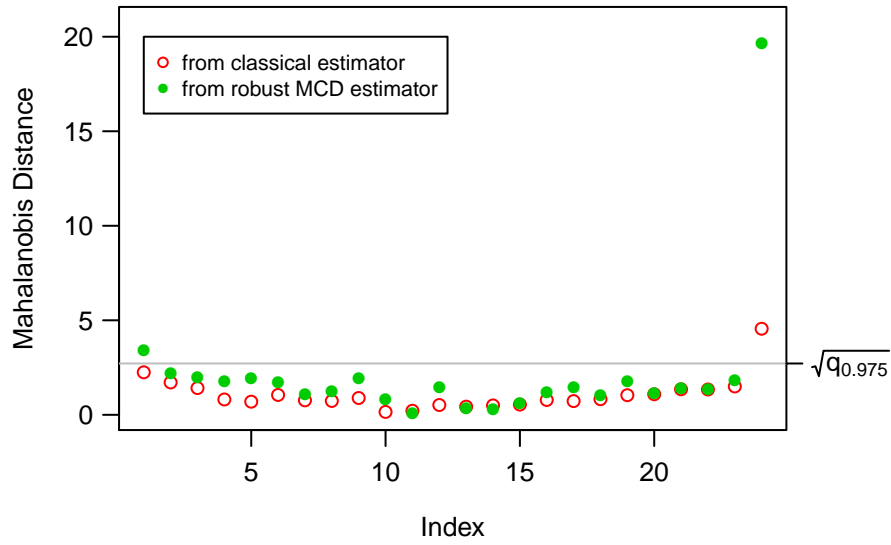


Figure 4.1.k.: Mahalanobis distances for the observations from the modified painted turtle example. Observations that lie above the square root transformed 97.5%- χ^2_2 quantile line can be considered to be outliers.

Other Approaches

- l S-Estimator.** The S-estimator $S\langle d_i \rangle$ is also based on a robust scale estimator and satisfies

$$\frac{1}{n-m} \sum_{i=1}^n \rho \left\langle \frac{d_i}{S\langle d_i \rangle} \right\rangle = \frac{1}{2}$$

where $\rho\langle u \rangle$ is the adequately adjusted bisquare function. It has a high breakdown-point but needs a stochastic search algorithm as well.

- m Stahel-Donoho Estimator.** If we want simultaneous a high breakdownpoint and a highly efficient estimator for normally distributed data we must inevitably consider the **Stahel-Donoho estimator**. It is defined as a weighted covariance matrix estimator, where the weights of each observation are a measure of its maximal one dimensional extremity. This measure is based on the idea that if an observation is a multivariate outlier, there must be some one dimensional projection in which the multivariate outlier is apparent as a one dimensional outlier.

A still serious drawback of this estimator is the effort involved in its calculation. The algorithmic complexity is at least an order of magnitude greater than for, for example, the MCD estimator.

- n Orthogonalized Gnanadesikan-Kettenring (OGK) Estimator.** If one wants to work on really high dimensional data, the above approaches are far too slow. Much faster estimates with high breakdown point can be computed if one gives up the requirements of affine equivariance of the covariance matrix. Such an algorithm was proposed based on pairwise covariances. A suitable robust estimate of pairwise covariances is

$$c^{(x,y)} = \frac{1}{4} \left((S\langle x+y \rangle)^2 - (S\langle x-y \rangle)^2 \right),$$

where $S\langle \cdot \rangle$ is a robust estimation of σ . Since a matrix of such estimates does not yield a semi-definite matrix, a correction is needed to obtain a suitable estimate of the

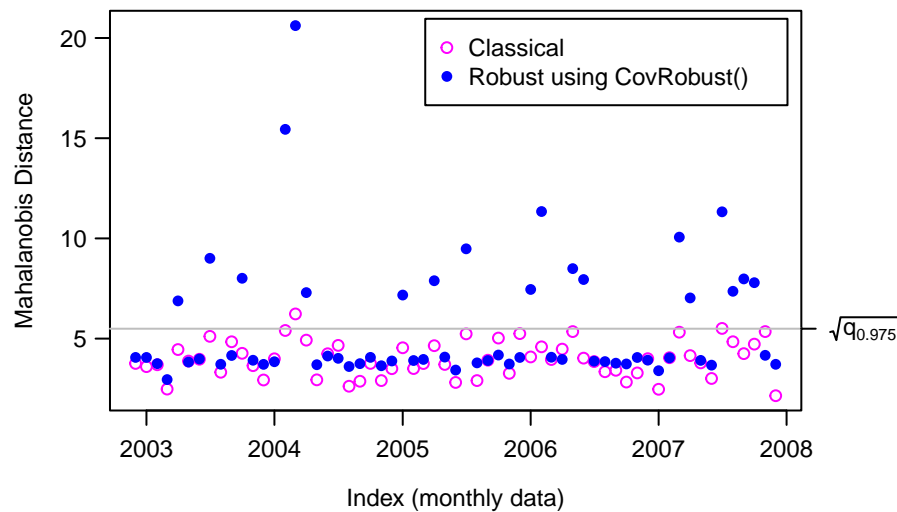


Figure 4.1.p.: Mahalanobis distances for the observations from 17 focused directional FoHF. Observations that lie above the square-root of the 97.5%- χ^2_2 quantile line can be considered to be outliers.

covariance matrix.

o R functions. We recommend to use the R function

- `CovRobust(..., control="auto")` from R package `rrcov`

Using "auto" selects an appropriate method according to the size of the dataset:

- Stahel-Donoho estimator if dataset $< n = 1'000 \times p = 10$ or $< 5'000 \times 5$
- S-estimator if dataset $< 50'000 \times 20$
- Orthogonalized Quadrant Correlation (=OGK) if $n > 50'000$ and/or $p > 20$

Implementation details can be found in the vignette of the R Package `rrcov` written by Valentin Todorov and Peter Filzmoser (see `library(help=rrcov)` to find information where the vignette is located).

Other estimator introduced above can be found in

- `covMcd(...)` and `covOGK(...)` from R package `robustbase`
- `cov.rob(..., method="mcd")` from R package `MASS`

Example p Focused Directional FoHF. In this example monthly returns of 17 funds of hedge funds (FoHF) are analysed. According to a self-declaration all of them run a “focused directional” strategy. The Mahalanobis distances of data covering 61 consecutive months are analysed classically and by a robust procedure (with `CovRobust` from package `robustbase`). Applying the classical procedure one weak outlier is visible (cf. Figure 4.1.p). On the other hand, the robust procedure identifies 18 strong outliers.

q Conclusion. Multivariate statistical analysis often is based on the covariance matrix, since the multivariate normal distribution is the only workable model conception. Classically optimal estimation, however, is very sensitive to outliers. Thus, robust estimation methods are also needed here. In the function `CovRobust` from the R package `rrcov` are suitable methods implemented depending on the size n and the dimension p of the data set.

4.2 Principal Component Analysis

- a Objectives of Principal Component Analysis.** The goals of principal component analysis can be various, e.g.
- Reduction of the dimension by elimination of directions (= linear combinations of the original variables), that only explain a small portion of the variance.
 - Identification of structures (subgroups and outliers).
 - Linear transformation of explanatory variables to avoid collinearity (cf. principal component regression).
 - Fit a q -dimensional (q is smaller than the dimensionality of the data) hyperplane which fits the data well; i.e., the orthogonal deviations of the observations to the hyperplane is as small as possible (original idea of Karl Pearson, 1901)
- b** The principal components give, in descending importance, the directions in which the data display a large (important) and respectively small (unimportant) spread. Each principal component is orthogonal to all the others.
- c Principal Component Analysis and Robustness.** As long as we use the principal components descriptively, the model of the multivariate normal distribution stays far in the background. This has its advantages, but also the disadvantage that we don't exactly know how to robustify this procedure, because there is no underlying model. We can thus hold the opinion that robust procedures are not necessary at all in principal component analysis. Explorative principal component analysis is based mainly on looking for unusual structures in two dimensional scatter plots of each two principal components. But just as outliers (and other structures) don't have to show up in the individual coordinates, they also don't necessarily show up in two dimensional projections.
- d Robust Standardized Variables.** A first possibility for applying ideas of robustness to principal component analysis is to standardize the individual variables with robust methods (so, e.g., median and MAD) and carry out the principal component analysis of the standardized variables. (Note that in this case, no further classical standardization should be done; i.e. the principal component analysis is carried out with the covariance matrix.) One dimensional deviations from the normal distribution in the individual variables thus become readily apparent.
- e Principal Component Analysis (PCA) with Robust Estimation.** A more cumbersome procedure is to base the principal component analysis on a robustly estimated covariance or correlation matrix. Outliers can then be determined with the Mahalanobis distances that are calculated with the robust estimate covariance matrix. The principal components, on the other hand, truly represent in descending order the directions where the variation of the majority of the data (i.e. without considering outliers) is large.
- In the scatter plots of the principal components, we can then investigate structures that are not based on projections influenced by outliers.

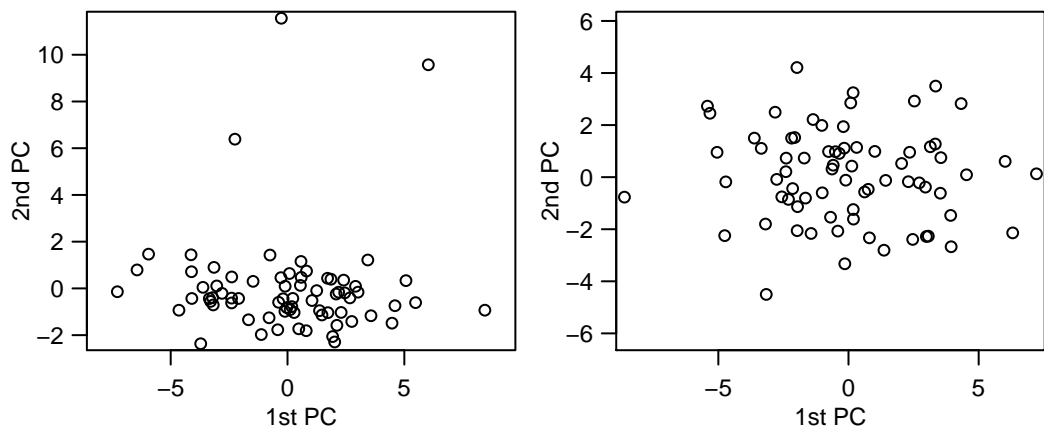


Figure 4.2.f.: Scatter plots of the first two principal components based on the classical approach (left) and on a robustly estimated covariance matrix (right). In the classical approach, the outliers are clearly visible whereas in the robust approach they are not.

Example f To illustrate the difference between classical PCA and PCA based on a robustly estimated covariance matrix, an example is simulated in R Output 4.2.f. In the scatter plot of the first two classical principal components, the three outliers are clearly visible, whereas these outliers cannot be identified in the scatter plot of the first two robust principal components. The Mahalanobis distances based on the robustly estimated covariance matrices (cf. 4.1.p) must be used. Consistently, the loading of this two approaches are very different. The order of the principal components based on the robust estimation corresponds with the simulated data without the outliers. The first principal component of the classical approach is dominated by the three outliers.

```
> library(MASS)
> library(rrcov)
> library(mvtnorm)
> set.seed(4711)
> mN <- rmvnorm(n=72, mean = c(-2,0, 1), sigma = diag(c(1,12,3)))
> mN[c(29,30,31),1] <- c(8, 5, 10)

> mN.pc <- princomp(mN, cor=FALSE)
> mN.Rpc <- PcaCov(mN, scale=FALSE)
> mN.pc.p <- predict(mN.pc)
> mN.Rpc.p <- predict(mN.Rpc)

> par(mfrow=c(1,2)), las=1
> eqscplot(mN.pc.p[,1:2])
> eqscplot(mN.Rpc.p[,1:2])

## Loadings: classical (left), robust (right)
> structure(cbind(loadings(mN.pc)[,1:3],
                    mN.Rpc@loadings[,1:3]), class="loadings")

Loadings:
      Comp.1  Comp.2  Comp.3      PC1      PC2      PC3
[1,]          0.998          0.996
[2,]  0.977      -0.212  0.978 -0.203
[3,]  0.212      0.975  0.206  0.976
```

R-Output 4.2.f: R-code to illustrate the difference in scatter plots of the first two principal components when they are based on the classical approach and on robustly estimated covariance matrix, respectively.

- g Conclusion.** Principal component analysis that is based on robust estimated covariance or correlation matrices is another tool in multivariate explorative data analysis. The motivation for such a procedure, however, is less clear than that for the robust fitting of regression models.

4.3 Linear Discriminant Analysis

- a Discriminant Analysis** is a multivariate data analysis technique where a (graphical) representation is sought so that the difference between various classes or groups can be clearly shown. This is a predominantly exploratory approach and is often used if there is no (or little) theoretical or causal knowledge available about the difference between the classes.

If we have found such a representation, it is also possible to assign new observations to the classes on the basis of this representation. Then we are talking about **classification**.

- b Fisher's Linear Discriminant Analysis** has the goal of finding the linear combination of descriptive variables that leads to maximum separation between the class centers. The distance between the class centers is determined relative to the variability within the classes. (Details about this can be found in any book about multivariate statistics.)

Let \mathbf{W} be the covariance matrix within the class and \mathbf{B} the covariance matrix of the class centers. The optimal linear combination \underline{a}_1 is then derived from

$$\underline{a}_1 = \arg \max_{\underline{a}} \frac{\underline{a}^T \mathbf{B} \underline{a}}{\underline{a}^T \mathbf{W} \underline{a}}.$$

The solution is $\underline{a}_1 = \mathbf{W}^{-1/2} \underline{e}_1$, where \underline{e}_1 is the eigenvector for the largest eigenvalue of the matrix

$$\mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2}.$$

It is obvious not only to consider the direction where the classes are separated best, but also directions $\underline{a}_2 = \mathbf{W}^{-1/2} \underline{e}_2$, $\underline{a}_3 = \mathbf{W}^{-1/2} \underline{e}_3$, ... ($\underline{e}_1, \underline{e}_2, \underline{e}_3, \dots$ are orthogonal to each other), which separate them next best. That is, we are facing a problem analogous to principal component analysis: the "covariance" matrix is $\mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2}$

- Example c Fleas.** Lubischew (1962) measured the Aedeagus (part of the male genital apparatus) of three subspecies of fleas, in order to identify the subspecies. On the one hand, the maximal width of the front part (variable `width` in micrometers) and on the other hand the front angle of the gripping apparatus in units of 7.5 deg (variable `angle`) were determined. In total, 74 fleas were measured, that belonged to one of the three subspecies *chaetocnema concinna*, *chaetocnema heikertingeri* or *chaetocnema heptapotamica*.

Can the fleas be separated into three types solely on the basis of these two body features? To demonstrate the effect of outliers, the existing data was contaminated with 11 outliers.

In Figure 4.3.c the flea data are shown in the first two discriminant variables (the k -th discriminant variable consists of the values $z_i^{(k)} = \underline{a}_k^T \underline{x}_i$, $i = 1, 2, \dots$). If all conditions are met, the individual classes must appear as round clusters that each represent a

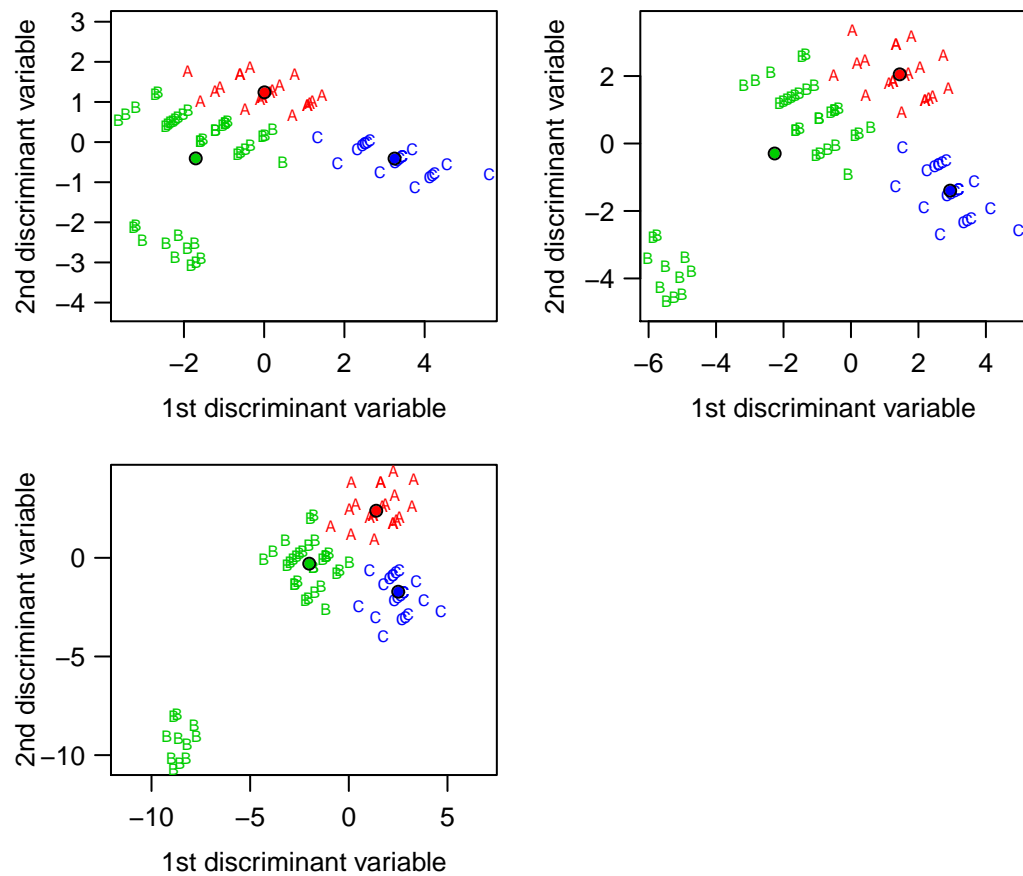


Figure 4.3.c.: Flea Example: Representation of the data in the first two discriminant variables. The discriminant variables are estimated classically (above left), with robust estimated covariance matrix \mathbf{W} (above right), and with robust estimations for the class centers and covariance matrix.

multivariate standard normal distribution. If the assumption that all classes can be described by the same (up to the location) multivariate normal distribution is violated, correspondingly distorted class clusters appear. As is obvious in Figure 4.3.c, the outliers strongly distort the class clusters in classical discriminant analysis.

- d First Approach to Robustification.** The covariance matrix \mathbf{W} obviously represents the normal distribution of the individual classes, in contrast to the matrix \mathbf{B} . If we are more interested in a graphical analysis of the group centers and their position in respect to each other, then we *barely have an underlying model*, of how the (usually few) class centers should behave (see discussion in principal component analysis). Thus a possibility for robustifying linear discriminant analysis could be that we use robust estimation only for the covariance matrix \mathbf{W} and keep the classical estimation of the matrix \mathbf{B} (with \mathbf{B} you would probably rather talk about a “geometric representation” than about an estimation of a covariance matrix!). This idea from Venables and Ripley (1999) is implemented in the R function `lda(..., method="mve")`. There the robust MVE estimator is used for the covariance matrix \mathbf{W} .

Example e Fleas. In Figure 4.3.c above right is shown the solution of the procedure with the function `lda(..., method="mve")` from the R package `MASS`. The class clusters are

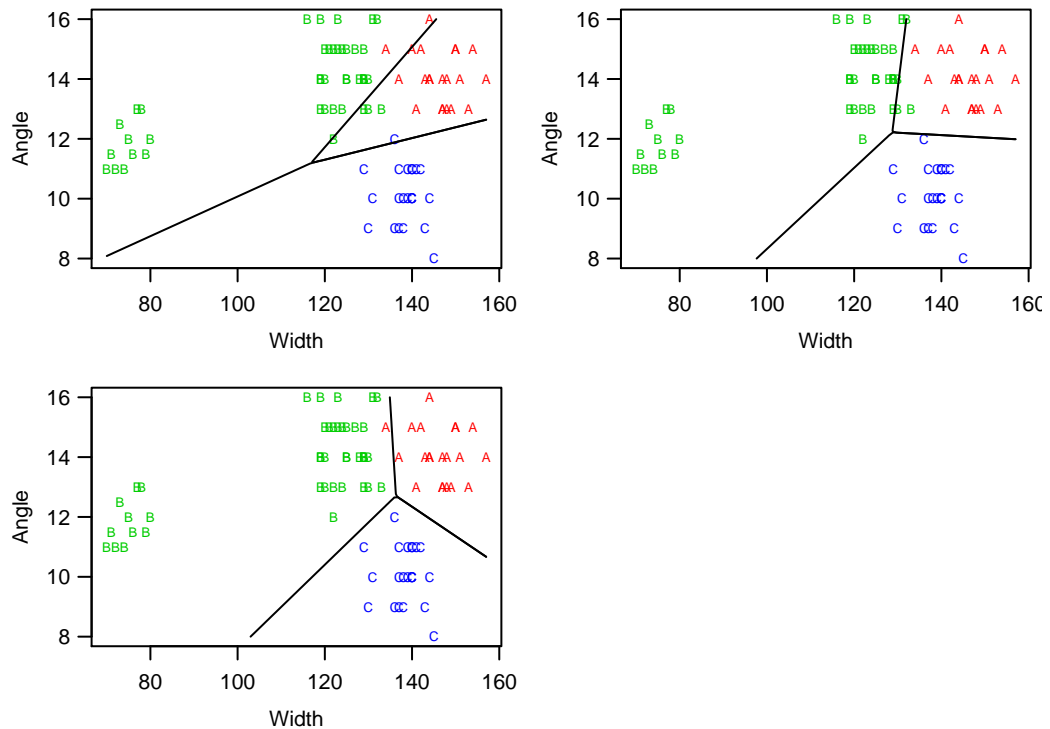


Figure 4.3.g.: Flea example: Illustration of the data in the two variables width and angle. The class borders are estimated classically (above left), with robust estimated covariance matrix \mathbf{W} (above right) and with robust estimations for the class centers and covariance matrix \mathbf{W} .

now more or less round, except for class B. However, the center of class B still doesn't lie in the main part of the data.

- f Refinement of the Robustification.** The procedure `lda(..., method="mve")` can still be improved if the class centers are also robustly estimated. This is realized in the R function `rlda(...)`. Both the estimation of the centers as well as the covariance matrix \mathbf{W} are based on the robust MCD estimator. In its present form, however, the implementation is very slow, because for K classes, the R function `rlda(...)` calls the function `cov.mcd(...)` $(K + 1)$ times.

The illustration below left in Figure 4.3.c shows the corresponding solution for the **flea example**. This is barely influenced by the outliers.

Example g Fleas. In the illustration of the two discriminant variables (Figure 4.3.c), the borders between the classes are given by the bisectors between the class centers. We back transform these into the scatter plots of the original variables and thus get lines as borders, as are shown in Figure 4.3.g. The solution from the procedure with robust estimations of both the class centers and the covariance matrix is the most convincing in this example.

- h Outlook.** In this chapter we have shown robustified discriminant analysis in a simple example. The space spanned by all possible discriminant variables here has dimension two ($= \min\{K - 1, p\}$, where K is the number of classes and p the number of descriptive variables). As soon as more than three groups and more than two descriptive variables are available, the discriminant space can be larger than two dimensional. Then in discriminant analysis we are also interested in, among other things, finding a

two dimensional subspace where the classes can be distinguished as well as possible. Whether and to what extent discriminant analysis is helpful in this context will not be discussed here. However, experience with principal component analysis can be applied here.

- Recap i** **Multivariate statistical analysis are often based on the covariance matrix,** because the multivariate Gaussian distribution is such a convenient model.
- Robust Estimators of the covariance matrix with breakdown point of $1/2$ are able to detect multidimensional outliers fast and reliably.
 - The clearer a procedure is based on a model the better the procedure can be robustified. Exploratory use of PCA is not based on a model.
 - Principal component analysis (PCA), which is based on a robustly estimated covariance matrix, may yield however additional insight.
 - If there are outliers, the *robustified* linear discriminant analysis (LDA) shows the difference between the groups clearer and estimates the class borders more reliable.

5 Baseline Removal Using Robust Local Regression

5.1 A Motivating Example

- Example a From Mass Spectroscopy.** The spectrum we consider as a motivating example was taken from a sample of sheep blood. The instrument used was a so called SELDI TOF (Surface Enhanced Laser Desorption Ionisation, Time Of Flight) Mass Spectrometer. The spectrum in Figure 5.1.a consists of sharp features superimposed upon a continuous, slowly varying baseline. The goal of the analysis of such spectra is to study features. But to do so we first must remove the baseline. There are many different approaches which solve this task. Some of them depend strongly on the properties of the spectra and their subject matter background.
- b** We will introduce here an approach which is applicable more subject matter independent using robust local regression. But let us first introduce the “LOWESS” approach as a version of robust local regression.

5.2 Local Regression

- Example a Chlorine.** The investigation involved a product A, which must have a fraction of 0.50 of available chlorine at the time of manufacture. The fraction of available chlorine in the product decreases with time. Since theoretical calculations are not feasible, a study was run to get some insight into the decrease. Figure 5.2.a shows the data we

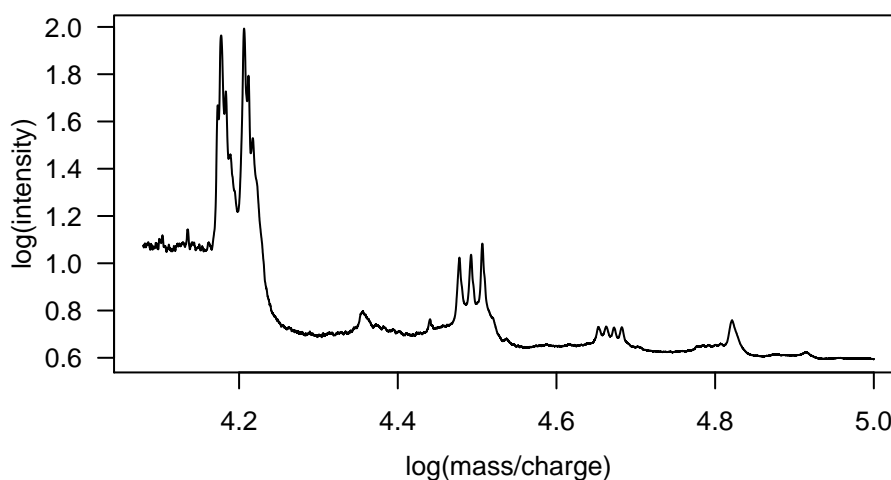


Figure 5.1.a.: The spectrum of a sample of sheep blood. The instrument used was a so called SELDI TOF (Surface Enhanced Laser Desorption Ionisation, Time Of Flight) Mass Spectrometer.

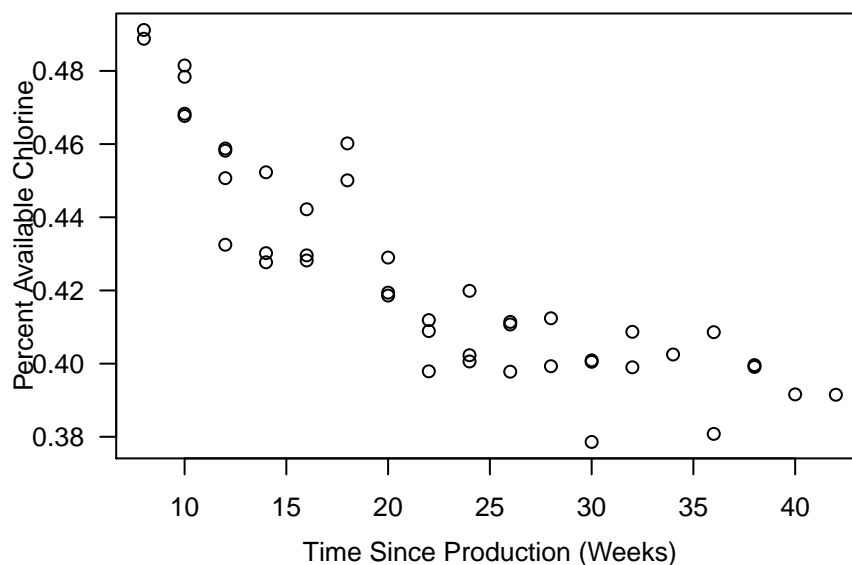


Figure 5.2.a.: Fraction of available chlorine in a product over time.

will analyse.

- b** In regression analysis we study

$$Y_i = h\langle x_i; \underline{\beta} \rangle + E_i.$$

The unstructured deviations from the function h are modelled by random errors E_i which are normally distributed with mean 0 and constant variance σ^2 .

In **linear** regression, the function $h\langle x_i; \underline{\beta} \rangle$ takes on the most simple structure – a linear structure in the unknown parameters:

$$h\langle x_i; \underline{\beta} \rangle = \beta_0 + \beta_1 \tilde{x}_i.$$

What can be done, if the function h is **nonlinear** with respect to the parameter $\underline{\beta}$? – There are two principally different approaches:

- If the structure of the nonlinear function is known but some parameters are unknown, then we can apply nonlinear regression analysis (cf. next block course). You can find good examples for that in chemistry.
- If there is little known about the structure, we would like to explore the relationship based on the given data by a “*smoother*”.

- c Local Regression – The Basic Idea.** The basic idea of local regression is that every function h can be approximated linearly within a small neighbourhood of a given point. Hence we choose a window around a point z_1 where we want to determine $h\langle z_1 \rangle$. The function h should be approximated very well by a straight line within this window. Then we fit a straight line on those points which are within the window, and predict h at z_1 . This prediction estimates the value $h\langle z_1 \rangle$. (See also Figure 5.2.c.)

This procedure is repeated for a set of points z_i , $i = 1, \dots, N$, which span the range of x values as well as possible. This results in predictions $\hat{h}\langle z_1 \rangle, \dots, \hat{h}\langle z_N \rangle$ at the points

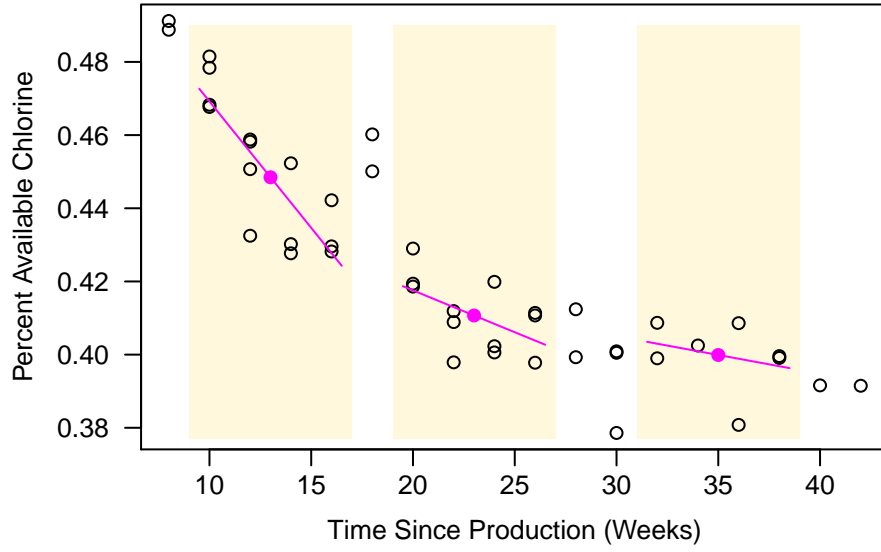


Figure 5.2.c.: Example Chlorine. Sketch of the idea of local regression based on three non-overlapping local windows.

z_1, \dots, z_N of the corresponding function values $h\langle z_1 \rangle$, $i = 1, \dots, N$. To visualize the function h , the points are connected by line segments.

- d** This procedure can be further improved by including a weight which downweights the influence of an observation according to its distance from the point z_1 . If the window around z_1 is defined by $z_1 \pm b_w$ (b_w is called **bandwidth**), then the regression problem within a window can be expressed as a weighted least-square problem. The estimated function value at z_1 is $\hat{h}(z_1) = \hat{\beta}_0$, where $\hat{\beta}_0$ is the first component of

$$\hat{\underline{\beta}}(z_1) = \arg \min_{\underline{\beta}} \sum_{i=1}^n \mathbf{w}_r \langle x_i \rangle K \left\langle \frac{x_i - z_1}{b_w} \right\rangle (y_i - (\beta_0 + \beta_1 (x_i - z_1)))^2.$$

$K \langle (x_i - z_1)/b_w \rangle$ is called **kernel weight** for the observation i . In the **lowess** procedure (and **loess**, the extension to multivariate problems) the weights are defined by applying Tukey's tricube kernel function

$$K \left\langle \frac{x_i - z_1}{b_w} \right\rangle = \left[\max \left\{ 1 - \left| \frac{x_i - z_1}{b_w} \right|^3, 0 \right\} \right]^3.$$

K is zero outside $z_1 \pm b_w$.

- e** To achieve robustness against outliers in y direction, robustness weights $\mathbf{w}_r \langle x_i \rangle$ can be included. They are implicitly defined, e.g., by Tukey's biweight robustness weight function

$$\mathbf{w}_r \langle x_i \rangle = \left(\max \left\langle 1 - \left(\frac{\tilde{r}_i}{b} \right)^2, 0 \right\rangle \right)^2 \quad \text{with } \tilde{r}_i = \frac{y_i - \hat{h} \langle x_i \rangle}{\hat{\sigma}_{\text{MAV}}} \quad \text{and } b = 4.05.$$

(Note that b is not the bandwidth but a tuning constant as used in the previous chapters.)

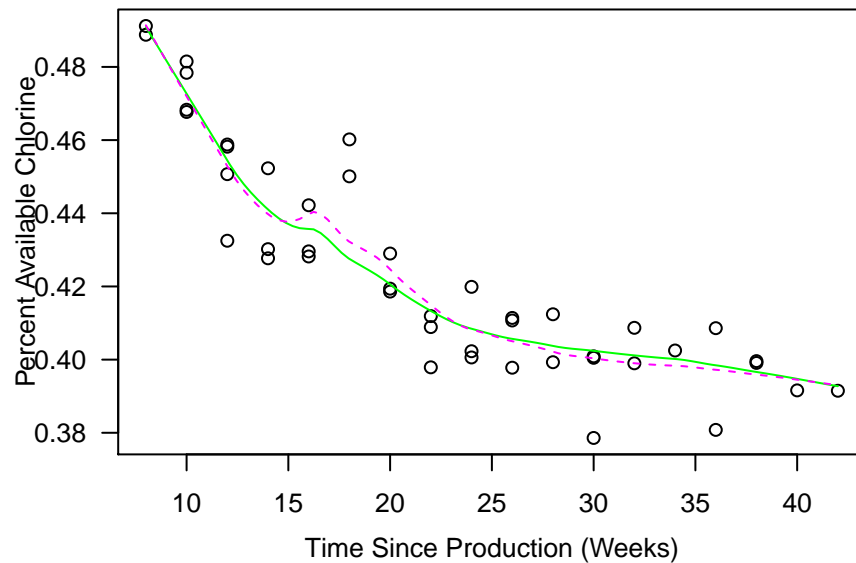


Figure 5.2.h.: Example Chlorine: LOWESS-solution for windows containing $f = 35\%$ of the data. The dotted line indicates the non-robust solution whereas the full line the robust solution.

- f** Far more important for the local regression estimator is the choice of the (half) window width b_w . If b_w is very small the approximation error will be very small - what is advantageous. But the disadvantage is that variance of the prediction $\hat{h}(z_1)$ is large because the number of points in the window will be small. If the window width b_w is (too) large, then the linear approximation may be insufficient and the approximation error becomes big. In other words: If b_w is small, the data are interpolated (not a very smooth estimate of h). However, if b_w is too big, the resulting curve is practically identical to the global least squares solution, i.e., a straight line.
- g** To ensure that there are always enough points in the window, the window width should be made adaptive. An obvious suggestion is to choose the window width b_w such that a fixed number d of points should be included in the window. It is recommended that about $d = 2/3 \cdot n$ points are in the window, where n is the number of observations. (Often we express the number of points in the window as a fraction f , hence $d = f \cdot n$.) But often it is more useful to try several different values.

Example h Chlorine. The LOWESS procedure described above is applied to the chlorine data with $f = 0.35$ with and without robustness weights, respectively (cf. Figure 5.2.h and R Output 5.2.h for the R code used). The solutions are almost identical except between 14 and 20 weeks. There are two outliers which influence the classical solution.

i * Outline of the LOWESS/LOESS Procedure

1. Select a grid of points z_1, \dots, z_N which cover the exploratory variable x .
2. Predict $\hat{h}(z_k)$ for each z_k by fitting a local regression. Use linear interpolation to predict $\hat{h}(x_i)$ at all points x_1, \dots, x_n .
3. Estimated the scale parameter by the median of absolute values ($\hat{\sigma}_{MAV}$) and calculate the robustness weights $w_r(x_i)$.
4. Run local regressions again on the grid z_k , $k = 1, \dots, N$ and predict $\hat{h}(x_i)$, $i = 1, \dots, n$ by linear interpolations.
5. If a robust estimation is needed, steps 3 and 4 are repeated until convergence is reached.

```

> plot(YY ~ x, data=Chlor, xlab="Time Since Production (Weeks)",
       ylab="Percent Available Chlorine")
> xnew <- seq(8, 42, length=100)
> clr <- loess(YY ~ x, data=Chlor, span=0.35, degree=1, family='gaussian')
> lines(xnew, predict(clr, xnew), col='magenta')
> rlr <- loess(YY ~ x, data=Chlor, span=0.35, degree=1,
              family='symmetric')
> lines(xnew, predict(rlr, xnew), col='green')

```

R-Output 5.2.h: R code for fitting and graphing local regression in the Chlorine Example.

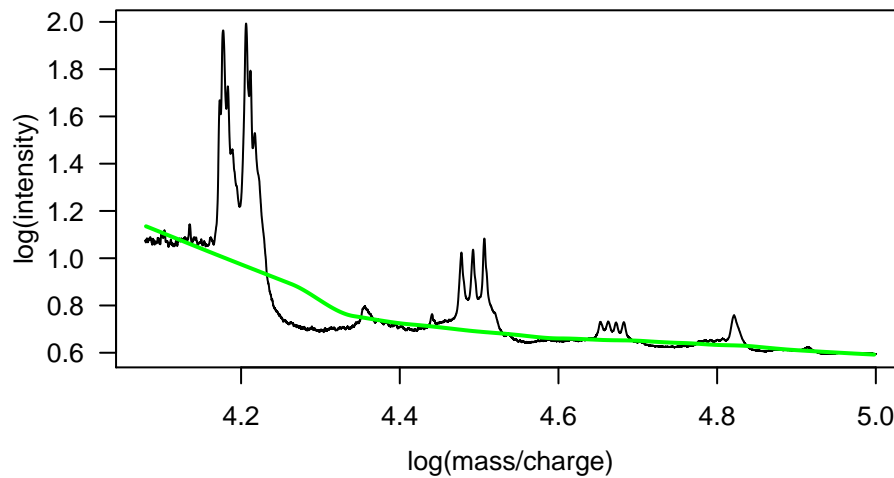


Figure 5.3.a.: Example From Mass Spectroscopy: Baseline is estimated by the LOWESS procedure with $f = 0.35$.

Usually 3 to 5 iterations are enough.

The resulting values $\hat{h}\langle z_i \rangle$ estimate the function h at the grid points z_1, \dots, z_N .

5.3 Baseline Removal

Example a From Mass Spectroscopy. Let us come back to the Mass Spectroscopy example in 5.1.a and apply the LOWESS procedure. We choose a high value of f to obtain a reasonable smooth curve. We hope so to catch the baseline in this spectrum. But as we can see in Figure 5.3.a, we miserably fail. Although we tried different values for f , we could not find any baseline estimation which was of any use. – This approach is much too naive.

b Modify LOWESS. Let us switch the view to

- The baseline is contaminated by the target signal.
- The contamination is one-sided.

Hence, we use an asymmetric robustness weight function in

$$\hat{\beta}(z_1) = \arg \min_{\underline{\beta}} \sum_{i=1}^n w_r(t_i) K\left(\frac{t_i - z_1}{b_w}\right) \cdot [y_i - \{\beta_0 + \beta_1 (t_i - z_1)\}]^2$$

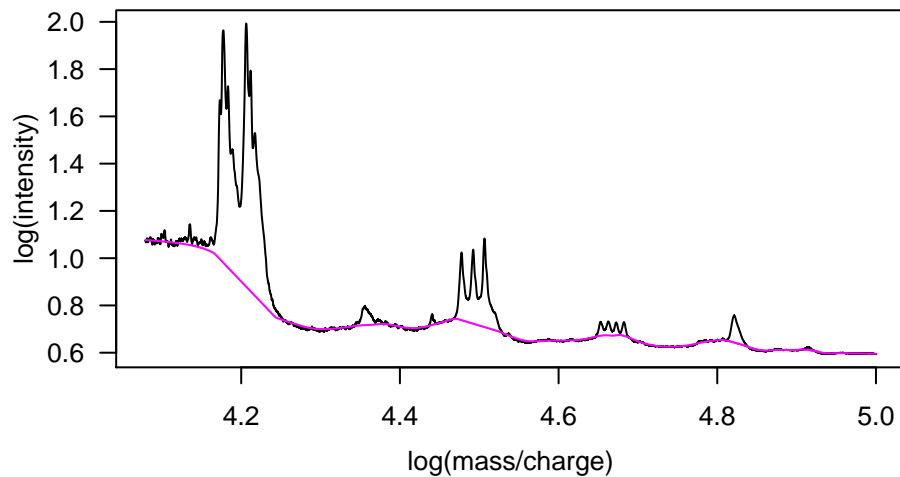


Figure 5.3.c.: Example From Mass Spectroscopy: The baseline is estimated by `rfbaseline()` with $d = 0.35$.

as, e.g.,

$$w_r(x_i) = \begin{cases} 1 & \text{if } r_i < 0 \\ [\max\{1 - (r_i/b)^2, 0\}]^2 & \text{otherwise,} \end{cases}$$

Again, we have to fix the tuning constant b and the bandwidth b_w . Our experience shows that

- a good choice for b is 3.5 (or any value between 3 and 4).
- the bandwidth b_w should be at least $2 \times$ the longest period in which the baseline is contaminated by the target signal.
- σ is estimated from the negative residuals.

This procedure is implemented in `rfbaseline()` of the R package `IDPmisc`. More details can be found in Ruckstuhl, Henne, Reimann, Steinbacher, Vollmer, O'Doherty, Buchmann and Hueglin (2012).

Example c From Mass Spectroscopy. In Figure 5.3.c, the introduced baseline removal procedure is applied the mass spectroscopy data of 5.1.a:

```
> library(IDPmisc)
> MS1.rfb4 <- rfbaseline(x=MS1$MZ, y=MS1$I, NoXP=1400, maxit=c(5,0),
                        DOT=TRUE, Scale=rfbaselineScale)
```

* `NoXP` is one way of specifying the amount of smoothing; `NoXP` is the number of `x` points used to compute each fitted value; it must be larger than 3. If `DOT` is `TRUE` outliers are disregarded totally; that is, observations with weight 0 are disregarded even when the neighbourhood is determined.

Contrary to the result shown in 5.3.a, we obtain a result which fits our ideas of a suitable solution. However, we cannot prove whether this solution is correct in any sense.

6 Some Closing Comments

6.1 General

- a Applications of Robustness Concepts to Other Models.** In this block, two robustness measures were introduced, namely the influence function and breakdown point, and some robust estimation methods for regression models (including the location model), generalized linear models (GLM) and covariance matrices were proposed and illustrated. There are also robust estimators for other models, e.g. for nonlinear regression models, in survival analysis or in time series (in particular ARIMA models).

Also, one of the most famous smoothers, LOWESS, and its extension LOESS, are built on a robust regression estimation.

- b Regression Analysis.** The challenge of robustness in linear regression models are quite well understood. Robust fitting and inference procedures are worked out and implemented, at least in R. But a subsequent model check is still indispensable. If there are outliers, you still must decide what to do with them. Preferably, you run another analysis on the outliers to find out what caused them or whether there are some regular patterns of occurrence in time or space.

There are also robust versions of regression estimators that have better predictive accuracy than ordinary least squares (OLS) in situations where the predictors are highly correlated and/or where the number of predictors p is large compared to the number of cases n , and even when $p > n$. These estimators are known as regularized regression estimators. Well-known estimators of this type are ridge regression, LASSO, and a combination of these two called the elastic net estimator. Robust versions of these regularized estimators are implemented in the function `pense` of package `Robot`.

- c General Thoughts.** The type of robustness that we have looked at in this block concerns itself with deviations from the probability distribution proposed in the model and must therefore be considered as distributional robustness. In general, robustness is concerned with “small” deviations from a specified (probability) model. The assumptions made for many models often include that the observations must be i.i.d. Therefore, we should investigate an inferential procedure (estimation, statistical test, confidence interval) for robustness (or stability) with respect to deviations from independence (i.), from the equality of the distributions (i.), and from the distribution (d.). The article from Stahel (1991) gives a good overview of what had already been explored in various areas up to the end of the 80s. A more recent overview is not known to the author.

6.2 Statistics Programs

- a R.** For all robust procedures that we have introduced in this block, there is at least one corresponding function in the open source statistics package R.


Linear Regression Models	<code>lmrob(...)</code> in the package <code>robustbase</code> <code>plot(lmrob object)</code>  residual analysis
GLM	<code>glmrob(...)</code> in the package <code>robustbase</code>
Model Comparision	<code>anova(lmrob - or glmrob object)</code> in the package <code>robustbase</code>
Covariance Matrices	<code>CovRobust(...)</code> in the package <code>rrcov</code>
Linear Discriminant Analysis	<code>rlda(...)</code> (own contribution)
Scatterplot Smoothing	<code>scatter.smooth(...)</code> or <code>lowess(...)</code>
Baseline Removal	<code>rfbaseline(...)</code> in the package <code>IDPmisc</code>
<i>and many more</i>	

Table 6.2.b.: Recommended R function for robust fitting.

- The package `robustbase` is specifically compiled for robust methods. The regression MM-estimator and improvements thereof are implemented in `lmrob(...)`. Generalized M-estimator for generalized linear models (GLM) can be applied with `glmrob(...)`. To run a regression M-estimator, you can use `glmrob(...)`. The R-function `anova(...)` can compare robustly fitted models. The robust MCD-covariance estimates in implemented in `covMcd(...)`. For more details, see the corresponding help files.
- The package `rrcov` contains a set of algorithms for computing robust multivariate location and scatter estimators, various robust methods for principal component analysis as well as for robust linear and quadratic discriminant analysis. The implementation is based on the S4 class system of the R programming environment.
- The R package `RobStatTM` was developed specifically for the book by Maronna et al. (2019) and contains all the robust estimators described in the book that are not yet included in `robustbase` or `rrcov`. Among others, there is in it the function `step.lmrobdetMM` to perform a robust variable selection by a robustified version of the final prediction error criterion and the R function `pense` that computes an MM version of the elastic net estimator. As a special case of the latter, it can compute the MM ridge and the MM lasso.
- In the R Package `MASS` there is the R function `rlm`. With `rlm(..., methods="M")`, the linear regression model is fitted by a regression M-estimator, whereas the function `rlm(..., methods="MM")` fits the model by a regression MM-estimator. However, it may fail if there are factor variables with many levels or very unbalanced factor variables. In the same package the R function `cov.rob` is found, with which you can carry out various robust estimations of the covariance matrices.

b Robust methods are essential in the daily business of statistical data analysis. I recommend to use R and the robust methods therein, given in Table 6.2.b.

- c **General Purpose Statistics Software.** Most of the widely used general purpose statistics packages have implemented some robust procedures and an interface to R, nowadays.

6.3 Literature

- a **Literature for Further Study.** There are two highly recommended new books on robust statistics, one from Maronna et al. (2019) and another from Heritier et al. (2009). The latter focuses on applications in Biostatistics. The two classic books on robust statistics are Huber and Ronchetti (2009) (it is the second edition of Huber (1981)) and Hampel et al. (1986). In the latter, the introductory chapter is also generally accessible. Hoaglin, Mosteller and Tukey (1983) as well as Rousseeuw and Leroy (1987) give an application oriented introduction to robust procedures, but don't contain the newer developments.

The foundations of robustness can be found in the famous articles from Huber (1964) and Hampel (1974).

More about rejection rules can be found in, for example, Barnett and Lewis (1978). But often rejection rules do not deliver what they promise (see Hampel, 1985).

A Appendix

A.1 Some Thoughts on the Location Model

- a What do we actually want?** What could we really want with a sample (or with two joined samples)?
- To estimate natural constants with measurements with random error.
 - Estimate changes for two linked samples with consideration for random fluctuations: Usually a "typical" change is interesting, e.g. as mentioned in 1.2.c in the example of sleep prolongation by medications.
 - Estimate expected value of unknown distribution (Taxes for budget planning purposes).

- b The location model is fundamental to these.** To achieve these we explicitly or implicitly (in some procedures this is not always said explicitly) fit the following location model:

$$X_i = \beta + E_i, \quad i = 1, \dots, n,$$

The X_i are the individual observations. The random deviations E_i , $i = 1, \dots, n$, are independent and all have the same distribution functions \mathcal{F} . The parameter β is the unknown location parameter that we want to estimate.

- c Consistent Estimations for the Expected Value.** The arithmetic mean \bar{X} is the only consistent estimator for the expected value for an unknown distribution F . Therefore \bar{X} is suitable for the "tax" estimation. However, the expected value is often a bad characterization for skewed distributions. (For budget planning, however, the distribution of tax revenue usually doesn't play a roll, since we are only interested in the total revenue!)
- d Nonparametric Statistics.** We are familiar with nonparametric statistics. There, the fewest possible assumptions are made about the error distribution. However, independence is still always assumed and sometimes so is symmetry of the measurement error.

The U-Test (Rank Sum Test from Wilcoxon, Mann and Whitney) is good for all distributions. It is therefore not surprising that the U-Test also has good robustness properties (like many nonparametric methods).

* An asymptotic result shows that the relative efficiency of the U-Test with respect to the t test is larger than 0.86. For the normal distribution this lower value is assumed. For small deviations from the normal distribution the relative efficiency becomes larger than one very quickly.

Limits:

- The U-Test is not applicable for small samples because it can't be significant.
- Often approximate prior knowledge about the distribution is available. This is

thrown away by nonparametric statistics.

- The nonparametric procedures can be generalized to other models only partially.

Thus, parametric approaches are definitely justified.

- e Error is normally distributed.** Parametric approaches to the error distribution are very useful, since they describe the data with a few numbers (parameters) (together with the model, naturally).

Experience shows that the normal distribution (with expected value 0) can often be assumed for the error distribution \mathcal{F} . Under an exact normal distribution, \bar{X} is the optimal maximum likelihood estimation.

But, real data is never *exactly* normally distributed. There are gross errors, or the “real” distributions are often longer tailed than the normal distribution. Experience shows that, e.g., for large data sets a t distribution with 3 to 9 degrees of freedom is more appropriate than the normal distribution. Why should small data sets behave otherwise? *However*, the t distribution is *more appropriate* but not the *only correct* distribution for describing the data.

* The arithmetic mean is also the solution that we get via application of the method of least squares. Then the optimality is also given by the Gauss-Markov theorem: “The estimations that are obtained via the least squares method are the most precise unbiased estimations from all the estimations that are *linear* functions of the measurements.” This optimality theorem does not require normal distribution of the measurements. However, without the assumption of linearity, the least squares estimator is only optimal under strong restrictions on the error distribution. The normal distribution fulfills these restrictions. Practically, this means that the requirement that the estimation be a linear function of the measurements gives approximately the same restrictions as the assumption that the measurement error is normally distributed.

- f Gross Error Model.** For any explicit distribution family we know the (asymptotically) optimal estimation (maximum likelihood).

But we never exactly know the distribution family. Therefore, instead of optimality for **one** distribution family, we should require good behavior (in the sense of almost optimal) for a set of distribution families. So, for example, instead of the normal distribution \mathcal{N} we take all distributions of the form

$$(1 - \varepsilon)\mathcal{N} + \varepsilon\mathcal{H},$$

where \mathcal{H} is an arbitrary distribution and ε is a fixed percentage (often between 0.01 and 0.2). This contaminated normal distribution is also known by the name gross error model. However, the goal remains to estimate the location parameter μ of the normal distribution.

- g** * **Literature for Further Study.** More discussion about the pros and cons of robust methods can be found in Hampel et al. (1986, Kap. 1 und 8.2).

A.2 Calculation of Regression M-Estimations

- a** Two solution procedures are presented here for solving the system of equations that defines the regression M-estimation in 2.1.b for $\underline{\beta}$.

b Iteratively Reweighted Least Squares. A first proposition is based on the fact that the system of equations in 2.1.b can also be written as weighted least squares, where the weights depend on the data.

1. Let $\underline{\beta}^{(m)}$ and $\sigma^{(m)}$ be the m th solution attempt of the system of equations in 2.1.b.
2. Then the residuals $r_i^{(m)} := y_i - \sum_{j=1}^p x_{ij}\beta_j^{(m)}$ and the weights

$$w_i = w\langle r_i, \underline{\beta}^{(m)} \rangle = \frac{\psi_c\langle r_i^{(m)} / \sigma^{(m)} \rangle}{r_i^{(m)} / \sigma^{(m)}}$$

are calculated.

3. The least squares problem

$$\sum_{i=1}^n w_i r_i \langle \underline{\beta} \rangle x_{ik} = 0, \quad k = 1, \dots, p,$$

is solved and we get a better approximate solution $\underline{\beta}^{(m+1)}$ of the system of equations in 2.1.b.

4. A robust scale estimation like, e.g. the standardized MAV (see 2.1.f), for the residuals $r_i^{(m+1)} := y_i - \sum_{j=1}^p x_{ij}\beta_j^{(m+1)}$ leads to an improved value $\sigma^{(m+1)}$.
5. Steps 2 through 4 are repeated until the solutions converge.

The converged values are a solution to the system of equations in 2.1.b and thus correspond to the regression M-estimations $\hat{\underline{\beta}}$ and $\hat{\sigma}$.

c Iterative Least Squares Estimator with Modified Observations. In the second proposition the residuals are modified (see Figure A.2.c). As with the first, this is also an iterative process. The procedure is similar to the first proposition, except that step 3 is replaced by the following step:

- 3'. Where the least squares problem

$$\sum_{i=1}^n (y_i^{(m)} - \sum_{j=1}^p x_{ij}\beta_j)^2 \stackrel{!}{=} \min_{\underline{\beta}}$$

is solved with the pseudo-observations

$$y_i^{(m)} := \sum_{j=1}^p x_{ij}\beta_j^{(m)} + \psi\langle r_i^{(m)} / \sigma^{(m)} \rangle \sigma^{(m)}$$

giving us a better approximate solution $\underline{\beta}^{(m+1)}$ of the system of equations in 2.1.b.

d Notes.

- Both approaches require start values. A possibility is to start out with the least squares solution. Another possibility is presented in section 2.4.
- The second approach clearly shows that with a regression M-estimator the influence of the residuals is bounded (**B**ounding the **i**nfluence of residuals: **BIR**).
- The convergence properties of these two algorithms can be referred to in Huber (1981).

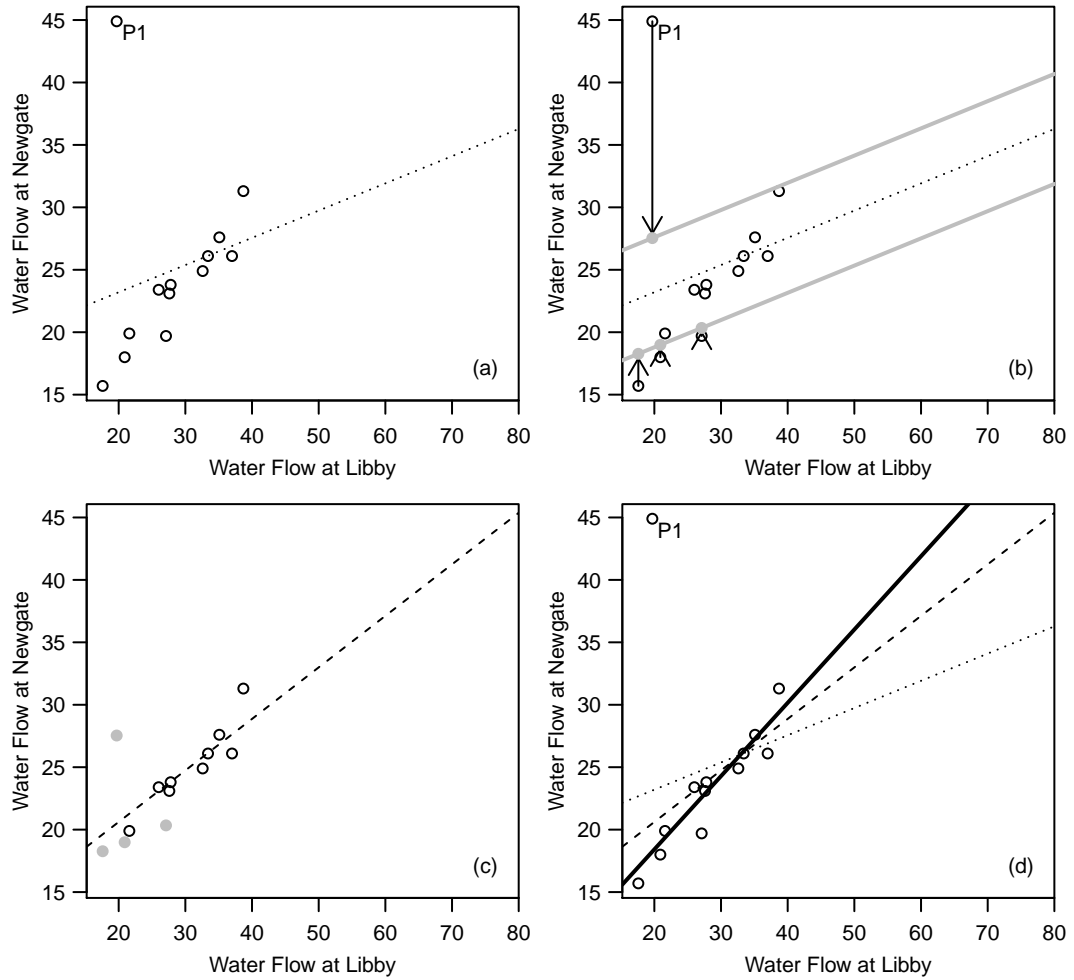


Figure A.2.c.: Schematic representation of the calculations of a regression M-estimator on the modified water flow example: (a) Beginning with the least squares solution as a starting value. After that, the observations are modified to pseudo-observations (b). With these pseudo-observations the least squares solution is again determined. In the last figure (d) are shown the 0. (---), the 1. (-----) and the last (—) iterations.

A.3 More Regression Estimators with High Breakdown Points

- a LMS Estimator.** From the location problem we know that there exists an estimator with breakdown point $\varepsilon^* = 1/2$ (e.g. the median). To achieve this in regression, we have to change our strategies. Until now, for

$$\sum_{i=1}^n (r_i(\underline{\beta}))^2 \stackrel{!}{=} \min_{\underline{\beta}}$$

we have tried to replace the quadratic function with something appropriate. Now we replace the summation. One proposition is to use the median:

$$\text{med}_i \{ (r_i(\underline{\beta}))^2 \} \stackrel{!}{=} \min_{\underline{\beta}}.$$

This estimator is called the **LMS estimator** (least median of squares estimator).

The LMS estimator is very resistant to deviations from the normal distribution. It is also advantageous that it doesn't need a scale estimation.

However, a serious disadvantage is that it has poor efficiency if the model assumptions are correct (e.g. no outliers) and a low asymptotic convergence rate ($1/\sqrt[3]{n}$ instead of $1/\sqrt{n}$). Additionally, it reacts very sensitively to unusual occurrences in the central region of the data. Because of these disadvantages, today the LMS estimator should no longer be used in practice!

b * **The LMS Estimator is an S-Estimator.** The variable

$$\sqrt{\text{med}_i \langle (r_i(\underline{\beta}))^2 \rangle} / 0.6745$$

is a robust scale estimator and corresponds to the standardized MAV (2.1.f). Thus the LMS belongs to the family of S-estimators, which minimize a (robust) scale estimation $\hat{\sigma}(r_1(\underline{\beta}), \dots, r_n(\underline{\beta}))$:

$$\hat{\underline{\beta}} := \arg \min_{\underline{\beta}} \hat{\sigma}(r_1(\underline{\beta}), \dots, r_n(\underline{\beta})).$$

For an appropriate choice of scale estimation $\hat{\sigma}(r_1(\underline{\beta}), \dots, r_n(\underline{\beta}))$ the corresponding S-estimator has a breakdown point of $\varepsilon^* = 1/2$. As with most S-estimators, the calculation of the LMS is very expensive. More about S-estimators and especially the LMS estimator can be found in Rousseeuw and Leroy (1987).

c **LTS Estimator.** As a direct alternative to the LMS, Rousseeuw has proposed the LTS estimator (least trimmed squares estimator):

$$\hat{\underline{\beta}} := \arg \min_{\underline{\beta}} \sum_{i=1}^q (r_{(i)}(\underline{\beta}))^2.$$

The summation is over the smallest $q = \lfloor (n + p + 1)/2 \rfloor$ squared residuals. This estimator has a better asymptotic convergence rate, but is still exactly as robust as the LMS. The efficiency remains low, comparable to the LMS or other robust S-estimators.

Bibliography

- Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data*, John Wiley & Sons, New York.
- Cushny, A. R. and Peebles, A. R. (1905). The action of optical isomers, *J. Physiology* **32**: 501–510.
- Dieke, G. H. and Tomkins, F. S. (1951). The $3p\ ^3\sigma \rightarrow 2s\ ^3\sigma$ -bands of TH and T_2 ., *Physical Review* **82**(6): 796–807.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**: 383–393.
- Hampel, F. R. (1985). The breakdown points of the mean combined with some rejection rules, *Technometrics* **27**(2): 95–107.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, New York.
- Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York.
- Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (eds) (1983). *Understanding Robust and Exploratory Data Analysis.*, John Wiley & Sons, New York.
- Huber, P. J. (1964). Robust estimation of a location parameter, *The Annals of Mathematical Statistics* **36**: 1753–1758.
- Huber, P. J. (1981). *Robust Statistics*, John Wiley & Sons, New York.
- Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, 2nd edn, John Wiley & Sons.
- Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, Prentice-Hall International, Inc.
- Koller, M. and Stahel, W. A. (2011). Sharpening wald-type inference in robust regression for small samples, *Computational Statistics & Data Analysis* **55**(8): 2504–2515.
- Koller, M. and Stahel, W. A. (2017). Nonsingular subsampling for regression S estimators with categorical predictors, *Computational Statistics* **32**(2): 631–646.
- Maronna, R. A., Martin, R. D., Yohai, V. J. and Salbian-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*, Wiley Series in Probability and Statistics, 2nd edn, John Wiley & Sons.

- Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M. and Maechler, M. (2011). *robustbase: Basic Robust Statistics*. R package version 0.7-0.
- Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression & Outlier Detection.*, John Wiley & Sons, New York.
- Ruckstuhl, A. F. (1997). Partial breakdown in two-factor models, *Journal of Statistical Planning and Inference* **57**: 257–271. Special Issue on Robust Statistics and Data Analysis, Part II.
- Ruckstuhl, A. F., Henne, S., Reimann, S., Steinbacher, . M., Vollmer, M. K., O'Doherty, S., Buchmann, B. and Hueglin, C. (2012). Robust extraction of baseline signal of atmospheric trace species using local regression, *Atmospheric Measurement Techniques* **5**(11): 2613–2624.
URL: <http://www.atmos-meas-tech.net/5/2613/2012/>
- Ruckstuhl, A. F., Stahel, W. A. and Dressler, K. (1993). Robust estimation of term values in high-resolution spectroscopy: Application to the $e^3\Sigma_u^+ \rightarrow a^3\Sigma_g^+$ spectrum of T_2 , *Journal of Molecular Spectroscopy* **160**: 434–445.
- Ruckstuhl, A. and Meier, P. (2009). Style exposure and leverage of funds of hedge funds with a factor model. Results of a project supported by CTI, the Swiss innovation promotion agency, as well as Complementa AG.
- Stahel, W. (2000). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 3. Auflage edn, Vieweg, Braunschweig/Wiesbaden.
- Stahel, W. A. (1991). Research directions in robust statistics, in W. Stahel and S. Weisberg (eds), *Directions in Robust Statistics and Diagnostics (Part II)*, Vol. 34 of *The IMA Volumes in Mathematics and Its Applications*, Springer-Verlag, New York, pp. 243–278.
- Stahel, W. A. (2007). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 5. Auflage edn, Vieweg+Teubner Verlag.
- Venables, W. N. and Ripley, B. (1999). *Modern Applied Statistics with S-Plus*, Statistics and Computing, third edn, Springer-Verlag, New York.
- Yohai, V., Stahel, W. A. and Zamar, R. H. (1991). A procedure for robust estimation and inference in linear regression, in W. Stahel and S. Weisberg (eds), *Directions in Robust Statistics and Diagnostics (Part II)*, Vol. 34 of *The IMA Volumes in Mathematics and Its Applications*, Springer-Verlag, New York, pp. 365–374.