Globally robust inference for the location and simple linear regression models

Jorge Adrover, Matias Salibian-Barrera, Ruben Zamar

Abstract

We define globally robust confidence intervals and p-values for the location and simple linear regression models. The need for robust inference has been noticed and partially addressed in the statistical literature (see for example the book by Barnett and Lewis (1994) and references therein). We construct intervals that are stable in the sense of achieving coverages near the nominal ones even in the presence of outliers and other departures from the parametric model. Moreover, our intervals are informative in the sense of having relatively short lengths. These globally robust confidence intervals constitute an improvement over previous robust intervals which do not take into account the potential bias of the estimates.

1 Introduction

To fix ideas we will start our discussion by considering the following simple *location-scale* model:

$$y = \theta + \sigma\varepsilon\,. \tag{1}$$

The simple linear regression model is discussed in Section 3. In the model (1) θ is an unknown location parameter, σ is a nuisance scale parameter, and ε has a specified distribution F_0. Correspondingly, the distribution of y is $F_\theta(y) = F_0((y-\theta)/\sigma)$. To allow for outliers and other departures from this model, we assume that the actual distribution F belongs to the ϵ*−contamination neighborhood* (see Tukey (1960)),

$$\mathcal{F}_\epsilon(F_\theta) = \{F = (1-\epsilon)F_\theta + \epsilon F^* : F^* \text{ any arbitrary distribution}\}, \quad 0 < \epsilon < 1/2. \tag{2}$$

According to this model, the majority of the data comes from a central parametric model but a minority may come from an unknown arbitrary distribution.

A robust confidence interval should maintain a high coverage level and a reasonable length over the whole contamination neighborhood. These two properties are formally stated in the following definition.

Definition 1 *A confidence interval* (L_n, U_n) *for* θ *is called* globally robust *of level* $(1-\alpha)$ *if it satisfies the following conditions:*

1. *(*Stable interval*) The minimum asymptotic coverage over the* ϵ*-contamination neighborhood is* $(1-\alpha)$*:*
$$\lim_{n\to\infty} \inf_{F\epsilon\mathcal{F}_\epsilon(F_\theta)} P_F\left(L_n < \theta < U_n\right) \geq (1-\alpha);$$

2. (Informative interval) *The maximum asymptotic length of the interval is bounded over the ϵ-contamination neighborhood:*

$$\lim_{n\to\infty} \sup_{F\epsilon\mathcal{F}_\epsilon(F_\theta)} [U_n - L_n] < \infty.$$

These properties have been previously considered in the literature. See, for example, Tukey and McLaughlin (1963), Dixon and Tukey (1968), Huber (1968, 1970) and Barnett and Lewis (1994, p. 74) and references therein.

Not surprisingly, *classical Student-t confidence intervals*, $\overline{X}_n \pm t_{(n-1)}(\alpha/2)S_n/\sqrt{n}$, fail Parts 1 and 2 of Definition 1. To see this, consider $F^{x_0} = (1-\epsilon)F_0 + \epsilon F^*$, where F^* is a point mass distribution at $x_0 > 0$. Then, L_n and U_n converge almost surely (F^{x_0}) to $\epsilon x_0 > 0$ and we get

$$\lim_{n\to\infty} \inf_{F\epsilon\mathcal{F}_\epsilon(F_0)} P_F\left(L_n < 0 < U_n\right) \le \lim_{n\to\infty} P_{F^{x_0}}\left(L_n < 0 < U_n\right) = 0,$$

showing that Student t-intervals fail Part 1. If we take now F^* above to be a point mass distribution at $\pm x_0$ (equally weighted) we have

$$\lim_{n\to\infty} \sup_{F\epsilon\mathcal{F}_\epsilon(F_0)} [U_n - L_n] \ge \lim_{n\to\infty} \sup_{x_0>0} 2\ t_{(n-1)}(\alpha/2)\frac{x_0}{\sqrt{n}} = \infty$$

failing Part 2.

The problem with the above confidence intervals does not solely lie on the lack of robustness of $\overline{X}_n$ and S_n. It can be shown that if we use a robust location estimate and a robust scale estimate the resulting confidence interval satisfies Part 2 but not Part 1 of the above Definition. In other words, if we do not take into account the effect of the asymptotic bias any such confidence interval will have asymptotic minimum coverage level over the contamination neighborhood equal to 0, regardless of the target level $1-\alpha$.

The effect of asymptotic bias on the coverage level of intervals of fixed length was first studied by Huber (1968). Let $\hat{\theta}_n$ be a robust location estimate and let $\hat{\theta}(F)$ be its asymptotic value under an asymmetric distribution F. Since $\hat{\theta}(F)$ is different from the actual value of the parameter, for large sample sizes most of the probability mass in the distribution of $\hat{\theta}_n$ concentrates on an interval that excludes the true location parameter θ. This lack of robustness due to asymptotic bias is more severe in the simple linear regression model (see Section 3).

The rest of the paper is organized as follows. Section 2 defines robust confidence intervals and p-values for the location parameter. Section 3 defines robust confidence intervals for simple linear regression. Section 4 gives an example. Section 5 presents some numerical results and Section 6 gives some concluding remarks. An appendix with proofs is also included.

2 Globally robust inference for location

In this section we show how to construct confidence intervals, bounds and p-values that satisfy our robustness requirements in terms of stability and informativeness (see Definition 1).

2.1 Confidence intervals

Suppose that we have a translation equivariant estimate $\hat{\theta}_n$ for the parameter θ in model (1) which satisfies

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}(F)) \ \to_d \mathrm{N}(0, v^2(F)), \tag{3}$$

where $|\hat{\theta}(F) - \theta| < \bar{\theta}$ and $\bar{\theta}$ is an upper bound for the bias of $\hat{\theta}(F)$ as defined in Berrendero and Zamar (2001). Salibian-Barrera and Zamar (2002) showed that, under certain regularity conditions on the estimating equations, MM-location estimators are asymptotically normally distributed uniformly on the contamination neighborhood (see also Fraiman *et al.* (2001)), and Salibian-Barrera (2000) showed that under certain regularity conditions MM-regression estimates are asymptotically normal for any distribution in a contamination neighborhood.

The fraction ϵ of contamination in the data is usually unknown. In some cases (see the example in Section 4) ϵ can be inferred from a careful screening of the data. Another approach is to find the largest ϵ for which the robust inference methods proposed below remain informative (e.g. the robust test rejects the null hypothesis). In this way these robust inference methods may render a better understanding of the data even in the absence of a precise knowledge of ϵ.

For any *fixed* $F \in \mathcal{F}_\epsilon(F_\theta)$, an asymptotic confidence interval of level $(1-\alpha)$ is given by

$$\left(\hat{\theta}_n - l_n, \hat{\theta}_n + r_n\right),$$

where $l_n = l_n(F)$ and $r_n = r_n(F)$ satisfy the equation

$$P_F\left(-r_n \leq \hat{\theta}_n - \theta \leq l_n\right) = 1 - \alpha. \tag{4}$$

We can write

$$\begin{aligned} P_F\left(-r_n \leq \hat{\theta}_n - \hat{\theta}(F) + \hat{\theta}(F) - \theta \leq l_n\right) &= \\ P_F\left(\frac{-r_n - b}{v_n} \leq \frac{\hat{\theta}_n - \hat{\theta}(F)}{v_n} \leq \frac{l_n - b}{v_n}\right) &= 1 - \alpha, \end{aligned} \tag{5}$$

where $\sqrt{n}v_n$ is a consistent estimate for $\sigma(F)$ for all $F \in \mathcal{F}_\epsilon(F_\theta)$, and $b = b(F) = \hat{\theta}(F) - \theta$.

From equation (3) we can use the normal distribution to approximate the left-hand side of (5). Hence, we can obtain estimates $\hat{l}_n$ and $\hat{r}_n$ solving the following equation:

$$\Phi\left(\frac{\hat{l}_n - b}{v_n}\right) + \Phi\left(\frac{\hat{r}_n + b}{v_n}\right) - 1 = 1 - \alpha, \tag{6}$$

where Φ is the standard normal cumulative distribution function. Since $F \in \mathcal{F}_\epsilon(F_\theta)$ is *unspecified,* the bias $b = b(F)$ in (6) is unknown and *cannot* be estimated from the data. Consequently, we will find endpoints $\hat{l}_n$ and $\hat{r}_n$ such that

$$\Phi\left(\frac{\hat{l}_n - b(F)}{v_n}\right) + \Phi\left(\frac{\hat{r}_n + b(F)}{v_n}\right) - 1 \geq 1 - \alpha, \quad \text{for all} \;\; F \in \mathcal{F}_\epsilon(F_\theta).$$

In this way we can obtain an *stable* confidence interval with coverage at least $1-\alpha$ over the entire neighborhood.

Using Equation (6) we find $\hat{l}_n$ as a function of $\hat{r}_n$:

$$\hat{l}_n = v_n \Phi^{-1}\left[2 - \alpha - \Phi\left(\frac{\hat{r}_n + b}{v_n}\right)\right] + b,$$

and choose $\hat{r}_n$ in order to minimize the resulting interval length:

$$\hat{l}_n + \hat{r}_n = v_n \Phi^{-1} \left[2 - \alpha - \Phi \left(\frac{\hat{r}_n + b}{v_n} \right) \right] + b + \hat{r}_n. \tag{7}$$

Differentiating the right-hand side of (7) with respect to $\hat{r}_n$ and setting the derivative equal to zero we have

$$-\frac{\varphi\left(\frac{\hat{r}_n+b}{v_n}\right)}{\varphi\left(\Phi^{-1}\left[2-\alpha-\Phi\left(\frac{\hat{r}_n+b}{v_n}\right)\right]\right)} + 1 = 0,$$

where φ is the standard normal density function. That is

$$\varphi\left(\frac{\hat{r}_n + b}{v_n}\right) = \varphi\left(\Phi^{-1}\left[2 - \alpha - \Phi\left(\frac{\hat{r}_n + b}{v_n}\right)\right]\right).$$

From this equation it follows that

$$\frac{\hat{r}_n + b}{v_n} = \pm\ \Phi^{-1}\left[2 - \alpha - \Phi\left(\frac{\hat{r}_n + b}{v_n}\right)\right].$$

The minus sign case is discarded because it can only be satisfied with $\alpha = 1$. From the other case we obtain

$$\Phi\left(\frac{\hat{r}_n + b}{v_n}\right) = 2 - \alpha - \Phi\left(\frac{\hat{r}_n + b}{v_n}\right) \tag{8}$$

which yields the solution

$$\begin{aligned} \hat{r}_n &= v_n \Phi^{-1}(1 - \alpha/2) - b = v_n z_{\alpha/2} - b \\ \hat{l}_n &= v_n z_{\alpha/2} + b. \end{aligned}$$

The corresponding confidence interval for θ of level greater than or equal to $(1 - \alpha)$ for fixed $F \in \mathcal{F}_\epsilon(F_\theta)$ is:

$$\begin{aligned} \hat{I}_n(F) &= \left(\hat{\theta}_n - v_n z_{\alpha/2} - b,\ \hat{\theta}_n + v_n z_{\alpha/2} - b\right) \\ &= \left(\hat{\theta}_n - v_n z_{\alpha/2},\ \hat{\theta}_n + v_n z_{\alpha/2}\right) - b \end{aligned}$$

Since $\hat{I}_n(F)$ still depends on the unknown F through b, a robust interval of level $(1 - \alpha)$ can be constructed as follows:

$$I_n = \bigcup_{F \in \mathcal{F}_\epsilon(F_\theta)} \left[\left(\hat{\theta}_n - v_n z_{\alpha/2},\ \hat{\theta}_n + v_n z_{\alpha/2}\right) - b\right] \tag{9}$$

Since $|b| = |b(F)| < \bar{\theta}$ for all $F \in \mathcal{F}_\epsilon(F_\theta)$ we have

$$I_n \subseteq \left(\hat{\theta}_n - v_n z_{\alpha/2} - \bar{\theta},\ \hat{\theta}_n + v_n z_{\alpha/2} + \bar{\theta}\right)$$

with equality if $\bar{\theta}$ is a sharp bias bound, that is if $\bar{\theta} = \sup_{\mathcal{F}_\epsilon(F_\theta)} |b(F)|$. Note that the above robust confidence intervals can be constructed for any estimate $\hat{\theta}_n$ that is asymptotically normal, has a consistent estimate of its standard error, and a know bias bound $\bar{\theta}$. In order to obtain the shortest robust confidence interval we will consider estimates where such a sharp bound is available. It is easy to verify that the above confidence interval is robust according to Definition 1.

The above derivation extends to the case of robust upper and lower confidence bounds (see Section 2.2) and will also allow us to derive robust p-values for one- and two-sided hypothesis tests (see Section 2.3).

An alternative robust confidence interval can be constructed following Fraiman *et al.* (2001). The main idea is to find q_n such that $P_F(|\hat{\theta}_n - \theta| \leq q_n) = 1 - \alpha$. Because of the asymptotic normality of $\hat{\theta}_n$ an estimate of q_n can be found by solving

$$\Phi\left(\frac{\tilde{q}_n - b}{v_n}\right) + \Phi\left(\frac{\tilde{q}_n + b}{v_n}\right) - 1 = 1 - \alpha, \tag{10}$$

Note that the solution $\tilde{q}_n$ to (10) is a strictly increasing function of $|b|$ for each fixed v_n (and of v_n for each fixed $|b|$). That is, the quantile $\tilde{q}_n$ is a monotone function of the bias (and standard deviation) and its largest possible value is obtained by replacing $|b|$ by $\bar{\theta}$:

$$\Phi\left(\frac{\bar{q}_n - \bar{\theta}}{v_n}\right) + \Phi\left(\frac{\bar{q}_n + \bar{\theta}}{v_n}\right) - 1 = 1 - \alpha. \tag{11}$$

This yields the robust confidence interval

$$\hat{\theta}_n \pm \bar{q}_n. \tag{12}$$

Note that $\bar{q}_n$ and $v_n z_{\alpha/2} + \bar{\theta}$ converge to $\bar{\theta}$ as $n \to \infty$, and hence the asymptotic lengths of these two confidence intervals equal $2\bar{\theta}$. This positive asymptotic length is the price we pay to mantain the nominal coverage level over the whole contamination neighborhood. However, note that if the upper bound $\bar{\theta}$ is sharp then $2\bar{\theta}$ is the smallest asymptotic length of any *robust* confidence interval (based on $\hat{\theta}_n$). In fact, any shorter confidence interval (based on $\hat{\theta}_n$) will not attain the nominal coverage level over the neighbourhood. The robust confidence interval for θ with shortest asymptotic length will be obtained when the estimator $\hat{\theta}_n$ minimizes the bias bound $\bar{\theta}$ (see Section 2.4 for the choice of $\hat{\theta}_n$, and Berrendero and Zamar (2001) for a definition and details on bias bounds).

The robust interval (9) is easier to compute but slightly longer than that in (12) (that is: $v_n z_{\alpha/2} + \bar{\theta} > \bar{q}_n$) as can be seen by sustituting $v_n z_{\alpha/2} + \bar{\theta}$ in (11).

2.2 One sided confidence bounds

In this section we derive robust lower and upper bounds for the parameter θ. To save space, we simultaneously derive both types of bounds, using square braquets to indicate the corresponding steps leading to robust upper bounds.

For any fixed $F \in \mathcal{F}_\epsilon(F_\theta)$, an asymptotic lower [upper] bound of level $(1-\alpha)$ for θ is given by

$$\left(\hat{\theta}_n - l_n, \infty\right) \qquad \left[\left(-\infty, \hat{\theta}_n + r_n\right)\right] \tag{13}$$

where $l_n = l_n(F)$ $[r_n = r_n(F)]$ satisfies the equation

$$P_F\left(\hat{\theta}_n - \theta \leq l_n\right) = 1 - \alpha \qquad \left[P_F\left(\hat{\theta}_n - \theta \geq -r_n\right) = 1 - \alpha\right]. \tag{14}$$

Similarly to (5), and recalling that $b = b(F) = \hat{\theta}(F) - \theta$, we have that l_n $[r_n]$ must satisfy

$$P_F\left(\frac{\hat{\theta}_n - \hat{\theta}(F)}{v_n} \leq \frac{l_n - b}{v_n}\right) = 1-\alpha \qquad \left[P_F\left(\frac{-r_n - b}{v_n} \leq \frac{\hat{\theta}_n - \hat{\theta}(F)}{v_n}\right) = 1-\alpha\right] \tag{15}$$

and, as before, for large n,

$$\hat{l}_n(F) = v_n\Phi^{-1}(1-\alpha) + b \qquad \left[\hat{r}_n(F) = v_n\Phi^{-1}(1-\alpha) - b\,\right]. \tag{16}$$

A robust lower [upper] bound of level $1-\alpha$ can now be defined as

$$(L_n^R, \infty) = \bigcup_{F\in\mathcal{F}_\epsilon(F_\theta)} \left(\hat{\theta}_n - \hat{l}_n(F), \infty\right) \qquad \left[(-\infty, U_n^R) = \bigcup_{F\in\mathcal{F}_\epsilon(F_\theta)} \left(-\infty, \hat{\theta}_n + \hat{r}_n(F)\right)\right]. \tag{17}$$

Note that

$$\left(\hat{\theta}_n - \hat{l}_n(F), \infty\right) = \left(\hat{\theta}_n - v_n\Phi^{-1}(1-\alpha) - b, \infty\right) \ \left[\left(-\infty, \hat{\theta}_n + \hat{r}_n(F)\right) = \left(-\infty, \hat{\theta}_n + v_n z_\alpha + b\right)\right]. \tag{18}$$

Hence we obtain

$$\left(L_n^R, \infty\right) = \left(\hat{\theta}_n - v_n z_\alpha - \bar{\theta}, +\infty\right) \qquad \left[\left(-\infty, U_n^R\right) = \left(-\infty, \hat{\theta}_n + v_n z_\alpha + \bar{\theta}\right)\right].$$

Note that these bounds are *stable* as in Definition 1.

It is easy to see that unlike in the confidence interval case, if we use a construction similar to the one in Fraiman *et al.* (2001), we obtain the same lower [upper] robust confidence bounds. An important observation is that this second construction allows us to show that $|\hat{l}_n - l_n| = o_p(1/\sqrt{n})$ $[|\hat{r}_n - r_n| = o_p(1/\sqrt{n})]$.

2.3 P-values

In this section we will show how to construct robust p-values and tests using the robust confidence intervals and bounds derived above.

We start by considering the following artificial hypothesis testing problem. Assume that the contaminating distribution H in $F = (1-\epsilon)F_\theta + \epsilon H$ is known, that is, that F belongs to the following "translated" family

$$\mathcal{F}_\epsilon(F_\theta, H) = \left\{F = (1-\epsilon)F_\theta + \epsilon H, \text{ fixed } H\ \right\}. \tag{19}$$

Suppose that we are interested in testing

$$H_0\ :\ \theta \leq \theta_0 \quad \text{ versus } \quad H_1\ :\ \theta > \theta_0 \tag{20}$$

for $F \in \mathcal{F}_\epsilon(F_\theta, H)$. Using the lower bounds in (18) define $\hat{\theta}_n(\alpha, F) = \hat{\theta}_n - v_n\Phi^{-1}(1-\alpha) - b(F)$. and a test for (20) is given by the rejection rule:

$$\text{Reject H}_0 \ \text{ if } \quad \hat{\theta}_n(\alpha, F) > \theta_0.$$

Notice that this is a size α test since

$$\sup_{\theta\leq\theta_0} P_{(1-\epsilon)F_\theta+\epsilon H}\left(\hat{\theta}_n(\alpha,(1-\epsilon)F_\theta+\epsilon H)>\theta_0\right)$$
$$\leq P_{(1-\epsilon)F_{\theta_0}+\epsilon H}\left(\hat{\theta}_n(\alpha,(1-\epsilon)F_{\theta_0}+\epsilon H)>\theta_0\right)=\alpha.$$

In other words, for the fixed family (19) the p-value is given by

$$\hat{p}_n(F)=\inf\left\{\alpha:\hat{\theta}_n(\alpha,F)>\theta_0\right\}. \tag{21}$$

We can calculate $\hat{p}_n(F)$ explicitely because $\hat{\theta}_n(\alpha,F)$ is increasing in α and hence $\hat{p}_n(F)$ satisfies $\hat{\theta}_n(\hat{p}_n(F),F)=\theta_0$,or equivalently,

$$\hat{p}_n(F)=1-\Phi\left(\frac{\hat{\theta}_n-\theta_0-b}{v_n}\right). \tag{22}$$

Analogously, for the case $H_0:\theta\geq\theta_0$ versus $H_1:\theta<\theta_0$, we have $\hat{p}_n(F)=\Phi\left(\frac{\hat{\theta}_n-\theta_0-b}{v_n}\right)$.

Consider now the 2-sided hypothesis

$$H_0:\theta=\theta_0 \quad \text{versus} \quad H_1:\theta\neq\theta_0.$$

Since the confidence interval $\hat{\theta}_n\pm\tilde{q}_n$, where $\tilde{q}_n$ satisfies (10) is shorter than that in (9), we construct the p-value for this case based on $\tilde{q}_n(\alpha,F)$ defined by the equation

$$\Phi\left(\frac{\tilde{q}_n(\alpha,F)-b}{v_n}\right)+\Phi\left(\frac{\tilde{q}_n(\alpha,F)+b}{v_n}\right)-1=1-\alpha\,. \tag{23}$$

The rejection rule would be of the form: reject H_0 if $|\hat{\theta}_n-\theta_0|>\tilde{q}_n(\alpha,F)$. The p-value is

$$\hat{p}_n(F)=\inf\left\{\alpha:\tilde{q}_n(\alpha,F)<|\hat{\theta}_n-\theta_0|\right\}.$$

We can easily verify that $\tilde{q}_n(\alpha,F)$ is a decreasing function of α. Therefore $\hat{p}_n(F)$ solves the equation

$$\tilde{q}_n(\hat{p}_n(F),F)=|\hat{\theta}_n-\theta_0|\,. \tag{24}$$

From (23) it is easy to see that the solution $g(x)$ to the equation $\tilde{q}_n(g(x),F)=x$ is given by

$$g(x)=2-\Phi\left(\frac{x-b}{v_n}\right)-\Phi\left(\frac{x+b}{v_n}\right). \tag{25}$$

We obtain

$$\hat{p}_n(F)=2-\Phi\left(\frac{|\hat{\theta}_n-\theta_0|-b}{v_n}\right)-\Phi\left(\frac{|\hat{\theta}_n-\theta_0|+b}{v_n}\right). \tag{26}$$

Recall that the p-values (22) and (26) were obtained for $F\in\mathcal{F}_\epsilon(F_\theta,H)$. We can then define a robust p-value for all $F\in\mathcal{F}_\epsilon(F_\theta)$ as

$$\hat{p}_n^R=\sup_{F\in\mathcal{F}_\epsilon(F_\theta)}\hat{p}_n(F). \tag{27}$$

Hypothesis	Robust p-values
$H_1 \ : \ \theta > \theta_0$	$1 - \Phi\left(\frac{\hat{\theta}_n - \theta_0 - \bar{\theta}}{\hat{\sigma}_n}\right)$
$H_1 \ : \ \theta < \theta_0$	$\Phi\left(\frac{\hat{\theta}_n - \theta_0 + \bar{\theta}}{\hat{\sigma}_n}\right)$
$H_1 \ : \ \theta \neq \theta_0$	$2 - \Phi\left(\frac{\lvert\hat{\theta}_n - \theta_0\rvert - \bar{\theta}}{\hat{\sigma}_n}\right) - \Phi\left(\frac{\lvert\hat{\theta}_n - \theta_0\rvert + \bar{\theta}}{\hat{\sigma}_n}\right)$

Table 1: Robust p-values for location-scale problems

It is easy to see that $\hat{p}_n(F)$ is a monotone function of $b = b(F)$ and hence, if the bias bound $\bar{\theta}$ is sharp, then the robust p-values are shown in Table 1.

Note that, in the same way the robust confidence intervals of Section 2.1 have a positive asymptotic length, the rejection rules associated with the p-values in (27) have robust Type I Errors at the expense of lower power. In other words, to guarantee a level-α test *with uncertainty in our model* we lose power, in particular for values of the parameter near the null hypothesis. From a practical point of view, however, lower power is a reasonable price to pay to achieve a *robust rejection rule.*

2.4 The choice of the point estimate

As discussed at the end of Section 2.1, the robust confidence intervals proposed in this paper have minimum asymptotic length when $\hat{\theta}_n$ minimizes the bias bound $\bar{\theta}$. However, for finite samples, the length of these robust confidence intervals also depends on the estimated standard error v_n, so the best choice for a point estimate should consider its efficiency as well. Moreover, in practice the particular way in which v_n is calculated will impact on the intervals coverage and lengths. Another point of practical importance is the ability to calculate the bias bound $\bar{\theta}$.

Under the simplifying assumption that the error scale is known Fraiman *et al.* (2001) find the M-estimate that minimizes the maximum asymptotic length of robust confidence intervals similar to the ones defined in (12). The main difference lies in the treatment of the estimator standard error. While we propose using an estimate v_n, Fraiman *et al* (2001) propose replacing it by an asymptotic upper bound. Their results suggest that a good choice for a location M-estimate, in terms of asymptotic length of the resulting confidence interval, solves

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \hat{\theta}_n}{s_n}\right) = 0, \tag{28}$$

with score function $\psi(u) = \psi_1^D(u/1.345)$, where ψ_1^D is a (twice continuously differntiable) approximation to Huber's score function $\psi_1^H(u) = \min\{\max\{u, -1\}, 1\}$. The scale estimate s_n is the scale S-estimate (Rousseeuw and Yohai, 1984) with score function χ also recommended in the given reference.

Following Berrendero and Zamar (2001), the bias bound for $\hat{\theta}_n$ is estimated by

$$\bar{\theta}_n = k \ s_n \ B(\epsilon), \tag{29}$$

with

$$k = \sup_{F \in \mathcal{F}_\epsilon} \frac{s(F_0)}{s(F)} = \frac{1}{s^-(\epsilon)}, \quad s_n = \text{shorth}(y_i) \quad \text{and} \quad s^-(\epsilon) = \inf_{F \in \mathcal{F}_\epsilon} \frac{s(F)}{s(F_0)},$$

and the shorth is the standardized length of the shortest half of the sorted data. For a precise definition of the shorth see Rousseeuw and Leroy (1987). Martin and Zamar (1993) gave formulae for $B(\epsilon)$ and $s^{-}(\epsilon)$

Finally, we need an estimate of the standard error of $\hat{\theta}_n$ which is consistent for all distributions in the neighborhood. A good choice is the fast robust bootstrap in Salibian-Barrera and Zamar (2002).

2.5 Numerical results

We will illustrate the performance of the robust confidence intervals of a location parameter by a small Monte Carlo simulation.

We considered two levels of point mass contamination δ_{x_0} which we call "mild contamination" in the case of $x_0 = 1.5$ and "strong contamination" in the case $x_0 = 4.0$. The results for the estimate proposed in (28) above are reported in Table 4. As we can expect from the theoretical considerations above, there is a clear superiority of the globally robust approach over the "naive" approach (see Remark 1), specially for larger values of n and ϵ.

ϵ	n	Globally Robust		Naive	
		$x_0 = 1.5$	$x_0 = 4.0$	$x_0 = 1.5$	$x_0 = 4.0$
0.00	20	0.94 (0.88)	0.94 (0.88)	0.94 (0.88)	0.94 (0.88)
	40	0.94 (0.64)	0.94 (0.64)	0.94 (0.64)	0.94 (0.64)
	100	0.94 (0.41)	0.94 (0.41)	0.94 (0.41)	0.94 (0.41)
	500	0.93 (0.18)	0.93 (0.18)	0.93 (0.18)	0.93 (0.18)
0.01	20	0.93 (0.89)	0.94 (0.88)	0.94 (0.89)	0.94 (0.89)
	40	0.95 (0.64)	0.95 (0.64)	0.95 (0.64)	0.95 (0.63)
	100	0.95 (0.41)	0.95 (0.41)	0.95 (0.41)	0.95 (0.41)
	500	0.94 (0.19)	0.95 (0.19)	0.94 (0.18)	0.94 (0.18)
0.05	20	0.94 (0.95)	0.93 (0.94)	0.93 (0.91)	0.93 (0.91)
	40	0.95 (0.74)	0.95 (0.73)	0.93 (0.67)	0.93 (0.67)
	100	0.95 (0.54)	0.95 (0.54)	0.88 (0.43)	0.88 (0.44)
	500	0.96 (0.35)	0.96 (0.35)	0.62 (0.20)	0.61 (0.20)
0.10	20	0.95 (1.15)	0.95 (1.21)	0.91 (0.96)	0.92 (1.01)
	40	0.97 (0.98)	0.97 (1.03)	0.86 (0.70)	0.88 (0.75)
	100	0.98 (0.81)	0.98 (0.86)	0.65 (0.44)	0.69 (0.49)
	500	1.00 (0.63)	1.00 (0.66)	0.05 (0.20)	0.05 (0.22)

Table 2: Percentages of coverage and median length (in parentheses) of naively robust and globally robust confidence intervals. The nominal level is 95%.

Note that the coverage levels achieved by robust confidence intervals are much better than the naive ones. Also note that for small sample sizes, the lenghts of the robust and naive confidence intervals are comparable. This fact can be explained observing that for small n the contribution of the asymptotic bias to the length of the intervals is small compared with that of the variance of the estimate. However, for large n, there is a signficant relative increase in length for the robust confidence intervals. The bias uncertainty incorporated in the robust methods prevents the length of these confidence intervals to converge to zero as $n \to \infty$.

3 Globally robust inference for simple linear regression

Consider the following simple linear regression model:

$$y_i = \beta_0 + \beta_1 (x_i - \mu) + \varepsilon_i, \quad 1 \le i \le n, \tag{30}$$

where μ is a location parameter, x_i and ε_i are independent, $\varepsilon_i \sim F_0$, $\text{Median}_{F_0}(\varepsilon_i) = 0$, $x_i \sim G_0$ and (x_i, y_i) are iid random vectors. The joint distribution of (x_i, y_i) under model (30) will be denoted by H_0. To allow for outliers and other departures from this central parametric model we will assume that the joint distribution H of (x_i, y_i) belongs to the ϵ–contaminated neighborhood

$$\mathcal{F}_\epsilon(H_0) = \{H = (1-\epsilon)H_0 + \epsilon H^* : H^* \text{ any distribution on } \Re^2\}, \quad 0 < \epsilon < 1/2. \tag{31}$$

The presence of contamination in the data may cause robust slope estimates to converge to a value different from β_1 (see (30)), producing *asymptotic bias.* Therefore, if the confidence interval does not account for bias, it will have unacceptably low coverage level for many distributions in the neighborhood.

In the following, to fix ideas, we will concentrate on the slope parameter β_1. Classical asymptotic $(1-\alpha)100\%$ confidence intervals for β_1 are of the form $\tilde\beta_1 \pm z_{\alpha/2}\ \text{SE}(\tilde\beta_1)$ where $\tilde\beta_1$ is the least squares estimate for β_1 and $\text{SE}(\tilde\beta_1)$ is the corresponding estimated standard error of $\tilde\beta_1$. Similarly to the location model, naive robust confidence intervals for β_1 could be obtained using a robust point estimate $\hat\beta_1$, and a robust estimate of its standard error. Again, this naive procedure produces confidence intervals satisfying Part 2 of Definition 1, but not Part 1. To illustrate this problem we generated 600 samples (x_i, y_i) of sizes $n = 20, 40, 60, 80$ and 100 from contaminated normal distributions $(1-\epsilon)N(\mathbf{0}, I) + \epsilon N(\boldsymbol{\mu}, \tau^2 I)$ with $\boldsymbol{\mu}' = (\mu_x, \mu_y)$, $\tau = 0.1$, $\mu_x = 3$ and $\mu_y = 1.5$ (2) for $\epsilon = 0.05$ (0.10). We refer to this case as the "mild contamination case". In the "strong contamination case" we take $\mu_x = 5$ and $\mu_y = 2.5$ for $\epsilon = 0.05$ and $\epsilon = 0.10$. We used Splus to calculate the high breakdown point regression MM–estimates (Yohai, 1987) and their asymptotic standard errors. The nominal confidence level in all the cases is 0.95.

% of Contamination	Sample Size	Mild Contamination	Strong Contamination
5	20	0.88	0.76
	40	0.82	0.57
	60	0.81	0.35
	80	0.62	0.26
	100	0.47	0.23
10	20	0.71	0.50
	40	0.46	0.25
	60	0.27	0.11
	80	0.15	0.07
	100	0.13	0.05

Table 3: Actual coverage levels of traditional confidence intervals for the slope parameter based on high breakdown point regression estimates.

The observed coverage levels (see Table 3) are well below the nominal level, specially for large sample sizes. This is in agreement with our discussion above and indicates that the low coverage levels are mainly due to the bias introduced by the contamination in the data. Notice that in general the variance of $\hat{\beta}_1$ is of order $1/n$ while its bias does not vanish as n goes to infinity.

As before, let $\hat{\beta}_1$ be an asymptotically normal and robust point estimate:

$$\sqrt{n}\left(\hat{\beta}_1 - \hat{\beta}_1(H)\right) \to_d N\left(0, v^2(H)\right).$$

Moreover, assume that $\hat{\beta}_1$ has a sharp and known bias bound $\bar{\beta}$, in other words: $\sup_H \left|\hat{\beta}_1(H) - \beta_1\right| = \bar{\beta}$. To build a $(1-\alpha)$ confidence interval for β_1 based on $\hat{\beta}_1$ we need to find q_n such that

$$P_H\left(\left|\hat{\beta}_1 - \hat{\beta}_1(H)\right| \leq q_n\right) = 1-\alpha.$$

Using the normal approximation to the distribution of $\hat{\beta}_1$ we can obtain an estimated quantile $\tilde{q}_n = \tilde{q}_n(H)$ solving

$$\Phi\left(\frac{\tilde{q}_n(H) - b}{v_n}\right) + \Phi\left(\frac{\tilde{q}_n(H) + b}{v_n}\right) - 1 = 1-\alpha\,, \tag{32}$$

where $\sqrt{n}v_n$ is a consistent estimate of $v^2(H)$ and $b = b(H) = \hat{\beta}_1(H) - \beta_1$. Again, the monotonicity of $\tilde{q}_n$ in b yields a $(1-\alpha)$ robust confidence interval of the form $\hat{\beta}_1 \pm \bar{q}_n$ where $\bar{q}_n$ is given by

$$\Phi\left(\frac{\bar{q}_n - \bar{\beta}}{v_n}\right) + \Phi\left(\frac{\bar{q}_n + \bar{\beta}}{v_n}\right) - 1 = 1-\alpha\,.$$

In principle, any robust estimate of β_1 which is asymptotically normal and has a known maxbias function could be used to construct the above robust confidence intervals. However, it is also clear that the informativeness of the interval will much depend on the size of the maxbias and corresponding bias bound $\bar{\beta}$.

It is easy to see that

$$\bar{\beta}_1 = \sup_{H\in\mathcal{F}_\epsilon(H_0)} |T_1(H) - \beta_1| = \frac{\sigma_\varepsilon}{\sigma_x} B_1(\epsilon),$$

where $B_1(\epsilon) = \sup_{H\in\mathcal{F}_\epsilon(H_0)} \left|\hat{\beta}_1(H) - \beta_1\right| \sigma_x/\sigma_\varepsilon$, and σ_ε and σ_x are the residuals and explanatory variable scales under the central model H_0. As before, following Berrendero and Zamar (2001) an estimate of the above bias bound is

$$\bar{\beta}_{1n} = \frac{\hat{\sigma}_{\varepsilon,n}}{\hat{\sigma}_{x,n}}\ k(\epsilon)\ B_1(\epsilon) = \frac{\hat{\sigma}_{\varepsilon,n}}{\hat{\sigma}_{x,n}}\ B_1^*(\epsilon) \tag{33}$$

where $\hat{\sigma}_{\varepsilon,n}$ and $\hat{\sigma}_{x,n}$ are robust estimates of σ_ε and σ_x, $k(\epsilon) = s^+(\epsilon)/s^-(\epsilon)$, and $s^+(\epsilon)$ and $s^-(\epsilon)$ are the explosion and implosion biases of the scale estimates $\hat{\sigma}_{\varepsilon,n}$ and $\hat{\sigma}_{x,n}$ respectively (see Martin and Zamar (1993)). Some values of $B_1^*(\epsilon)$ for the GMS in the case of contamination neighborhoods of a Gaussian regression model using the shorth for $\hat{\sigma}_{\varepsilon,n}$ and $\hat{\sigma}_{x,n}$ are displayed in column 4 of Table 4.

3.1 Choice of the point estimates

To obtain relatively short robust confidence intervals we need to use point estimates with small bias bounds. Unfortunately, the estimates considered by Berrendero and Zamar (2001) have asymptotic biases that are too large for our purpose. We then turn our attention to the Generalized Median of Slopes (GMS) estimate first proposed by Brown and Mood (1951) and recently studied by Adrover and Zamar (2000). This estimate has considerably smaller maxbias. However there is one limitation: these maximum biases have been derived assuming that the explanatory variable has a symmetric central distribution (for example, a contaminated normal or contaminated double exponential). It is of course of practical interest to be able to surpass this difficulty.

The GMS is defined as the solutions $(\hat{\alpha}_n, \hat{\beta}_n)$ to the equations

$$0 = \frac{1}{n} \sum_{i=1}^{n} \operatorname{sign}\left(y_i - \hat{\alpha}_n - \hat{\beta}_n (x_i - \hat{\mu}_n)\right) \operatorname{sign}(x_i - \hat{\mu}_n), \tag{34}$$

$$0 = \frac{1}{n} \sum_{i=1}^{n} \operatorname{sign}(y_i - \hat{\alpha}_n - \hat{\beta}_n (x_i - \hat{\mu}_n)). \tag{35}$$

where $\hat{\mu}_n = \text{Median}\{x_i\}$. The corresponding regression fit and the vertical line $x = \hat{\mu}_n$ split the plane into four quarters containing equal number of points. Equations (34) and (35) are a natural generalization of the minimax–bias estimate of regression through the origin, Median$\{y_i/x_i\}$, see Martin *et al.* (1989).

The slope estimate $\hat{\beta}_n$ satisfies the fixed point equation

$$g(\hat{\beta}_n) = \text{Median}\left\{ \frac{y_i - \text{Median}\{y_i - \hat{\beta}_n (x_i - \hat{\mu}_n)\}}{x_i - \hat{\mu}_n} \right\} - \hat{\beta}_n = 0. \tag{36}$$

A solution to (36) can be found with an iterative algorithm (for details see Adrover and Zamar, 2000).

Table 4 below (extracted from Adrover and Zamar, 2000) displays the maxbiases of two median-based regression estimates assuming normally distributed explanatory variable and regression error under the central model. The values in the second column (labeled MS) correspond the the minimax bias estimate in the class of affine and regression equivariant estimates, when the intercept is equal to zero (Martin *et al.*, 1989).

ϵ	MS	GMS	$B_1^*(\epsilon)$
0.010	0.014	0.016	0.016
0.025	0.039	0.041	0.044
0.050	0.081	0.088	0.100
0.100	0.174	0.201	0.261
0.150	0.282	0.361	0.547
0.200	0.411	0.639	1.160
0.240	0.538	1.299	2.787
0.250	0.574	∞	∞

Table 4: Maximum bias of the MS and GMS slope estimates, and bias bounds for the GMS slope estimate.

We see that the GMS has a computable and fairly low bias bound. Note that the breakdown point of the GMS is $\epsilon^* = 1/4$. We decided to use GMS for the construction of our robust confidence intervals. The choice was guided by (i) the good bias performance shown by GMS; and (ii) the asymptotic normality of these estimates which can be proved under general conditions (see next Section).

3.1.1 Asymptotic properties of the GMS estimate

When studying the asymptotic distribution of the GMS estimate we introduce a minor modification to its definition to avoid technical problems caused by values of x too close to their location parameter μ. In practice these problems are avoided by restricting attention to points x_i which are not equal to $\hat{\mu}_n$.

The slight modification consists of multiplying the right hand side of (34) by "weights" $w((x_i - \hat{\mu}_n)/s_n)$ where s_n is an S–estimate of the scale of x_i and $w(t) = 1$ for $|t| > d$ (some small positive value) and zero otherwise. In summary, $(\hat{\alpha}_n, \hat{\beta}_n)$ are defined as a solution to

$$0 = \frac{1}{n}\sum_{i=1}^{n} \operatorname{sign}\left(y_i - \hat{\alpha}_n - \hat{\beta}_n(x_i - \hat{\mu}_n)\right) \operatorname{sign}(x_i - \hat{\mu}_n)\, w\left(\frac{x_i - \hat{\mu}_n}{s_n}\right) \tag{37}$$

$$0 = \frac{1}{n}\sum_{i=1}^{n} \operatorname{sign}\left(y_i - \hat{\alpha}_n - \hat{\beta}_n(x_i - \hat{\mu}_n)\right) \tag{38}$$

Adrover and Zamar (2000) proved the consistency of the GMS under certain regularity conditions that include the continuity of the contaminated distribution H. The following theorem shows that the GMS estimate is asymptotically normal for continuous distributions H in a contamination neighborhood of a symmetric distribution H_0.

Theorem 1 (Asymptotic Normality**):** *Suppose that H_0 satisfies:*

A1. F_0 *is a symmetric distribution with unimodal density function* f_0*;*

A2. G_0 *is a symmetric distribution with unimodal density function* g_0*;*

and that the random vector (x, y) *satisfies*

A3. $(x, y) \sim H$ *where* H *is absolutely continuous and belongs to the contamination neighborhood* $\mathcal{F}_\epsilon(H_0)$.

Then,

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_{0,n} - b_0 \\ \hat{\beta}_{1,n} - b_1 \end{pmatrix} \to_d \mathrm{N}(\mathbf{0}, \Sigma) \tag{39}$$

holds with (b_0, b_1) *given by*

$$E_H\left[\operatorname{sign}(y_i - b_0 - b_1(x_i - \hat{\mu}_n)) \operatorname{sign}(x_i - \mu) w\left(\frac{x_i - \mu}{\sigma_x}\right)\right] = 0$$

$$E_H\left[\operatorname{sign}(y_i - b_0 - b_1(x_i - \mu))\right] = 0.$$

While in general $\Sigma = (\sigma_{ij})$ *is quite involved, if* x *and* ε *are independent and* G *and* F *are symmetric, then* $\sigma_{12} = 0$ *(that is* $\hat{\beta}_{0,n}$ *and* $\hat{\beta}_{1,n}$ *are asymptotically independent) and*

$$\sigma_{11} = \frac{1}{4f_0^2(0)} \quad \textit{and} \quad \sigma_{22} = \frac{1}{4f_0^2(0)\mathrm{E}_{G_0}^2\{|x|w(x)\}}. \tag{40}$$

A sketch of the proof of Theorem 2 can be found in the Appendix.

4 Example

The response variable is the difference between monthly returns of Motorola shares and the returns on 30-day US Treasury bills, for the period January 1978 to December 1987. The explanatory variable is the difference between the monthly market return (based on transactions of the New York Stock Exchange and the American Exchange) and the returns on 30-day US Treasury bills, over the same period of time. These data were published in Berndt (1994).

Financial economists fit a straight line to this type of data. The slope measures the riskiness of the stock: the larger the slope the riskier the stock. The estimated slopes using three different estimation methods are given in Table 8. Figure 1 contains a scatter plot of the data and the three fitted lines.

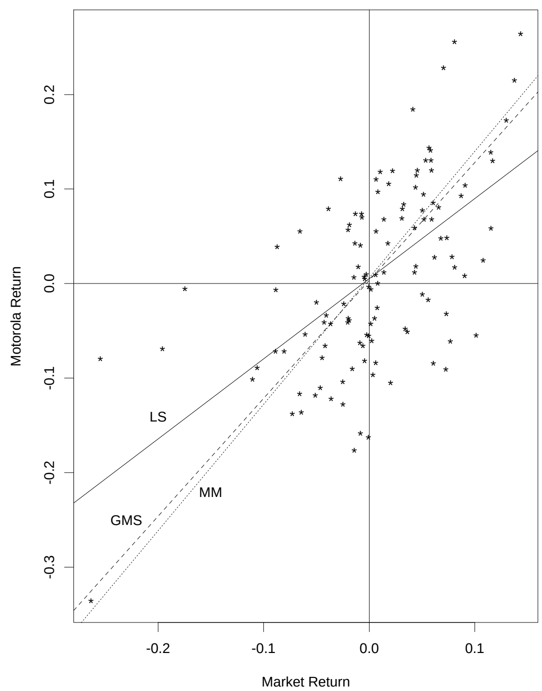

Figure 1: Motorola returns versus market returns. Three different fits.

The plot of the empirical autocorrelation function of the residuals from the GMS shows that there is no significant serial correlation in these data.

The estimates for the intercept parameter are equal to zero (to the second decimal place) for the three methods. On the other hand the slope estimates range from 0.85 (LS) to 1.34 (MM). The conclusions from a financial point of view widely differ depending on the regression method being used. Motorola's stocks are riskier than the market, according to GMS and MM, and safer than the market, according to LS. In summary, it is difficult to determine the risk of Motorola's stocks from these data without the help of an inferential method that takes into account the variability

Method of Estimation	$\hat{\beta}$
GMS	1.25
MM	1.34
LS	0.85

Table 5: Different regression fits based on several robust and non-robust estimates.

and possible bias of the estimates.

One hypothesis of possible interest in this case is

$$H_0 : \beta > 1 \quad \text{versus} \quad H_1 : \beta \leq 1, \tag{41}$$

that is, whether Motorola stocks are a safe investment or not. The conclusions from the analyses using the classical, naively robust and globally robust approaches are:

- The LS approach rejects H_0 at level $\alpha = 0.1$ (the p-value is 0.074) and the corresponding residual plot does not indicate the presence of outliers.
- The naive robust approach using lmRobMM from Splus is indecisive as the "outliers test" is significant and the recommendation is not to perform inference based on the final estimate. If we were to ignore the warning and proceed to test $H_0 : \beta \leq 1 \quad$ versus $\quad H_1 : \beta > 1$ we would then reject H_0 at level $\alpha = 0.1$ (the p-value is 0.095).
- The GMS approach rejects $H_0 : \beta \leq 1 \quad$ in favor of $\quad H_1 : \beta > 1$ at level $\alpha = 0.1$ with p-value 0.089 (see explanation below).

To calculate the robust p-value we need to determine a plausible value for ϵ. The plot of the standardized residuals using the GMS-fit suggests that out of 120 observations there is one clear outlier. It seems reasonable to use $\epsilon = 0.01$. Since $\hat{\sigma}_{\varepsilon,n}/\hat{\sigma}_{x,n} = 1.41$ the bias bound for $\epsilon = 0.01$ is $\hat{\overline{\beta}}_n = 0.0225$. The estimated standard deviation of the GMS slope estimate is $\hat{\sigma}_n = 0.169$, using the Method 3 described below. Therefore, the robust p-value is (see Table 3)

$$\hat{p}^R = 1 - \Phi\left[(0.25 - 0.0225)/0.169\right] = 0.089.$$

Since the procedure is conservative, and the noise-to-signal ratio is big ($\hat{\sigma}_{\varepsilon,n}/\hat{\sigma}_{x,n} = 1.41$), the interpretation of this p-value must be taken accordingly.

5 Numerical results

In this section we present the results of a Monte Carlo study to investigate the finite sample coverage level of robust confidence intervals for the slope of a simple linear regression model. Based on the considerations of the previous section we used the GMS slope estimate.

We present the results for three different choices of v_n, the estimate of the standard error of the GMS estimate in (32). In particular we consider

Method 1 (*Empirical Asymptotic Variance*): The variability is estimated using formula (40) with $w = 1$. By (40) and assuming $F_0 = \mathrm{N}(0, \sigma_\varepsilon^2)$ and $G_0 = \mathrm{N}(\mu, \sigma_x^2)$, we have

$$\mathrm{AV}(\hat{\beta}) = \frac{\pi}{2n} \frac{\sigma_\varepsilon^2}{\mathrm{E}^2 \, |x - \mu|} = \frac{\pi^2}{4n} \frac{\sigma_\varepsilon^2}{\sigma_x^2}.$$

Then,

$$v_n = \frac{\pi}{2\sqrt{n}} \frac{\hat{\sigma}_{\varepsilon,n}}{\hat{\sigma}_{x,n}}. \tag{42}$$

Method 2 (*Classical Bootstrap*): The variability is estimated using the standard deviation of the bootstrap distribution of $\hat{\beta}_n$.

Method 3 (*"Robust" Bootstrap*): The variability is estimated using the shorth of the bootstrap distribution of $\hat{\beta}_n$.

To check the performance of the robust confidence intervals using (33) and (42) we conducted a small Monte Carlo simulation using $m = 1000$ replicates of the following sampling situations: sample sizes $n = 20, 40, 60, 80, 100$ and 200 from contaminated normal distributions $(1-\epsilon)N(\mathbf{0}, I) + \epsilon N(\boldsymbol{\mu}, \tau^2 I)$ with $\boldsymbol{\mu}' = (\mu_x, \mu_y)$, $\tau = 0.1$, $\mu_x = 3$ and $\mu_y = 1.5$ (2) for $\epsilon = 0.05$ (0.10). We refer to this case as the "mild contamination case". For the "medium contamination case" we take $\mu_x = 5$ and $\mu_y = 2.5$ for $\epsilon = 0.05$ and $\epsilon = 0.10$. Finally, we considered a "strong contamination case" with $\mu_x = 5$ and $\mu_y = 15$ for $\epsilon = 0.05$ and $\epsilon = 0.10$. The nominal confidence level in all the cases is 0.95.

The numerical results shown below indicate that although the coverage of the robust confidence intervals are in general fairly good (in comparison with those achieved by the "naive" procedure) they are somewhat affected by the underestimation of the standard deviation of $\hat{\beta}_n$.

We now show the empirical coverage and average lengths of robust confidence intervals when we estimate the variance with the methods described above.

Method 1 Most of the observed coverage levels in Table 6 are between 90% and 95% (nominal level). These results constitute a major improvement in comparison with those of Table 3 (based on the "naive" approach).

Further numerical analysis showed that the undercoverage in Table 6 is mainly due to the underestimation of the variability of the GMS by Method 1.

Method 2 The coverages in Table 7 are now mostly above the nominal 95% level (all of them in the case of strong contamination and most of them in the case of small sample sizes). Again, further numerical analysis shows that the overcoverages and increases in average length observed in Table 7 are mainly due to overestimation of the GMS variability by Method 2.

Method 3 The coverages reported in Table 8 now are mostly below the nominal 95% level (except in the case of strong contamination with $\epsilon = 0.10$). Again, further numerical analysis shows that these few cases of overcoverage are associated with overestimation of the GMS variability by Method 3.

In summary, Method 2 gives confidence intervals with very high coverage level at the expense of average length. This is due to the fact that the bootstrap standard errors overestimate the estimator variability. Methods 1 and 3 are more difficult to compare. For $\epsilon = 0.05$ Method 3

% of Contamination	Sample Size	Mild Contamination	Medium Contamination	Strong Contamination
5%	20	0.90 (1.28)	0.90 (1.30)	0.89 (1.33)
	40	0.92 (0.98)	0.92 (1.00)	0.92 (1.02)
	60	0.91 (0.84)	0.91 (0.85)	0.93 (0.87)
	80	0.91 (0.75)	0.92 (0.77)	0.94 (0.78)
	100	0.90 (0.69)	0.92 (0.71)	0.94 (0.72)
	200	0.93 (055)	0.94 (0.56)	0.95 (0.57)
10%	20	0.90 (1.42)	0.91 (1.46)	0.92 (1.64)
	40	0.91 (1.13)	0.91 (1.18)	0.94 (1.33)
	60	0.91 (1.01)	0.90 (1.09)	0.95 (1.20)
	80	0.89 (0.95)	0.89 (1.01)	0.96 (1.10)
	100	0.89 (0.89)	0.88 (0.96)	0.97 (1.05)
	200	0.89 (0.76)	0.90 (0.84)	0.98 (0.90)

Table 6: Monte Carlo mean coverage and median length (in parenthesis) of robust confidence interval for the slope β using the asymptotic variance formula (40) with F_0 =N$(0, \sigma_\varepsilon^2)$ and G_0 =N(μ, σ_x^2). The parameters μ, σ_x^2 and σ_ε^2 were robustly estimated from the data.

has better coverage levels with a minimal increase in average lengths. However, the situation is reversed in some cases when $\epsilon = 0.10$. Considering the overall performance we are inclined to recommending Method 3.

6 Concluding Remarks

We have proposed globally robust inference methods for location and simple linear regression. In general, the coverage and significance levels of our methods are notably better than other naive methods that ignore the uncertainty due to the bias of the point estimate. We apply the smallest possible adjustment to maintain the nominal coverage and significance level over the entire contamination neighborhood.

Our results are fully satisfactory for the location model. In this case we recommend the use of the optimal M-estimates derived in Fraiman *et al.* (2001). These estimates have good bias and asymptotic variance behaviour.

The situation for the simple linear regression is less satisfactory. We chose to base our inference on the GMS estimate because it nearly minimizes the maximum bias for small fractions of contamination. However, the discontinuous estimating equations used to define this estimate lower its efficiency and, worst of all, yield an asymptotic variance which is very difficult to estimate. Moreover, we need stringent regularity assumptions (in particular, the symmetry of the central distribution of the regressor variables, and the continuity of the contaminated H) which we would very much like to remove. Unfortunately, we could not push our results beyond these limits due to the lack of continuity of the estimating equations defining the GMS. We conjecture that to relax these assumptions other regression estimates might need to be considered. Therefore, robust confidence intervals based on smoother estimating equations may lead to better results despite the somewhat larger bias of the point estimate. This question needs further study.

To implement these methods in either models (location or simple linear regression) the fraction ϵ of contamination in the data needs to be specified. We propose to construct several confidence

% of Contamination	Sample Size	Mild Contamination	Medium Contamination	Strong Contamination
5%	20	0.99 (2.01)	0.99 (2.01)	0.99 (2.01)
	40	0.97 (1.19)	0.97 (1.19)	0.97 (1.27)
	60	0.96 (0.97)	0.96 (0.98)	0.97 (1.03)
	80	0.94 (0.85)	0.94 (0.85)	0.97 (0.89)
	100	0.94 (0.76)	0.95 (0.77)	0.97 (0.81)
	200	0.95 (0.61)	0.96 (0.61)	0.97 (0.62)
10%	20	0.99 (2.00)	0.99 (2.02)	0.99 (2.43)
	40	0.98 (1.34)	0.98 (1.36)	0.99 (1.69)
	60	0.95 (1.17)	0.94 (1.19)	0.99 (1.44)
	80	0.93 (1.08)	0.92 (1.11)	0.99 (1.30)
	100	0.93 (1.01)	0.92 (1.05)	0.99 (1.20)
	200	0.93 (0.87)	0.93 (0.92)	0.99 (1.00)

Table 7: Monte Carlo mean coverage and median length (in parenthesis) of robust confidence interval for the slope β using the standard deviation of the bootstrap distribution of $\hat{\beta}_n$.

intervals and p-values for different (small) fractions of contamination. The largest fraction of contamination for which a test remains significant or a confidence interval remains informative is of particular interest.

The case $p > 1$ is of course of interest and not covered by this work.

7 Appendix

Proof of Lemma 1: First notice that

$$\hat{\sigma}_\varepsilon(H) = \sup\left\{ s : \mathrm{E}_H \chi_a \left(\frac{y - \hat{\alpha}(H) - \hat{\beta}(H)x}{s} \right) > \frac{1}{2} \right\},$$

where $a = \Phi^{-1}(3/4)$, and $\chi_a(u) = 1$, if $|u| > a$, and zero otherwise. Since

$$\frac{1}{2} = (1-\epsilon)\mathrm{E}_\Phi \chi_a \left(\frac{y}{s^-} \right)$$

and

$$\frac{1}{2} \geq (1-\epsilon)\mathrm{E}_\Phi \chi_a \left(\frac{y}{\hat{\sigma}_\varepsilon(H)} \right)$$

we have that $\hat{\sigma}_\varepsilon(H) \geq s^-$. To see that this lower bound is sharp, let $H^0 = (1-\epsilon)H_0 + \epsilon\delta_{(0,0)}$, where $\delta_{(0,0)}$ is a point mass distribution at $(0,0)$. Since, by definition of the GMS estimate it is clear that $\hat{\alpha}(H^0) = \hat{\beta}(H^0) = 0$,

$$\hat{\sigma}_\varepsilon(H^0) = s^-.$$

The following lemma is needed to prove Theorem 3. The proof is straightforward and ommited.

Lemma 2: Let $T_{c,n}$ be a sequence of Huber's M-estimate satisfying the equation

$$\frac{1}{n} \sum \psi_c(z_i - t) = 0,$$

% of Contamination	Sample Size	Mild Contamination	Medium Contamination	Strong Contamination
5%	20	0.94 (1.41)	0.94 (1.42)	0.91 (1.42)
	40	0.92 (1.01)	0.92 (1.01)	0.93 (1.04)
	60	0.92 (0.86)	0.92 (0.86)	0.94 (0.90)
	80	0.92 (0.76)	0.92 (0.77)	0.94 (0.79)
	100	0.92 (0.71)	0.92 (0.71)	0.94 (0.73)
	200	0.95 (057)	0.95 (0.58)	0.96 (0.58)
10%	20	0.95 (1.54)	0.95 (1.56)	0.95 (1.79)
	40	0.89 (1.17)	0.87 (1.18)	0.96 (1.43)
	60	0.87 (1.05)	0.85 (1.08)	0.97 (1.28)
	80	0.86 (1.00)	0.85 (1.04)	0.98 (1.18)
	100	0.87 (0.95)	0.86 (1.00)	0.98 (1.10)
	200	0.91 (0.83)	0.92 (0.89)	0.99 (0.95)

Table 8: Monte Carlo mean coverage and median length (in parenthesis) of robust confidence interval for the slope β using the shorth of the bootstrap distribution of $\hat{\beta}_n$.

where

$$\psi_c(z) = \begin{cases} -1 & \text{if} \quad z \leq -c \\ z/c & \text{if} \quad |z| < c \\ 1 & \text{if} \quad z \geq c. \end{cases}$$

Let $\{c_n\}$ be a sequence of positive numbers and $M_n \equiv \text{Median}\,\{z_i\}$. Then, there exists a sequence $T_{c_n,n} \equiv T_n$ such that

$$|M_n - T_n| \leq c_n$$

Proof of Theorem 2: The proof goes along the following steps:

1- Consider the sequence of "smooth" estimating equations:

$$\begin{aligned} g_n^{(1)}(t,b,u,s) &= \frac{1}{n} \sum \eta_n \left(\frac{y_i - t}{x_i - u} - b \right) w_n \left(\frac{x_i - u}{s} \right) = 0 \\ g_n^{(2)}(t,b,u) &= \frac{1}{n} \sum \eta_n \left(y_i - t - b(x_i - u) \right) = 0 \\ g_n^{(3)}(u) &= \frac{1}{n} \sum \eta_n \left(x_i - u \right) = 0 \end{aligned}$$

where η_n and w_n are continuously twice differentiable approximations to ψ_{c_n} and w.

2- If $\lim_{n\to\infty} nc_n = \infty$ and $\lim_{n\to\infty} \sqrt{n} c_n = 0$, then

$$\begin{aligned} g_n^{(1)}(\hat{\alpha}_n, \hat{\beta}_n, \hat{\mu}_n, \hat{s}_n) &= O_p(c_n) \\ g_n^{(2)}(\hat{\alpha}_n, \hat{\beta}_n, \hat{\mu}_n) &= O_p(c_n) \\ g_n^{(3)}(\hat{\mu}_n) &= O_p(c_n). \end{aligned}$$

3- By 1- and 2- above, and setting $\hat{\alpha} = \hat{\alpha}(H), \hat{\beta} = \hat{\beta}(H), \hat{\mu} = \hat{\mu}(H)$ and $\hat{s} = \hat{s}(H)$,

$$\begin{aligned}
\mathbf{O}_p(c_n) &= \begin{pmatrix} g_n^{(1)}(\hat{\alpha}_n, \hat{\beta}_n, \hat{\mu}_n, \hat{s}_n) \\ g_n^{(2)}(\hat{\alpha}_n, \hat{\beta}_n, \hat{\mu}_n) \\ g_n^{(3)}(\hat{\mu}_n) \end{pmatrix} = \begin{pmatrix} g_n^{(1)}(\hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{s}) + g_{s,n}^{(1)}(\hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{s})(\hat{s}_n - \hat{s}) \\ g_n^{(2)}(\hat{\alpha}, \hat{\beta}, \hat{\mu}) \\ g_n^{(3)}(\hat{\mu}) \end{pmatrix} \\
&+ \begin{pmatrix} g_{t,n}^{(1)} & g_{b,n}^{(1)} & g_{u,n}^{(1)} \\ g_{t,n}^{(2)} & g_{b,n}^{(2)} & g_{u,n}^{(2)} \\ g_{t,n}^{(3)} & g_{b,n}^{(3)} & g_{u,n}^{(3)} \end{pmatrix} \begin{pmatrix} \hat{\alpha}_n - \hat{\alpha} \\ \hat{\beta}_n - \hat{\beta} \\ \hat{\mu}_n - \hat{\mu} \end{pmatrix} + \mathbf{O}_p(c_n) \\
&= \mathbf{Z}_n + \mathbf{C}_n \mathbf{d}_n + \mathbf{O}_p(c_n)
\end{aligned}$$

where $g_{\varsigma,n}^{(i)}$ is the first partial derivative of $g_n^{(i)}$ with respect to $\varsigma = t, b, u, s$.

4- $\mathbf{Z}_n - \mathrm{E}_H(\mathbf{Z}_n) + \mathbf{C}_n \mathbf{C}_n \mathbf{d}_n + \mathbf{O}_p(c_n) = \mathbf{O}_p(c_n)$ and $\sqrt{n}[\mathbf{Z}_n - \mathrm{E}_H(\mathbf{Z}_n)] \to_d N(\mathbf{0}, \mathbf{\Sigma})$.

5- $\mathbf{C}_n - \mathrm{E}_H(\mathbf{C}_n) = \mathbf{o}_p(1)$, and $\mathrm{E}_H(\mathbf{C}_n) \to \mathbf{C} = (c_{i,j})$.

6- Then, $\sqrt{n}\mathbf{d}_n \to_d N(\mathbf{0}, \mathbf{C}^{-1}\mathbf{\Sigma}(\mathbf{C}^{-1})')$.

7- Under the central model H_0,

$$\mathbf{C}^{-1}\mathbf{\Sigma}(\mathbf{C}^{-1})' = \begin{pmatrix} \frac{1}{4f_0^2(0)} & 0 & 0 \\ 0 & \frac{2[1-G_0(d)]}{4f_0^2(0)E_{G_0}^2(|x|w(x))} & 0 \\ 0 & 0 & \frac{1}{4g_0^2(0)} \end{pmatrix},$$

where $d > 0$ is the cutoff value for w.

8 References

Adrover, J.G. and Zamar, R.H. (2000). Bias robustness of three median-based regression estimates. Technical Report No 194, Department of Statistics, University of British Columbia. Canada.

Barnett, B. and Lewis, T. (1994). *Outliers in statistical data.* Wiley & Sons. New York.

Berndt, E.R. (1994). *The practice of econometrics: classic and contemporary.* Addison-Wesley Publishing Company. New York.

Berrendero, J.R. and Zamar, R.H. (2001). Maximum bias curves for robust regression with non-elliptical regressors. *Ann. Statist.*, **29**.

Brown, G.W. and Mood, A. M. (1951). On median tests for linear hypotheses, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Univ. of California Press, Berkeley, 159–166.

Dixon, W. and Tukey, J. (1968). Approximate behavior of the distribution of Winsorized t (trimming/Winsorization). *Technometrics*, **10**, 83-98.

Fraiman, R., Yohai, V.J. and Zamar, R. (2001). Optimal robust M–estimates of location. *Ann. Statist.*, **29**.

Huber, P.J. (1968). Robust confidence limits. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **10**, 269–278.

Huber, P.J. (1970. Studentizing robust estimates, in *Nonparametric Techniques in Statistical Inference,* M. L. Puri, Ed., Cambridge University Press, Cambridge, England.

Martin, R. D., Yohai, V. J. and Zamar, R. H. (1989). Min–max bias robust regression. *Ann. Statist.*, **17**, 1608–1630.

Martin, R. D. and Zamar, R. H. (1993). Bias-robust estimates of scale. *Ann. Statist.*, **21,** 991-1017.

Rousseeuw, P.J and Leroy, A.M. (1987). *Robust regression and outliers detection.* Wiley, New York.

Rousseeuw, P.J and Yohai, V.J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series.* (J. Franke, W. Hardle and D. Martin, eds.) *Lecture Notes in Statistics*, **26**, 256-272. Berlin: Springer-Verlag.

Salibian-Barrera, M. (2000). Contributions to the theory of robust inference. PhD Thesis. Department of Statistics, University of British Columbia.

Salibian-Barrera, M. and Zamar, R.H. (2002). Bootstrapping robust estimates of regression. *Ann. Statist.*, **30**, No. 2, April 2002.

Tukey, J. (1960). A survey of sampling from contaminated distributions, in: *Contributions to Probability and Statistics*, I. Olkin, Ed., Stanford University Press, Stanford.

Tukey, J. and McLaughlin, D. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. *Sankhyã*, **A 25**, 331-352.

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