ETH
Swiss Federal Institute of Technology Zurich

Seminar for
Statistics

Sharpening Wald-type Inference in Robust Regression for Small Samples

Manuel Koller*, Werner A. Stahel†

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

February 2011

Abstract

The datasets used in statistical analyses are often small in the sense that the number of observations n is less than 5 times the number of parameters p to be estimated. In contrast, methods of robust regression are usually optimized in terms of asymptotics with an emphasis on efficiency and maximal bias of estimated coefficients. Inference, i.e., determination of confidence and prediction intervals, is proposed as complementary criteria. An analysis of MM-estimators leads to the development of a new scale estimate, the *Design Adaptive Scale Estimate*, and to an extension of the MM-estimate, the *SMDM-estimate*, as well as a suitable ψ-function. A simulation study shows and a real data example illustrates that the SMDM-estimate has better performance for small n/p and that the use of the new scale estimate and of a slowly redescending ψ-function is crucial for adequate inference.

Keywords: MM estimator, Robust Regression, Robust Inference

*koller@stat.math.ethz.ch
†stahel@stat.math.ethz.ch

1 Introduction

The literature provides many proposals for robust linear regression. In this paper, we focus on variations of MM-estimators, which consist of an initial S-estimate followed by an M-estimate of regression. They have the benefit of allowing for the specification of a breakdown point as well as asymptotic efficiency at the normal distribution. Some properties of such estimators depend critically on the choice of the ψ-function of the M-step.

The papers we found on the topic of choosing ψ for MM-estimators focus on properties of the estimated parameters only. The criteria applied are the maximum asymptotic bias or a similar, simplified version like the contamination sensitivity. Here, we take the whole statistical analysis into account. This means to ask, apart from high efficiency, for accurate inference – determination of confidence and prediction intervals – for a realistically small sample size. In this paper, we consider only Wald-type inference. Other approaches such as robustified likelihood-ratio tests or saddle-point tests would also be worth considering, but require more work for anything other than tests for hypotheses on single parameters.

Accurate estimation of the scale proves essential in achieving this goal. It is known that the scale determined by an S-estimator suffers bias if there are a sizable number of parameters p compared to the number of observations n, a situation often encountered in everyday statistical analyses. Maronna and Yohai (2010) suggested an empirical correction to the scale. We propose a novel scale estimate based on the MM-estimate's residuals. We call it *Design Adaptive Scale Estimate* or *D-scale* for short. Based on this scale we propose to reestimate the coefficients by an additional M-estimation step, thereby achieving a good match between estimated scale and coefficients. Since this procedure combines an S-estimation, an M-step, a D-estimation of scale and another M-step, we call it *SMDM-estimator*.

There are a number of papers discussing the choice of ψ-functions for MM-estimators. Martin et al. (1989) derive the maximum asymptotic bias of the S-estimate over an ϵ-contamination neighborhood of the central model as well as lower bounds for MM-estimates without an intercept. In Hennig (1995) various types of ψ-functions are discussed and an upper bound for the maximum asymptotic bias for MM-estimates is presented. The latter is improved by Berrendero et al. (2007). Their maximum asymptotic bias bounds coincide up to moderately large ϵ (about 0.3). The so-called *optimal* ψ-function is developed in Yohai and Zamar (1997) based on maximizing

efficiency with a bound on contamination sensitivity. Svarc et al. (2002) show that the *optimal* ψ-function is almost identical to the optimal ψ-function based on efficiency with a bound on the maximum asymptotic bias, at least for moderate contamination $\epsilon \leq 0.2$.

We will show that minimizing the maximum asymptotic bias, which entails a quickly redescending ψ-function, leads to poor inference properties when p/n is *not very small*. As our simulations show, a remedy for this problem is to use slowly redescending ψ-functions, corresponding to an early insight of Frank Hampel, who set the maximal rate of descent of the *Hampel* ψ-function to half the maximal ascent. For further reference see Hampel et al. (1986). Intuitively, this comes from a positive feedback loop: If observation i has a positive residual with $\psi'(r_i/\hat{\sigma}) < 0$ and if the parameters are changed to slightly decrease the fitted value $\hat{y}_i$, then the residual r_i increases and the influence of the observation on $\hat{y}_i$ decreases, which in turn decreases $\hat{y}_i$. Hampel's ψ-function has sharp corners. It turns out in simulations that this is a disadvantage; therefore, we introduce a new family of ψ-functions called *lqq*.

In Section 2, the MM-estimators are briefly reviewed. The Design Adaptive Scale Estimate and SMDM-estimates are defined in Sections 3 and 4. Wald-type inference is discussed in Section 5 and ψ-functions in Section 6. The simulation study is outlined and discussed in Section 7. A real data set is analyzed in Section 8 and Section 9 concludes the paper. Details to the calculation of the Design Adaptive Scale estimate are included as an Appendix.

2 MM-estimates

Our notation for the linear regression model is

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n,$$

where the e_i are i.i.d. and independent of $\boldsymbol{x}_i$ with $e_i \sim \mathcal{N}(0, \sigma^2)$ under the *central model*. Here we adopt the attitude that we want to fit this model as well as possible in the presence of contamination. We denote the residuals as $r_i = r_i(\hat{\boldsymbol{\beta}}) = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$.

M-estimates of regression are defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right), \tag{1}$$

where $\rho(r)$ is assumed to be a nondecreasing function of $|r|$, with $\rho(0) = 0$ and strictly increasing for $r > 0$ where $\rho(r) < \rho(\infty)$. Maronna et al. (2006) restrict the term *ρ-functions* to this type of functions. If ρ is bounded, it is assumed that $\rho(\infty) = 1$ and the estimate defined by (1) is then called *redescending* M-estimate of regression. The scale σ is required to gain scale equivariance and can either be an external scale estimate or estimated simultaneously. By differentiating (1) we get the estimating equation

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\sigma}\right) \boldsymbol{x}_i = 0,$$

where ψ is proportional to ρ' and is usually chosen to have $\psi'(0) = 1$.

An *M-estimate of scale* of $\boldsymbol{e} = (e_1, \ldots, e_n)$ is the solution $\hat{\sigma}$ to the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \chi\left(\frac{e_i}{\sigma}\right) = \kappa,$$

where κ is a tuning constant and $\chi(e)$ fulfills the same properties as ρ does.

S-estimates of regression are the parameter values $\hat{\boldsymbol{\beta}}_{\mathrm{S}}$ that minimize the M-estimate of scale $\hat{\sigma}_{\mathrm{S}} = \hat{\sigma}_{\mathrm{S}}(\boldsymbol{r}(\hat{\boldsymbol{\beta}}_{\mathrm{S}}))$ of the associated residuals,

$$\hat{\boldsymbol{\beta}}_{\mathrm{S}} = \arg\min_{\boldsymbol{\beta}} \hat{\sigma}_{\mathrm{S}}(\boldsymbol{r}(\boldsymbol{\beta})).$$

The maximal breakdown point $(1-p/n)/2$ of the S-estimate is attained when using $\kappa = (1 - p/n)/2$. See Maronna et al. (2006) for details.

It is impossible for S-estimates to achieve a high breakdown point as well as a high efficiency. Following the proposal of Yohai (1987), arbitrarily high efficiency is possible by using *MM-estimates*. They are defined as a local minimum of (1), obtained by using an iterative procedure started at an initial S-estimate $\hat{\boldsymbol{\beta}}_{\mathrm{S}}$. The corresponding $\hat{\sigma}_{\mathrm{S}}$ is used as scaling factor in (1). For a suitable choice of ρ in (1), the MM-estimate preserves the breakdown point of $\hat{\beta}_S$. The functions ρ, χ are usually taken from the same family. The tuning constant for ρ is determined such that the estimator reaches a desired value for the asymptotic efficiency.

It will be shown that the standard MM-estimate has three main problems for designs with a high p/n ratio:

- bias in the S-scale estimate,

- loss of efficiency of the estimated parameters,
- failure to keep the levels of tests at the desired value.

The first one is solved by using the *Design Adaptive Scale Estimate* while the two latter ones depend on how quickly the ψ-function redescends.

3 Design Adaptive Scale Estimate

It is well known that for linear regression using the least squares estimate, the residuals are correlated and heteroskedastically distributed. Therefore, the maximum likelihood estimate of the variance parameter of the errors is corrected to render it unbiased. This is done by dividing the sum of the squared residuals by $n - p$ instead of n. In the case of MM-estimates this is also an issue. To the knowledge of the authors, all proposed corrections depend only on n and p, but not on the design itself. Due to the nonlinear nature of the M-estimate, such a correction will not suffice and further correction factors are required at later stages of the analysis. Huber (1973) developed elaborate small sample correction factors for the covariance matrix estimate. It will be shown that appropriate standardization of the residuals renders these corrections obsolete.

We propose to estimate scale by the estimating equation

$$\sum_{i=1}^{n} \tau_i^2 w\left(\frac{r_i}{\tau_i \sigma_{\mathrm{D}}}\right)\left[\left(\frac{r_i}{\tau_i \sigma_{\mathrm{D}}}\right)^2 - \kappa\right] = 0, \tag{2}$$

where w is a weighting function, τ_i will be defined below, and κ is used to ensure Fisher consistency at the central model. We call this scale estimate the *Design Adaptive Scale Estimate*. In the case of ordinary least squares, $w(r) = 1$, $\tau_i = \sqrt{1 - h_i}$ and $\kappa = 1$, and $\hat{\sigma}_{\mathrm{D}}$ reduces to the standard scale estimate. In accordance with the M-step, an appropriate choice for the weighting function is $w(r) = \psi(r)/r$ (cf. Maronna et al., 2006, Section 2.2.3).

The correction factors τ_i are designed to reflect the heteroskedasticity of the distributions of the residuals r_i. They depend on the leverage h_i of the ith observation as well as the ψ-function used. For each observation, the τ_i is chosen such that the expected value of the ith summand in (2) is zero. The distribution of the residuals is approximated using a *von Mises expansion* of

$\hat{\beta}$,

$$r_i = y_i - \boldsymbol{x}_i' \left(\boldsymbol{\beta} + \frac{1}{n} \sum_{h=1}^{n} \mathrm{IF}(e_h, \boldsymbol{x}_h, \sigma) + \text{remainder} \right). \tag{3}$$

Details on how to calculate τ_i are given in the Appendix. By simulation we show that the standardization by τ_i removes the bias in the scale almost completely for all ratios $p/n \leq 1/3$ at least for suitable ψ-functions.

An important detail is the inclusion of κ inside the sum in the given weighted form. This has the benefit that outlying observations have no influence on the scale estimate and thus are rejected. The drawback of this approach is that there might be multiple solutions to the estimating equation, for example when there is a large cluster of outlying observations. The problem can be solved with the same trick as for the MM-estimate itself, by choosing a suitable starting point for the calculation of the D-scale. This will allow keeping the breakdown properties equal to those of the prior estimates.

4 SMDM-Estimates

We propose to extend the standard MM-estimate with two additional steps to solve the problems outlined at the end of Section 2. Following the MM-estimation step, calculate the *Design Adaptive Scale Estimate*. Then reestimate the regression parameters based on this new scale using the MM-estimate as initial estimate. We will call this estimate the *SMDM-estimate*. Other candidates of estimates could be SDM or SMD, i.e., MM-estimates with an additional D-step either before or after the M-regression estimate. Both of them turned out to be inferior to the SMDM-estimate. The D-scale is only unbiased if the τ factors are estimated correctly. If applied following the S-step, estimating τ proved difficult. For a more detailed explanation, we refer to the Appendix. Consequently, the SDM estimate was not included in the simulations. As for the SMD, the simulation study revealed that it does not reach the desired efficiency. Since the S-scale used in the first M-step is biased for larger p/n ratios, the effective tuning of the ψ-function differs from the intended one and thus the efficiency of the regression estimate is lowered.

5 Wald-type inference

Under some regularity conditions, MM-estimates are asymptotically normal, thereby allowing for Wald-type tests and confidence intervals. The covariance matrix of the estimated parameters,

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \gamma \boldsymbol{V}_{\boldsymbol{X}}^{-1},$$

consists of three parts: a scale σ, a correction factor γ depending on the ψ-function used, and a matrix part, $\boldsymbol{V}_{\boldsymbol{X}} = \boldsymbol{X}^T \boldsymbol{X}$.

We will estimate the three parts as correctly as possible, separately from each other. That is, we do not want to rely on canceling effects or use any correction factors later on. The first part, σ, is taken care of by using the scale of SMDM-estimates. The D-scale estimate is suitable for inference and does not require any correction factors.

The correction factor γ is given by the asymptotic normality theorem and is usually estimated empirically by

$$\hat{\gamma} = \frac{\frac{1}{n-p} \sum_{i=1}^{n} \psi\left(\frac{r_i}{\hat{\sigma}}\right)^2}{\left[\frac{1}{n} \sum_{i=1}^{n} \psi'\left(\frac{r_i}{\hat{\sigma}}\right)\right]^2},$$

where the factor $1/(n-p)$ is used instead of $1/n$ in order to recapture the classical formula in the classical case ($\psi(x) = x$), see Huber and Ronchetti (2009), Section 7.6. This formula is usually used together with Huber's small sample correction, which will be discussed later. Here, we propose to use the τ-standardized residuals again,

$$\hat{\gamma} = \frac{\frac{1}{n} \sum_{i=1}^{n} \psi\left(\frac{r_i}{\tau_i \hat{\sigma}}\right)^2}{\left[\frac{1}{n} \sum_{i=1}^{n} \psi'\left(\frac{r_i}{\tau_i \hat{\sigma}}\right)\right]^2}. \tag{4}$$

For the third part, we follow the proposal of Yohai et al. (1991), namely to use a weighted empirical covariance matrix estimate. The robustness weights of the final M-step are used as weights,

$$\widehat{\boldsymbol{V}}_{\boldsymbol{X}} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} w_i} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}, \tag{5}$$

where $w_i = w(r_i/\hat{\sigma})$ and $\boldsymbol{W} = \operatorname{diag}(w_1, \ldots, w_n)$. This has expectation $\boldsymbol{X}^T \boldsymbol{X}$. The τ-standardized residuals are not used here since to estimate the covariance matrix shape, we want to give the observations the same weights as used in the estimation of $\hat{\boldsymbol{\beta}}$.

Under the assumptions of a symmetric error distribution, a symmetric ρ-function and a balanced matrix $\boldsymbol{X}$, i.e., all leverages equal to p/n, Huber (1973) showed that $\gamma\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ contains a bias of the order $O(p/n)$ (see also Huber and Ronchetti (2009)). He derived a correction factor K^2 that makes $\gamma\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ unbiased up to $O(p^2/n^2)$ terms,

$$K^2 = \left[1 + \frac{p}{n}\frac{\frac{1}{n}\sum_{i=1}^{n}\left(\psi'\left(\frac{r_i}{\hat{\sigma}}\right) - \frac{1}{n}\sum_{i=1}^{n}\psi'\left(\frac{r_i}{\hat{\sigma}}\right)\right)^2}{\left[\frac{1}{n}\sum_{i=1}^{n}\psi'\left(\frac{r_i}{\hat{\sigma}}\right)\right]^2}\right]^2.$$

This correction is only valid when taking $\boldsymbol{V}_{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{X}$. The variant of Equation (5) was also mentioned, but deemed to be too complicated to calculate a correction factor. Nevertheless, many implementations of MM-estimators use this correction in combination with the matrix part as in (5).

6 ψ-functions

We propose a new, fully tunable ψ-function which we call *lqq* ("linear quadratic quadratic").

$$\psi(x) = \begin{cases} x & |x| \le c \\ \operatorname{sign}(x)\left(|x| - \frac{s}{2b}\left(|x| - c\right)^2\right) & c < |x| \le b + c \\ \operatorname{sign}(x)\left(c + b - \frac{bs}{2} + \frac{s-1}{a}\left(\frac{1}{2}\tilde{x}^2 - a\tilde{x}\right)\right) & b + c < |x| \le a + b + c \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{x} = |x| - b - c$ and $a = (bs - 2b - 2c)/(1 - s)$. The parameter c determines the width of the central identity part. The sharpness of the bend is adjusted by b while the maximal rate of descent is controlled by s ($s = 1 - |\min_x \psi'(x)|$). The length a of the final descent to 0 is determined by b, c and s.

We constructed this function as follows. Our initial proposal was the Generalized Gauss-Weight function, or *ggw* for short,

$$\psi(x, a, b, c) = \begin{cases} x & |x| \le c \\ \exp\left(-\frac{1}{2}\frac{(|x| - c)^b}{a}\right)x & |x| > c, \end{cases}$$

which is continuously differentiable. This function was appealing to us because it is possible to fix the maximal rate of descent and has the property of

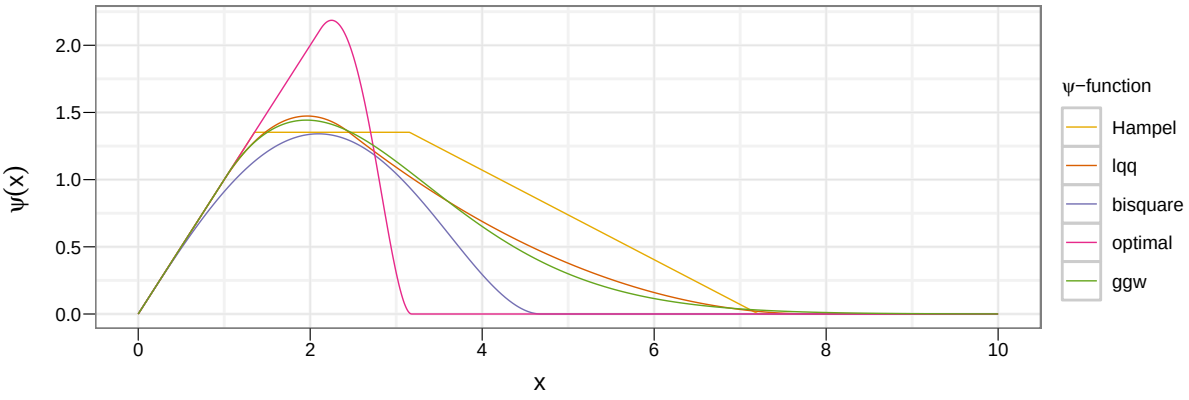

Figure 1: ψ-functions for MM-estimation with the same asymptotic efficiency.

reaching 0 only asymptotically. Moreover, since the function has a continuous derivative, the sensitivity curves of various statistics, e.g., t-values, show no jumps and less other irregularities as compared to the so-called *optimal* or *Hampel* ψ-functions.

The *ggw* ψ-function is only numerically integrable, however. This posed a problem in the implementation, because the computationally expensive S-step relies on many evaluations of the corresponding ρ-function. Even though shortcuts like pre-calculating the integral exist, it is more straightforward to simplify ψ. The *lqq* ψ-function is designed to mimic *ggw*. The function is constructed by integrating from ψ', which is continuous and consists of a constant and two linear parts. The continuity of ψ' is required to yield a continuous sensitivity curve of the t test statistic.

As mentioned before, the tuning constants will be chosen such that the decreasing slope for large residuals is less than the slope at 0. This is illustrated in Fig. 1. A list of recommended tuning constants for the *lqq* and *ggw* ψ-function is given in Table 1. The use of slowly redescending ψ-functions requires a price to be paid with the maximum asymptotic bias. The latter is considerably larger than for the *optimal* ψ-function with equal asymptotic efficiency. Fig. 2 shows the maximum asymptotic bias bounds, calculated as outlined in Berrendero et al. (2007).

		lqq			*ggw*		
	eff.	b	c	s	a	b	c
S-estimate		0.4015	0.2677	1.5000	0.2037	1.5000	0.2959
M-estimate	0.85	1.0582	0.7055	1.5000	0.8372	1.5000	0.7594
	0.9	1.2137	0.8092	1.5000	1.0283	1.5000	0.8709
	0.95	1.4735	0.9823	1.5000	1.3864	1.5000	1.0628

Table 1: Recommended tuning constants for *lqq* and *ggw* ψ-functions.

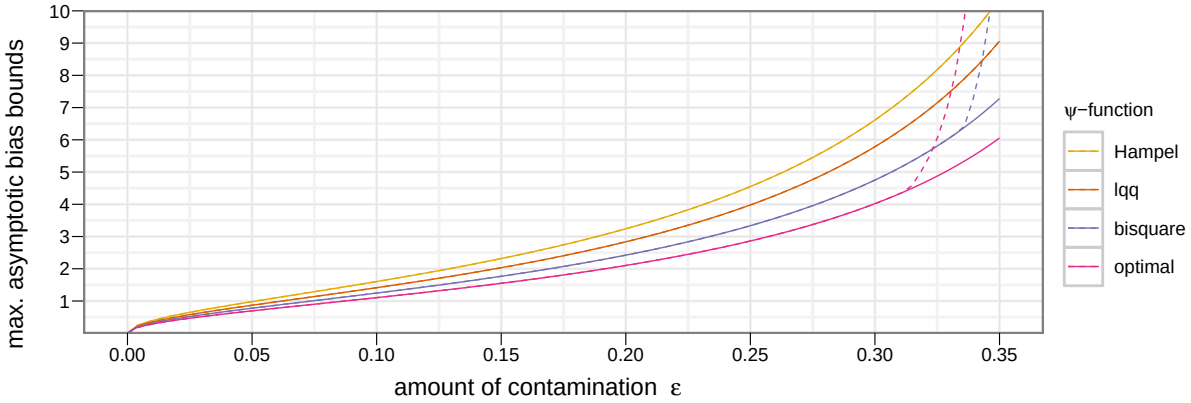

Figure 2: Maximum asymptotic bias bounds for the ψ-functions used in the simulation. The upper and lower bounds coincide for most of the plot.

7 Simulation Study

To compare the proposed methods with other robust regression procedures, we performed a simulation study. It was conducted with `R` version 2.11.1 (R Development Core Team, 2010). The methods proposed have been implemented in the `R`-package **`robustbase`** (Rousseeuw et al., 2011). Because of space constraints, we present here only a reduced set of results. The full simulation study is available as a vignette of the `R`-package **`robustbase`** (Koller, 2011).

7.1 Estimators

We compare the following estimators for $\hat{\beta}$ and $\hat{\sigma}$.

- MM, SMD, SMDM-estimates as implemented in the method `lmrob` of

the R-package `robustbase` (Rousseeuw et al., 2011).

- MM-estimate as implemented in the method `lmRob` of the package `robust` (Wang et al., 2010). We will denote it by *MMrobust* (and $\hat{\sigma}_{\text{robust}}$ for just the scale estimate) later on.

- MM-estimate using the S-scale correction q_E as proposed by Maronna and Yohai (2010). This method was implemented on the basis of `lmrob`. The scale estimate is multiplied by q_{E} which is defined as

$$q_{\text{E}} = \frac{1}{1 - (1.29 - 6.02/n)p/n}$$

  for *bisquare* ψ. When using q_{E} it is necessary to adjust the tuning constants of χ to account for the dependence of κ on p. We denote this estimator by $q_{\text{E}}\hat{\sigma}_S$ and MMq_{E}.

- ordinary least squares (OLS).

The covariance matrices are estimated as follows. For the standard MM-estimator, we compare Avar_1 of Croux et al. (2003) and the weighted empirical covariance matrix estimate corrected by Huber's *s*mall *s*ample *c*orrection (*Wssc* in the plot legend) as described in Huber and Ronchetti (2009). The latter is also used for MMq_{E}. For the SMD and SMDM variants we use the covariance matrix estimate as described in Section 5 (denoted by $W\tau$).

We compare the *bisquare*, *optimal*, *lqq*, and *Hampel* ψ-functions. They are illustrated in Fig.1. Note that the *Hampel* ψ-function is tuned to have a downward slope of $-1/3$ instead of the originally proposed $-1/2$. This was set to allow for a comparison to a more slowly redescending ψ-function.

7.2 Designs

The simulation setting used here is modeled on the one in Maronna and Yohai (2010). The designs used in the simulation are random designs without an intercept column. The distribution used to generate the errors is also used to generate the designs. We simulate $n = 25, 50$ and 100 with predictor–observation ratios of $p/n = 1/20, 1/10, 1/5, 1/3, 1/2$. We round p to the nearest smaller integer if necessary.

We simulate the following error distributions.

- Standard normal distribution,

- t_3,
- centered skewed t, as introduced by Fernández and Steel (1998) and implemented in the `R` package `skewt` (King and Anderson, 2004), with $\gamma = 2$, and $df = 5$ (denoted by $cskt(5, 2)$) and
- contaminated normal, $\mathcal{N}(0, 1)$ contaminated with 10% $\mathcal{N}(4, 1)$ (asymmetric, $cnorm(0.1, 4, 1)$).

We simulate 1000 replicates.

7.3 Criteria

The simulated methods are compared using the following criteria.

Scale estimates. The criteria for scale estimates are all calculated on the log-scale. The bias of the estimators is measured by the 10% trimmed mean. To recover a meaningful scale, the results are exponentiated before plotting. It is easy to see that this is equivalent to calculating geometric means. Since the methods are all tuned at the central model, $\mathcal{N}(0, 1)$, a meaningful comparison of biases can only be made for $\mathcal{N}(0, 1)$ distributed errors. The variability of the estimators, on the other hand, can be compared over all simulated error distributions. It is measured by the 10% trimmed standard deviation, rescaled by the square root of the number of observations.

Coefficients. The efficiency of estimated regression coefficients $\hat{\boldsymbol{\beta}}$ is characterized by their mean squared error (MSE). Since we simulate under $H_0 : \boldsymbol{\beta} = 0$, this is determined by the covariance matrix of $\hat{\boldsymbol{\beta}}$. We use $\mathbf{E}\left[\|\hat{\boldsymbol{\beta}}\|_2^2\right] = \sum_{j=1}^{p} \mathrm{var}(\hat{\beta}_j)$ as a summary. When comparing to the MSE of the ordinary least squares estimate (OLS), this gives the efficiency, which, by the choice of tuning constants of ψ, should yield

$$\frac{\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})}{\mathrm{MSE}(\hat{\boldsymbol{\beta}})} \approx 0.95$$

for standard normally distributed errors. The simulation mean of $\sum_{j=1}^{p} \mathrm{var}(\hat{\beta}_j)$ is calculated with 10% trimming. For other error distributions, this ratio should be larger than 1, since by using robust procedures we expect to gain efficiency at other error distributions (relative to the least squares estimate).

Covariance matrix estimate. The covariance matrix estimates are compared indirectly over the performance of the resulting test statistics. We

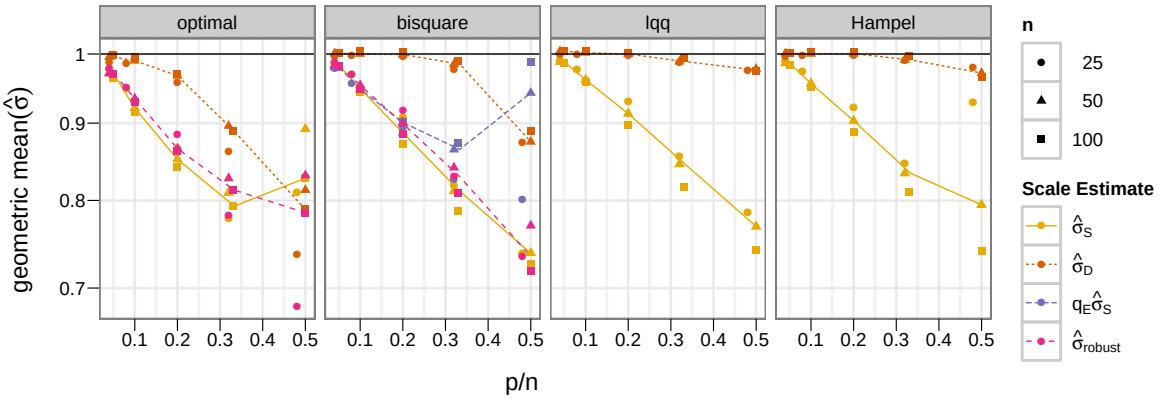

Figure 3: Geometric mean of scale estimates for normal errors and different ψ-functions. The mean is calculated with 10% trimming. The lines connect the median values for each simulated ratio p/n. $q_{\mathrm{E}}\hat{\sigma}_S$ applies for *bisquare* ψ only.

compare the empirical level of the hypothesis tests $H_0 : \beta_j = 0$ for some $j \in \{1, \ldots, p\}$. The power of the tests is compared by testing for $H_0 : \beta_j = b$ for several values of $b > 0$. The formal power of a more liberal test is generally higher. Therefore, in order for this comparison to be meaningful, the critical value for each test statistic was corrected such that all tests have the same simulated level of 5%.

7.4 Results

The results are presented as a series of plots. The results for the different ψ-functions are each plotted in a different facet, except for Fig. 5, where the facets distinguish the scale estimators, as well as Fig. 9 and 11, where the facets show the results for various error distributions. The plots are augmented with auxiliary lines to ease the comparison of the methods. The lines connect the median values over the values of n for each simulated ratio p/n. In many plots the y-axis has been truncated. Points in the gray shaded area represent truncated values using a different scale.

Scale. Fig. 3 shows a clear dependence of the bias of $\hat{\sigma}$ on p/n for the S-scale estimate $\hat{\sigma}_{\mathrm{S}}$. This bias seems to be independent of which ψ-function is used. Surprising is the similarity between the uncorrected $\hat{\sigma}_{\mathrm{S}}$ and the q_{E} corrected scale estimate. The performance of the D-scale estimate $\hat{\sigma}_{\mathrm{D}}$

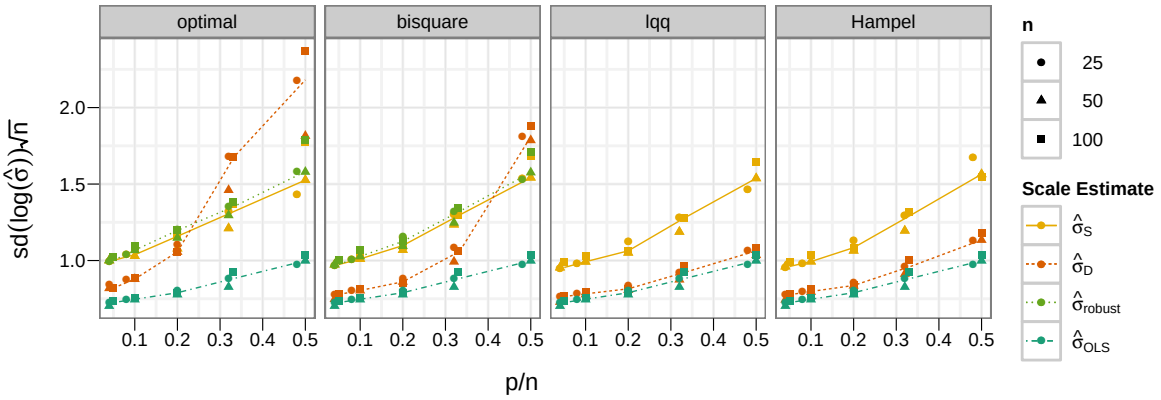

Figure 4: Variability of the scale estimates for normal errors and different ψ-functions. Standard deviations are calculated with 10% trimming.

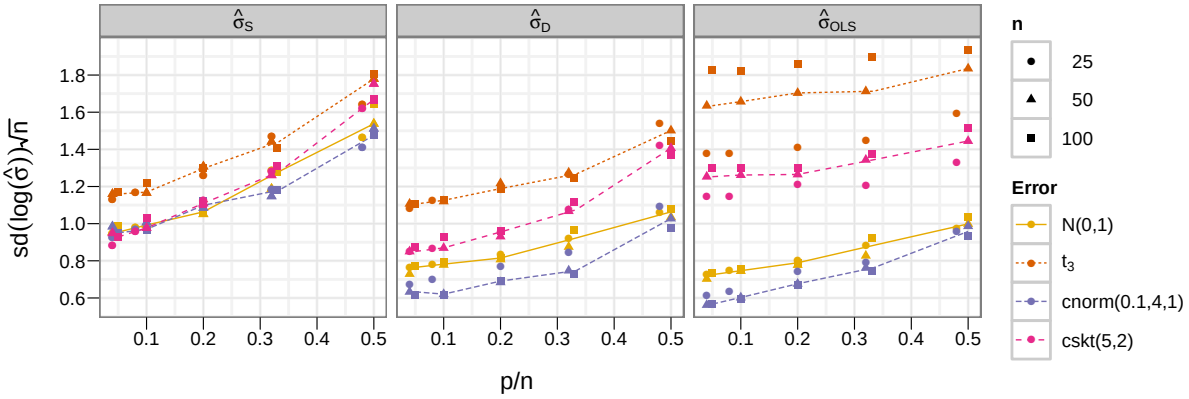

Figure 5: Variability of the D-scale, S-scale with *lqq* ψ-function, and for OLS-estimators.

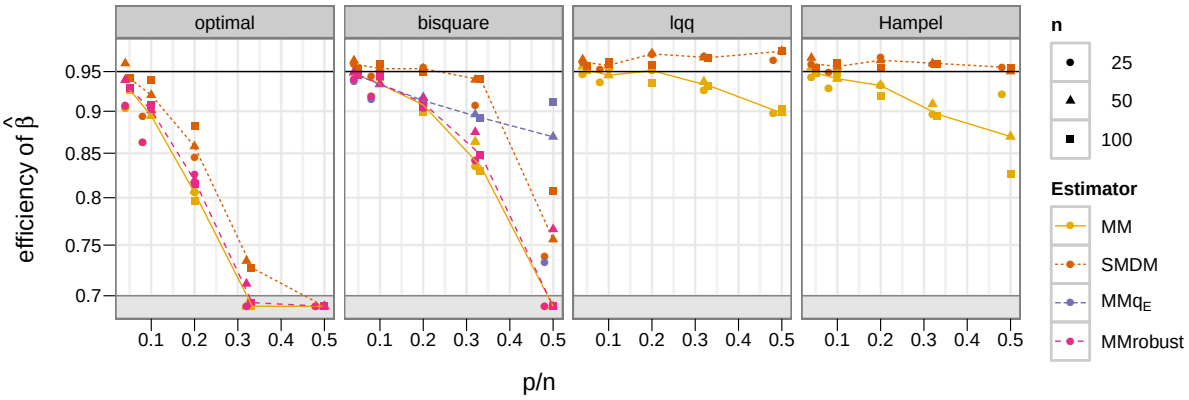

Figure 6: Efficiency of $\hat{\beta}$ with respect to OLS for normal errors and different ψ-functions. The vertical axis is truncated at 0.7. Extreme values for the vertical axis (shaded area) are shown on different scale.

depends on the ψ-function used. For the slowly descending ψ-functions, the bias is almost zero over the whole range of p/n. For these ψ-functions, the D-scale estimate is also more efficient at the central model than the S-scale estimate, as Fig. 4 shows. For the other simulated distributions, this is also the case (Fig. 5), except for t_1 distributed errors – a quite extreme situation not shown on plots – where the S-scale estimate is more efficient. All scale estimates suffer a loss of efficiency for larger p/n ratios. To keep the efficiency high for all values of p/n, the D-scale estimate requires a slowly descending ψ-function.

Coefficients. With increasing bias and a loss of efficiency of the scale estimate, there is also a loss of efficiency of the regression estimate $\hat{\boldsymbol{\beta}}$. Fig. 6 shows the results at the central model. For the *optimal* ψ-function, the loss is quite dramatic, independent on what estimator is used. The results for the MM and the SMD-estimates are the same since the D-scale estimate is not used for estimating $\hat{\boldsymbol{\beta}}$. The reestimation of the regression estimate by the last M-step of the SMDM-estimate improves the results considerably. Here the requirement of the slowly redescending ψ-functions is also clearly visible. While for the *optimal* and *bisquare* ψ-functions there is a considerable loss of efficiency, there is no visible loss for the *lqq* and *Hampel* ψ-functions.

In Fig. 7, empirical efficiencies of the estimates are plotted for different error distributions. The differences between the methods are marginal compared to the variance introduced by the different error distributions. In some

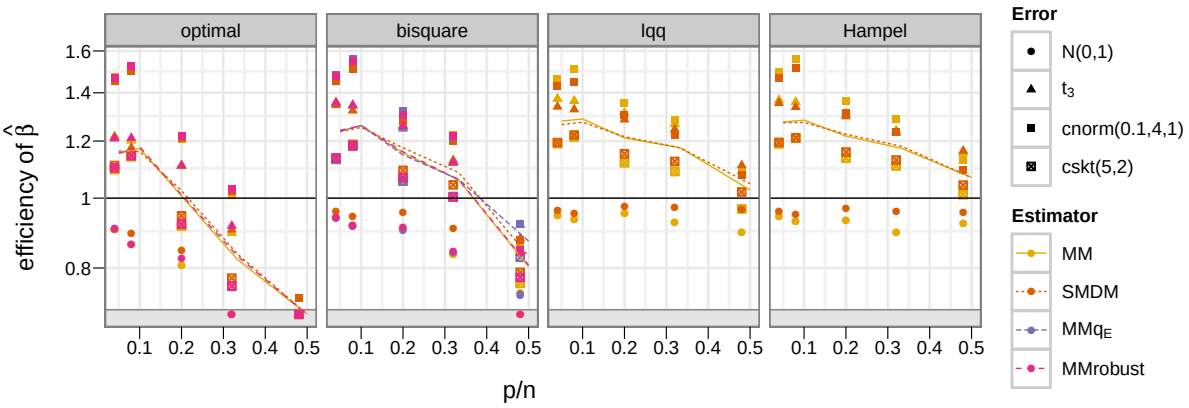

Figure 7: Efficiency of $\hat{\beta}$ with respect to OLS for different ψ-functions. The lines connect the median values over the simulated error distributions for each value of p/n. Results for $n = 25$ only.

cases the estimates using a quickly redescending ψ-function are even less efficient than the OLS-estimate. This does not happen for slowly redescending ψ-functions.

Levels. As can be seen from Fig. 8, there is a considerable difference in the behavior of the estimate as implemented in `lmRob` (denoted by *MMrobust.Wssc*) to the `lmrob` MM-estimate using Huber's small sample correction (*MM.Wssc*). Comparing the implementations of the two methods, the difference is quite subtle. While the latter uses the final residuals, the former uses the residuals of the initial S-estimate for estimating the covariance matrix. The method `lmRob`, in order to gain maximal robustness, uses the most robust residuals available to estimate the covariance matrix. This is a case where too much robustification leads to an undesirable erratic behavior.

There is a slight improvement when using (4) in combination with the SMDM-estimate (*SMDM.W*τ). However, most of the improvement comes from the slowly redescending ψ-function. The behavior for most of the estimates is quite similar, also when considering other error distributions (Fig. 9). The results for Avar_1 covariance matrix estimate depends strongly on the number of observations n. While for a low number of observations the empirical levels are a lot higher than the desired 5%, the results are much better for larger numbers of observations, especially for t_3 distributed errors. It is worth mentioning that the test based on the OLS-estimate keeps the empirical level of 5% for all the simulated error distributions (not shown on

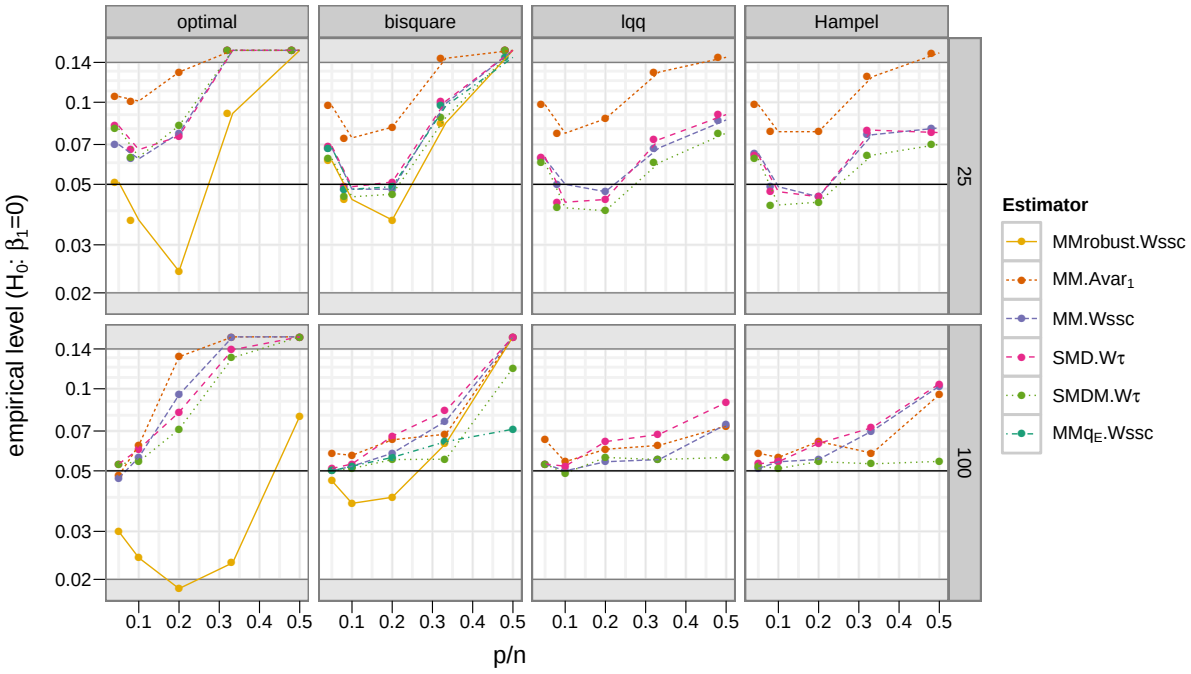

Figure 8: Empirical levels of test $H_0 : \beta_1 = 0$ for normal errors and different ψ-functions and sample sizes $n = 25$ (upper panel) and 100 (lower panel).

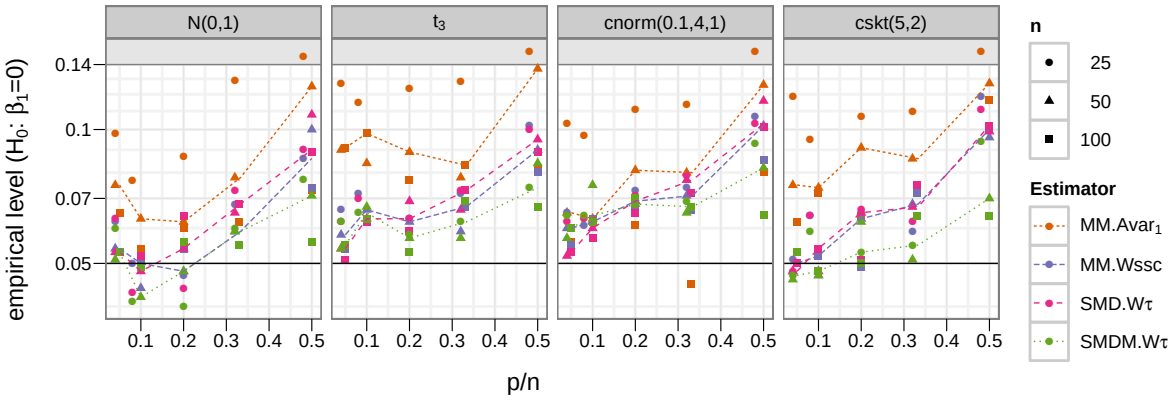

Figure 9: Empirical levels of test $H_0 : \beta_1 = 0$ for *lqq* ψ-function and different error distributions.

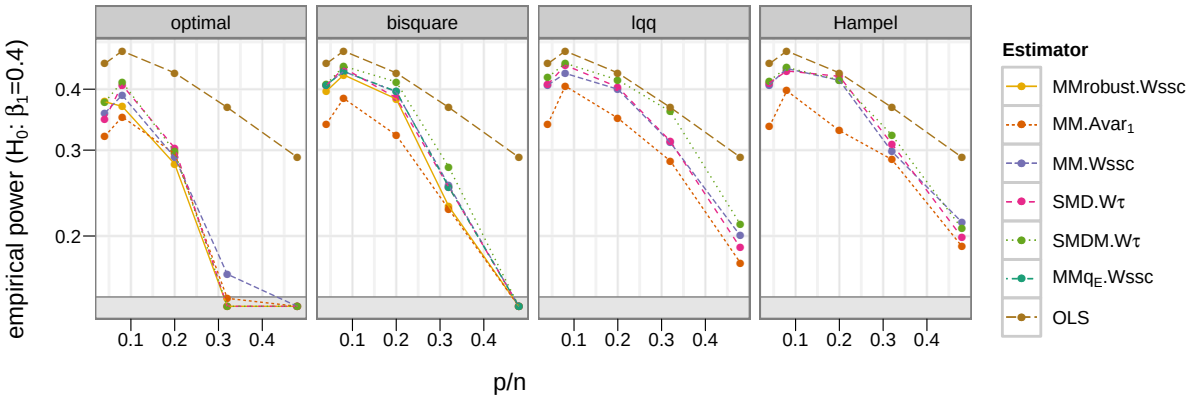

Figure 10: Empirical power of test $H_0 : \beta_1 = 0.4$ for different ψ-functions. Results for $n = 25$ and normal errors only.

plots).

Power. For comparing the power of the tests, we show only Fig. 10 and 11, where we test $H_0 : \beta_1 = b = 0.4$. Remember that the tests' critical values have been adjusted to keep the level of 0.05 – which would be tedious in practice. Aside from higher power, the plots for larger values b look similar. For lower values of b, the differences between the methods disappear.

For normally distributed errors, all tested methods lose a lot of power for increasing ratios of p/n, as can be seen in Fig. 10. The more robust estimator Avar_1 comes with the price of lower power than the other tested covariance matrix estimators. In the situation shown in Fig. 11, Avar_1 behaves better for asymmetric error distributions, but is still only equal in performance to the other methods. Interestingly, for larger values of n, this improvement disappears. All the simulated methods perform very badly for skewed t-distributed errors, while for the other error distributions, the power is slightly higher than for normally distributed errors.

8 A Real Data Example

As a complement to the simulation study, we also applied the proposed method to the nuclear power station dataset used in Cox and Snell (1981), which has been used by Davison and Hinkley (1997) and Brazzale et al. (2007) to compare their methods with the classical approach. The data is

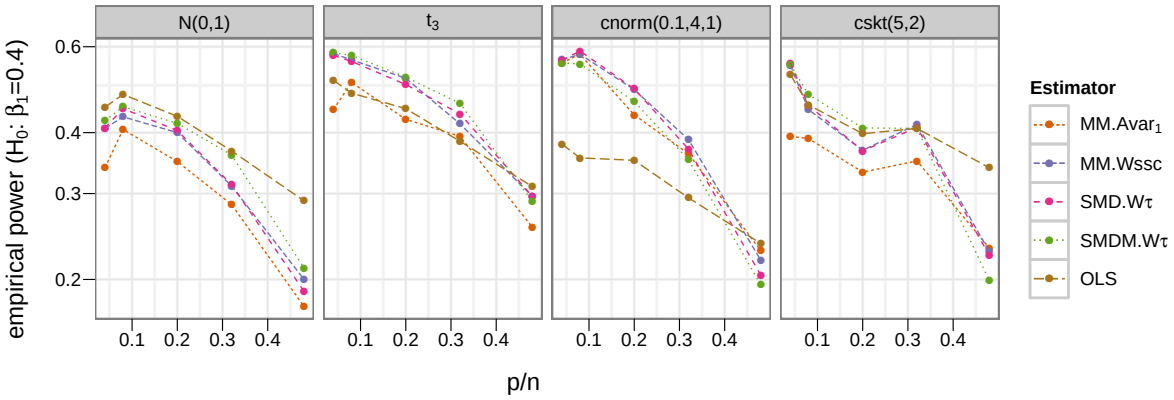

Figure 11: Empirical power for test $H_0 : \beta_1 = 0.4$ for different error distributions. Results for $n = 25$ and *lqq* ψ-function only.

on the construction cost of 32 light water reactor plants built in the USA in 1967 - 71. The full dataset contains ten possible explanatory variables, four of which can be dropped by some model selection technique as shown in the above cited references. Under the remaining six explanatory variables, there are two variables which are borderline significant. Subsequently, we will limit the discussion to these two variables, the number of power plants constructed by the same architect-engineer before ($\log(N)$) and the indicator for plants for which there was a partial turnkey guarantee (PT). The significance of the latter is crucial for the alleged basic question for which the study was undertaken.

When performing a residual analysis for the OLS fit, one will notice in the qq plot of the residuals that there is a tendency to a longer tailed error distribution. Davison and Hinkley (1997) consider the problem of predicting a new observation, calculating prediction intervals via ordinary bootstrap. We modified their approach to calculate confidence intervals for the parameter estimates. Brazzale et al. (2007) fitted a model where the responses have a Student t-distribution with 4 degrees of freedom. Standard errors and confidence intervals for the latter model were calculated using higher order asymptotics. We compare the fits of these models with MM and SMDM-estimates with *bisquare* and *lqq* ψ-functions. We used the covariance matrix estimates described in Section 5.

The parameter estimates and associated 95% confidence intervals are shown in Fig. 12. Since the correlation between $\log(N)$ and PT is -0.6,

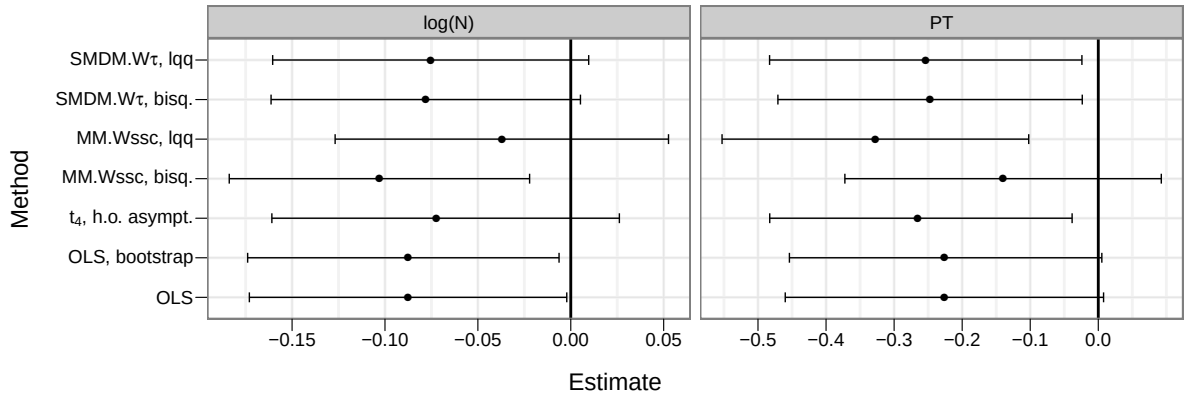

Figure 12: Estimates and 95% confidence intervals for selected explanatory variables of the nuclear power plant data. The naming of the methods follows the lines of Section 7.

the differences between the methods are quite similar for the two variables. The MM-estimates stand out, especially because of the strong influence of the ψ-function. While the initial S-estimate rejects observations 22 and 26 for both ψ-functions, after the M-step only the estimate using the *bisquare* ψ-function rejects the two observations. In fact, the MM-estimates for the two ψ-functions are quite different from each other, bracketing all the other results. After the D- and M-steps, because the D-scale estimate is substantially larger than the initial S-scale estimate, the two observations are not rejected anymore and the results are very similar for the two ψ-functions. For the SMDM-estimates, the minimum robustness weights are about 0.7, i.e., there are no observations rejected as outliers.

The comparison of all the methods does not really give any insights on which one of $\log(N)$ and PT should be treated as significant variable. The differences in the widths of the intervals are small compared to the differences in the parameter estimates. Therefore it is the location of the estimate that tips the scales and not the estimate of uncertainty. The upper ends of the bootstrap confidence intervals are so close to 0 that their sign depends on the random seed used. Thus the bootstrap results are also inconclusive. (To interested readers looking for a more satisfying answer, we recommend to plot *date* versus N using PT to color the points.)

9 Conclusions

The rate of redescent of the ψ-function used for MM-estimation is crucial for preserving the properties of the estimates with realistically small sample sizes. While all the compared ψ-functions have the same asymptotic efficiency, we have shown that there are large differences for larger ratios of number of predictors to observations. It is worth using a ψ-function not reaching the minimal possible maximal asymptotic bias. The proposed D-scale estimate in combination with a second M-step improves the performance of the estimates. The SMDM-estimate has the advantage of not requiring any correction factors for further inference. Using the τ-standardized residuals is enough, even for small samples.

The proposed method has been implemented in the function `lmrob` of the `R`-package `robustbase` (Rousseeuw et al., 2011, version 0.6-5 or younger) With the argument `setting="KS2011"`, the recommended parameters are set automatically: SMDM-estimator with the *lqq* ψ-function tuned for 95% asymptotic efficiency and the covariance estimate using (4) and (5) as described in Section 5. This setting was denoted as $\hat{\sigma}_{\mathrm{D}}$, *SMDM* and *SMDM.W*$\tau$ above.

Appendix. Details on the Design Adaptive Scale Estimate

For OLS, the scale estimator can be written as the solution of

$$\sum_{i=1}^{n} \tau_i^2 \left(\left(\frac{r_i}{\tau_i \sigma} \right)^2 - 1 \right) = 0,$$

with $\tau_i = \sqrt{1 - h_i}$. This equation, which can be solved explicitly, makes the argument for dividing the sums of squares by $n - p$ instead of n more transparent: Every summand in this implicit equation has the expectation 0, which leads to the unbiasedness of $\hat{\sigma}^2$. We robustify this equation by introducing the robustness weights to get (2), where κ is needed to obtain consistency at the central model as usual, i.e.,

$$\kappa = \frac{\mathbf{E}_0 \left[w(e) e^2 \right]}{\mathbf{E}_0 \left[w(e) \right]}.$$

The subscript 0 indicates the evaluation at the central model with scale parameter 1. The D-scale can be reliably calculated by means of an iterative reweighing algorithm. The starting value,

$$\hat{\sigma}_0 = \sqrt{\frac{\sum_{i=1}^{n} w_i r_i^2}{\kappa \sum_{i=1}^{n} w_i \tau_i^2}},$$

using the robustness weights w_i from the last M-step proved to be efficient.

We define τ_i as the value that zeroes the expectation of the ith summand in (2),

$$\mathbf{E}\left[w\left(\frac{r_i}{\tau_i \sigma}\right)\left(\frac{r_i}{\tau_i \sigma}\right)^2 - \kappa w\left(\frac{r_i}{\tau_i \sigma}\right)\right] = 0. \tag{6}$$

Since the exact distribution of the residuals is unknown, we approximate it using a *von Mises Expansion* of $\hat{\beta}$, see (3). After splitting the contributions of the ith observation and the other observations, and approximating the latter by a normally distributed random variable, we get

$$r_i \approx e_i - \frac{1}{n}\boldsymbol{x}_i' \,\mathrm{IF}(e_i, \boldsymbol{x}_i, \sigma) + u_{-i},$$

with

$$u_{-i} \sim \mathcal{N}\left(0, \frac{\mathbf{E}_0\left[\psi^2(e)\right]}{\mathbf{E}_0\left[\psi'(e)\right]^2}\left(h_i - h_i^2\right)\right).$$

We are then able to solve (6) using the above approximations and standard numerical integration and root-search procedures. Finding the root of (6) can be difficult for ψ-functions with very small support. In this case the curve around the real root is very flat and therefore a small error in the numerical integration can translate to a comparably large error in the solution. Because the ψ-functions of the M-step are tuned for high efficiency and thus have a quite large support, this causes no concern in applications.

It turns out that τ_i depends only on the ψ-function used as well as the leverage of the ith observation. It can be shown that the values are well approximated by a function of the form

$$\tau_i \approx (1 - c_1 h_i)\sqrt{1 - c_2 h_i}, \tag{7}$$

where

$$c_2 = 2\frac{\mathbf{E}_0\left[\psi(e)e\right]}{\mathbf{E}_0\left[\psi'(e)\right]} - \frac{\mathbf{E}_0\left[\psi(e)^2\right]}{\mathbf{E}_0\left[\psi'(e)\right]^2},$$

while c_1 is determined empirically by fitting (7) (with, e.g., a MM-estimator) to the exact solutions of (6) for a set of leverages ranging from 0 to 0.8. Since robustness is needed with regard to the leverage, we use the robustified estimator $\hat{\boldsymbol{V}}_X$ to calculate the h_is,

$$h_i = w_i \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{x}_i.$$

References

Berrendero, J., Mendes, B., Tyler, D., 2007. On the maximum bias functions of MM-estimates and constrained M-estimates of regression. Annals of Statistics 35 (1), 13.

Brazzale, A., Davison, A., Reid, N., 2007. Applied asymptotics: case studies in small-sample statistics. Cambridge University Press.

Cox, D. R., Snell, E. J., 1981. Applied Statistics. Chapman and Hall, London.

Croux, C., Dhaene, G., Hoorelbeke, D., 2003. Robust standard errors for robust estimators. Tech. rep., Dept. of Applied Economics, K.U. Leuven.

Davison, A. C., Hinkley, D. V., 1997. Bootstrap methods and their application. Cambridge University Press.

Fernández, C., Steel, M., 1998. On Bayesian modeling of fat tails and skewness. Journal of the American Statistical Association 93 (441), 359–371.

Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, N.Y.

Hennig, C., 1995. Efficient high breakdown point estimators in robust regression: which function to choose. Statist. Decisions 13, 221–241.

Huber, P., 1973. Robust regression: asymptotics, conjectures and Monte Carlo. The Annals of Statistics 1 (5), 799–821.

Huber, P. J., Ronchetti, E. M., 2009. Robust Statistics. Wiley, N.Y.

King, R., Anderson, E., 2004. skewt: The Skewed Student-t Distribution. R package version 0.1.
URL http://CRAN.R-project.org/package=skewt

Koller, M., 2011. Simulations for Sharpening Wald-type Inference in Robust Regression for Small Samples. Vignette of robustbase: Basic Robust Statistics, R package version 0.7-0.
URL http://CRAN.R-project.org/package=robustbase

Maronna, R. A., Martin, R. D., Yohai, V. J., 2006. Robust Statistics, Theory and Methods. Wiley, N.Y.

Maronna, R. A., Yohai, V. J., 2010. Correcting MM estimates for "fat" data sets. Computational Statistics & Data Analysis 54 (12), 3168–3173.

Martin, R., Yohai, V., Zamar, R., 1989. Min-max bias robust regression. The Annals of Statistics 17 (4), 1608–1630.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Maechler, M., 2011. robustbase: Basic Robust Statistics. R package version 0.7-0.
URL http://CRAN.R-project.org/package=robustbase

Svarc, M., Yohai, V., Zamar, R., 2002. Optimal bias-robust M-estimates of regression. In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L1 Norm and Related Methods. Birkhäuser, pp. 191–200.

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., Zivot, E., Rocke, D., Martin, D., Maechler, M., Konis, K., 2010. robust: Insightful Robust Library. R package version 0.3-11.
URL http://CRAN.R-project.org/package=robust

Yohai, V., 1987. High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics 15 (2), 642–656.

Yohai, V., Stahel, W., Zamar, R., 1991. A procedure for robust estimation and inference in linear regression. In: Stahel, Weisberg (Eds.), Directions in Robust Statistics and Diagnostics. Springer, N.Y., pp. 365–374.

Yohai, V., Zamar, R., 1997. Optimal locally robust M-estimates of regression. Journal of Statistical Planning and Inference 64 (2), 309–323.