Robust and Explicit Estimators for Weibull Parameters

Kris Boudt[a,b] Derya Caliskan[a,c] Christophe Croux[a,*]

[a] *Faculty of business and economics, K.U.Leuven, Belgium*
[b] *Department of business studies, Lessius University College, Belgium*
[c] *Department of statistics, Hacettepe University, Turkey*

Abstract

The Weibull distribution plays a central role in modeling duration data. Its maximum likelihood estimator is very sensitive to outliers. We propose three robust and explicit Weibull parameter estimators: the quantile least squares, the repeated median and the median/Q_n estimator. We derive their breakdown point, influence function, asymptotic variance and study their finite sample properties in a Monte Carlo study. The methods are illustrated on real lifetime data affected by a recording error.

Keywords: Breakdown point, Influence function, Outliers, Robustness, Weibull distribution.

*Corresponding author. E-mail: christophe.croux@econ.kuleuven.be

1 Introduction

The Weibull distribution plays a central role in lifetime models in medical and biological sciences as well as in engineering. If the data are contaminated with outliers, the maximum likelihood estimator can be very unreliable (see e.g. Adatia and Chan 1982; Shier and Lawrence 1984; Seki and Yokoyama 1996 and He and Fung 1999). Several robust alternatives have been proposed in the literature. Lingappaiah (1976) and Dixit (1994) propose a Bayesian approach to handle these outliers. Their estimation methods assume however that the number of outliers and their distribution family is known. In practice, this is never the case. Robust M-type estimators of the Weibull parameters have been studied and among them the method of medians estimator of He and Fung (1999). This estimator has attractive robustness and efficiency properties, but is not explicit.

We propose three robust Weibull parameter estimators that are an explicit function of the data and easy to calculate: the quantile least squares, the repeated median and the median/Q_n estimator. We derive their breakdown point, influence function and asymptotic variance, and study their finite sample properties in a Monte Carlo study. We also compute these robustness and efficiency measures for the quantile estimator proposed by Marks (2005) and for the median/MAD estimator of Olive (2006).

The remainder of the paper is organized as follows. In Section 2 we describe the proposed robust and explicit estimators of the Weibull parameters. In Sections 3 and 4 we derive their influence function and efficiency. The simulation study in Section 5 and the empirical application on lifetime data in Section 6 further document the robustness of the proposed estimators against outlier contamination. Section 7 concludes.

2 Estimators

The main theme of the paper is the robust estimation of the parameters λ and β of the Weibull density function

$$f_{\lambda,\beta}(x) = \frac{\beta}{\lambda}\,(x/\lambda)^{\beta-1}\exp[-(x/\lambda)^{\beta}],$$

where $x, \lambda, \beta > 0$. Since $f(x;\beta,\lambda) = \frac{1}{\lambda}f(\frac{x}{\lambda};\beta,1)$, the parameter λ is called a scale parameter. The Weibull cumulative distribution function equals

$$F_{\lambda,\beta}(x) = 1 - \exp[-(x/\lambda)^{\beta}].$$

The parameter β is the shape parameter. When $\beta = 1$ the Weibull distribution becomes an exponential distribution. For $\beta = 3.4$ the Weibull distribution appears similar to a normal distribution.

It is standard to use the Maximum Likelihood (ML) method for the estimation of the Weibull parameters. A robust and rather efficient alternative for the ML estimator is the method of medians proposed by He and Fung (1999). These estimators solve

$$\begin{aligned} \text{median}_i\{(1-(x_i/\hat{\lambda})^{\hat{\beta}})\log(x_i/\hat{\lambda})^{\hat{\beta}}\} &= c \\ \hat{\lambda} &= \text{median}_i\{x_i\}/(\log 2)^{1/\hat{\beta}}, \end{aligned}$$

where $c = \text{median}((1-Y)\log Y) \approx -0.51$ and Y has an exponential distribution with mean one. The solution to this system of equations requires to use iterative methods.

Like the method of medians, the estimators we propose are robust to outliers, but they have the additional advantage of being an explicit function of the data. To characterize the robustness of the proposed estimators, we derive their asymptotic breakdown point, defined as the smallest proportion of observations (for $n \to \infty$) that needs to be replaced with arbitrary values in order for the estimation of λ or β to be arbitrarily close to zero (implosion) or infinity (explosion). The breakdown point of the ML estimator is $1/n \to 0$. The method of medians has a 50% breakdown point (see He and Fung 1999). As a second robustness measure we consider in Section 3 the influence function which quantifies the effect of small contaminations on the estimator. The ML estimator has un unbounded influence function, while the influence function of the method of medians shape and scale estimators is bounded.

The proposed estimators all have a high breakdown point and bounded influence function. They are based on the quantiles of the log-transformed observations from the Weibull distribution. The α-quantile of a log-Weibull random variable is given by

$$G_{\lambda,\beta}^{-1}(\alpha) = \beta^{-1} \log(-\log(1-\alpha)) + \log \lambda.$$

Let $G^{-1}(\alpha) = G_{1,1}^{-1}(\alpha)$. We have the following linear relationship between the quantiles of the general and standard log-Weibull distribution

$$G_{\lambda,\beta}^{-1}(\alpha) = \beta^{-1} G^{-1}(\alpha) + \log \lambda. \tag{2.1}$$

Note that the log-Weibull distributions $G_{\lambda,\beta}$ form a location-scale family with location parameter $\mu = \log \lambda$ and scale $\sigma = 1/\beta$:

$$G_{\lambda,\beta}(\log x) = F_{\lambda,\beta}(\exp(x)) = G\left((\log x - \mu)/\sigma\right), \tag{2.2}$$

for all $x > 0$. A log-Weibull random variable can thus always be rewritten as

$\log X = \mu + \sigma U$, with U a random variable having distribution function G and density $g(u) = \exp(u - \exp(u))$.

2.1 Quantile estimator

Denote $\hat{q}_\alpha$ the empirical α-quantile of the observations $x_1, \ldots, x_n$. As noted by Marks (2005), it follows from (2.1) that the difference of the logs of any two Weibull quantiles q_{α_2} and q_{α_1} ($0 < \alpha_1 < \alpha_2 < 1$) depends only on the shape parameter β. Replacing the theoretical quantiles $G^{-1}_{\lambda,\beta}(\alpha_1)$ and $G^{-1}_{\lambda,\beta}(\alpha_1)$ in (2.1) by the corresponding empirical quantiles $\log \hat{q}_{\alpha_1}$ and $\log \hat{q}_{\alpha_2}$ yields the following so-called *quantile estimator* of shape

$$\hat{\beta}_{\mathrm{Q}} = \frac{G^{-1}(\alpha_2) - G^{-1}(\alpha_1)}{\log \hat{q}_{\alpha_2} - \log \hat{q}_{\alpha_1}}. \tag{2.3}$$

The corresponding scale estimator is then obtained by plugging the quantile estimator for β in (2.1). After some algebra, this yields the following estimate for the scale parameter

$$\hat{\lambda}_{\mathrm{Q}} = \hat{q}_\alpha / [-\log(1-\alpha)]^{1/\hat{\beta}_{\mathrm{Q}}}, \tag{2.4}$$

for any $0 < \alpha < 1$. In Appendix A we prove the following results regarding the breakdown point of these quantile-estimators. We assume that all observations are distinct.

Proposition 1 *The asymptotic breakdown point of the quantile estimator of shape $\hat{\beta}_Q$ equals*

$$\min(\alpha_2 - \alpha_1, 1 - \alpha_2, \alpha_1).$$

The highest breakdown point possible for this estimator is $1/3$ and is achieved for $\alpha_1 = 1/3$ and $\alpha_2 = 2/3$.

In the remainder of the paper, we take the optimal values $\alpha_1 = 1/3$ and $\alpha_2 = 2/3$.

Proposition 2 *The asymptotic breakdown point of the quantile estimator of scale $\hat{\lambda}_Q$, using the quantile estimator of shape $\hat{\beta}_Q$ with $\alpha_1 = 1 - \alpha_2 = 1/3$, equals*

$$\begin{cases} \min(\alpha, 1-\alpha, 1/3) & \text{for } \alpha \neq 1 - e^{-1} \\ \min(\alpha, 1-\alpha) & \text{for } \alpha = 1 - e^{-1}. \end{cases}$$

In the sequel, we take $\alpha = 0.5$ yielding an overall breakdown point of 1/3.

2.2 Quantile least squares and repeated median estimators

The quantiles of the general log-Weibull distribution in (2.1) are linearly related to the quantiles of the standard log-Weibull distribution, with intercept $b_0 = \log \lambda$ and slope $b_1 = 1/\beta$. Estimates for the Weibull parameters are thus given by a robust fit of the empirical quantiles against the corresponding quantiles of the standard log-Weibull distribution. Replacing the theoretical quantiles with their empirical counterparts in (2.1) yields a linear regression equation

$$y_i = b_0 + b_1 z_i + \varepsilon_i,$$

where $y_i = \log \hat{q}_{i/(n+1)}$ and $z_i = G^{-1}(i/(n+1))$. We consider two robust and explicit regression estimators for b_1 and b_0: the *Quantile Least Squares* (QLS) and the *Repeated Median* (RM) estimators. The corresponding estimates of scale and shape of the Weibull distribution are then directly given by

$$\hat{\lambda} = \exp(\hat{b}_0) \quad \text{and} \quad \hat{\beta} = 1/\hat{b}_1. \tag{2.5}$$

A similar regression based approach to estimation of the Weibull parameters was taken by other authors, e.g. Shier and Lawrence (1984) and Li (1994). They provided simulation-based evidence for the performance of different types of regression estimators, but did not develop any formal robustness study, and neither computed the asymptotic variance of the estimators.

The QLS estimator minimizes a weighted sum of residuals, whereby the observations for which the y_i's that are more extreme than the $\overline{\alpha}$ and $1-\overline{\alpha}$ empirical quantile receive a zero weight

$$\hat{b}_1 = \frac{\tilde{n}\sum_{j=\lfloor\overline{\alpha}n\rfloor+1}^{n-\lfloor\overline{\alpha}n\rfloor} z_j y_j - \sum_{j=\lfloor\overline{\alpha}n\rfloor+1}^{n-\lfloor\overline{\alpha}n\rfloor} z_j \sum_{j=\lfloor\overline{\alpha}n\rfloor+1}^{n-\lfloor\overline{\alpha}n\rfloor} y_j}{\tilde{n}\sum_{j=\lfloor\overline{\alpha}n\rfloor+1}^{n-\lfloor\overline{\alpha}n\rfloor} z_j^2 - (\sum_{j=\lfloor\overline{\alpha}n\rfloor+1}^{n-\lfloor\overline{\alpha}n\rfloor} z_j)^2} \tag{2.6}$$

$$\hat{b}_0 = \frac{1}{\tilde{n}}\sum_{j=\lfloor\overline{\alpha}n\rfloor+1}^{n-\lfloor\overline{\alpha}n\rfloor} y_i - \frac{1}{\tilde{n}}\hat{b}_1 \sum_{j=\lfloor\overline{\alpha}n\rfloor+1}^{n-\lfloor\overline{\alpha}n\rfloor} z_j, \tag{2.7}$$

where $0<\overline{\alpha}<1/2$ and $\tilde{n}=n-2\lfloor\overline{\alpha}n\rfloor$. The higher $\overline{\alpha}$, the more robust the estimator is to outliers. Clearly the OLS estimator (QLS with $\overline{\alpha}=0$) is not robust. Note from (2.5) that the scale estimator $\hat{\lambda}$ tends to zero or infinity if and only if $\hat{b}_0$ tends to $+\infty$ or $-\infty$. Similarly, the shape estimator $\hat{\beta}$ implodes or explodes if and only if $\hat{b}_1$ tends to $+\infty$ or zero. We have then the following result for the breakdown point of the QLS estimator, using similar arguments as for Proposition 1.

Proposition 3 *The asymptotic breakdown point of the QLS shape and scale estimators equals* $\min(\overline{\alpha}, 1-2\overline{\alpha})$. *The highest breakdown point possible for this estimator is* $1/3$ *and is obtained for* $\overline{\alpha}=1/3$.

In the remainder of the paper, we use the QLS estimator with $\overline{\alpha}=1/3$. In Section 4 we show that the QLS estimator has a relatively low efficiency. Therefore we also consider the repeated median estimator introduced by Siegel (1982). Let $med_i(z_i) = \text{median}(z_1,\ldots,z_n)$. The repeated median slope and intercept estimates equal

$$\hat{b}_1 = \underset{j}{med}\,\underset{i\neq j}{med}\,\frac{y_j-y_i}{z_j-z_i} \quad\text{and}\quad \hat{b}_0 = \underset{j}{med}\,\underset{i\neq j}{med}\,\frac{z_j y_i - z_i y_j}{z_j-z_i}. \tag{2.8}$$

Note that the slopes $(y_j - y_i)/(z_j - z_i)$ are always positive, hence $\hat{\beta} \geq 0$. Siegel (1982) showed that the asymptotic breakdown point of $\hat{b}_1$ and $\hat{b}_0$, defined as the smallest proportion of data one needs to replace to let the regression estimator tend to $\pm\infty$, equals 50%. Hence the breakdown point of the scale estimator $\hat{\lambda}$ equals also 50%. The shape estimator $\hat{\beta}$ explodes if $\hat{b}_1$ tends to zero. Since the z_i values are fixed, this can only happen when half of the y_i observations coincide. For this 50% of contamination is needed. We can thus conclude that the RM estimators $\hat{\lambda}$ and $\hat{\beta}$ inherit the 50% breakdown property of the RM regression estimators.

2.3 Median/MAD and median/Q_n location-scale estimators

The log-Weibull distribution belongs to a location-scale family with location $\mu = \log \lambda$ and scale $\sigma = 1/\beta$, see (2.2). Estimation of Weibull parameters can thus be seen as an estimation problem of the location and scale of the observations $\log x_1, \ldots, \log x_n$. Note that the asymptotic breakdown point of the scale and shape estimators $\hat{\lambda} = \exp(\hat{\mu})$ and $\hat{\beta} = 1/\hat{\sigma}$ equals the one of the location and scale estimators $\hat{\mu}$ and $\hat{\sigma}$. Standard location and scale estimators with 50% breakdown point are the median and median absolute deviation

$$
\begin{aligned}
\hat{\sigma} &= 1.3037 \mathrm{med}_j |\log x_j - \mathrm{med}_i \log x_i| && (2.9) \\
\hat{\mu} &= \mathrm{med}_i \log x_i - \hat{\sigma} \log \log 2. && (2.10)
\end{aligned}
$$

This estimator, called the *median/MAD* estimator, was considered by Olive (2006), who presents the correction factors making these estimators consistent, but does not derive the influence function and asymptotic variance of these estimators. As a more efficient alternative with the same breakdown point of 50%, we recommend to estimate σ using the Q_n scale-estimator proposed by Rousseeuw and Croux (1993).

It is given by

$$\hat{\sigma} \quad = \quad 1.9577\{|\log x_i - \log x_j|; i < j\}_{(l)}, \tag{2.11}$$

where the last part is the lth ordered value among this set of $\binom{n}{2}$ differences, where $l = \binom{h}{2} \approx \binom{n}{2}/4$ with $h = \lfloor n/2 \rfloor + 1$. The correction factor 1.9577 ensures consistency. It equals the inverse of the 1/4 quantile of the distribution of the absolute difference between two log-Weibull random variables. The corresponding estimators for the scale and shape parameters are called the *median/Q_n* estimators.

3 Influence function

In Section 2 it is shown that the proposed estimators have a high breakdown point. In this Section we derive their influence function (IF) and show that it is bounded. Hence, the proposed estimators are B- (or bias) robust, which means that their influence function is bounded. The IF is based on the representation of the estimator as a functional T of the empirical distribution function. The IF of the functional T at the distribution F measures the effect on T of adding a small probability mass to the point x_0, standardized by the mass of the contamination. If we denote the point mass distribution at x_0 by Δ_{x_0} and consider the contaminated distribution $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{x_0}$, then the influence function is given by

$$IF(x_0; T, F) = \lim_{\varepsilon \to 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}$$

(see Hampel et al. 1986). A desirable robustness property for an estimator is that it has a bounded IF, if not the estimator can be severely distorted by a small proportion of outliers.

In Appendix B we derive and present expressions for the influence functions of all estimators considered in this paper. They are pictured in Figures 1 and 2 for the case of $\beta = 1$ and $\lambda = 1$. We find that the influence functions of the maximum likelihood and ordinary least squares estimators are unbounded functions of x_0. They converge to $\pm$ infinity as x_0 moves towards zero or infinity. The influence functions of all other estimators considered in the paper are bounded. Note that the influence functions of the quantile, method of medians and median/MAD shape and scale estimators are step functions. The influence functions of the repeated median shape and scale estimators and of the median/Q_n shape estimator are smooth. The influence function of the median/Q_n scale estimator has a discontinuity because of the discontinuity in the influence function of the median.

4 Statistical efficiency

The proposed estimators have a bounded influence function and a high breakdown point. Here we present their asymptotic and finite-sample variance. Let $\theta = (\lambda, \beta)'$. The asymptotic covariance matrix of the maximum likelihood estimator is the inverse of the Fisher information matrix given by

$$I_\theta^{-1} = \begin{pmatrix} 1.109(\lambda/\beta)^2 & -0.257\lambda \\ -0.257\lambda & 0.608\beta^2 \end{pmatrix}. \tag{4.1}$$

For regular asymptotically normal estimators, the asymptotic covariance matrix can be computed as the expectation of the outer product of the influence functions

$$\mathrm{ASV}_{F_{\lambda,\beta}}(\hat{\theta}) = E_{F_{\lambda,\beta}}[IF(x; \hat{\theta}, F_{\lambda,\beta})IF'(x; \hat{\theta}, F_{\lambda,\beta})]. \tag{4.2}$$

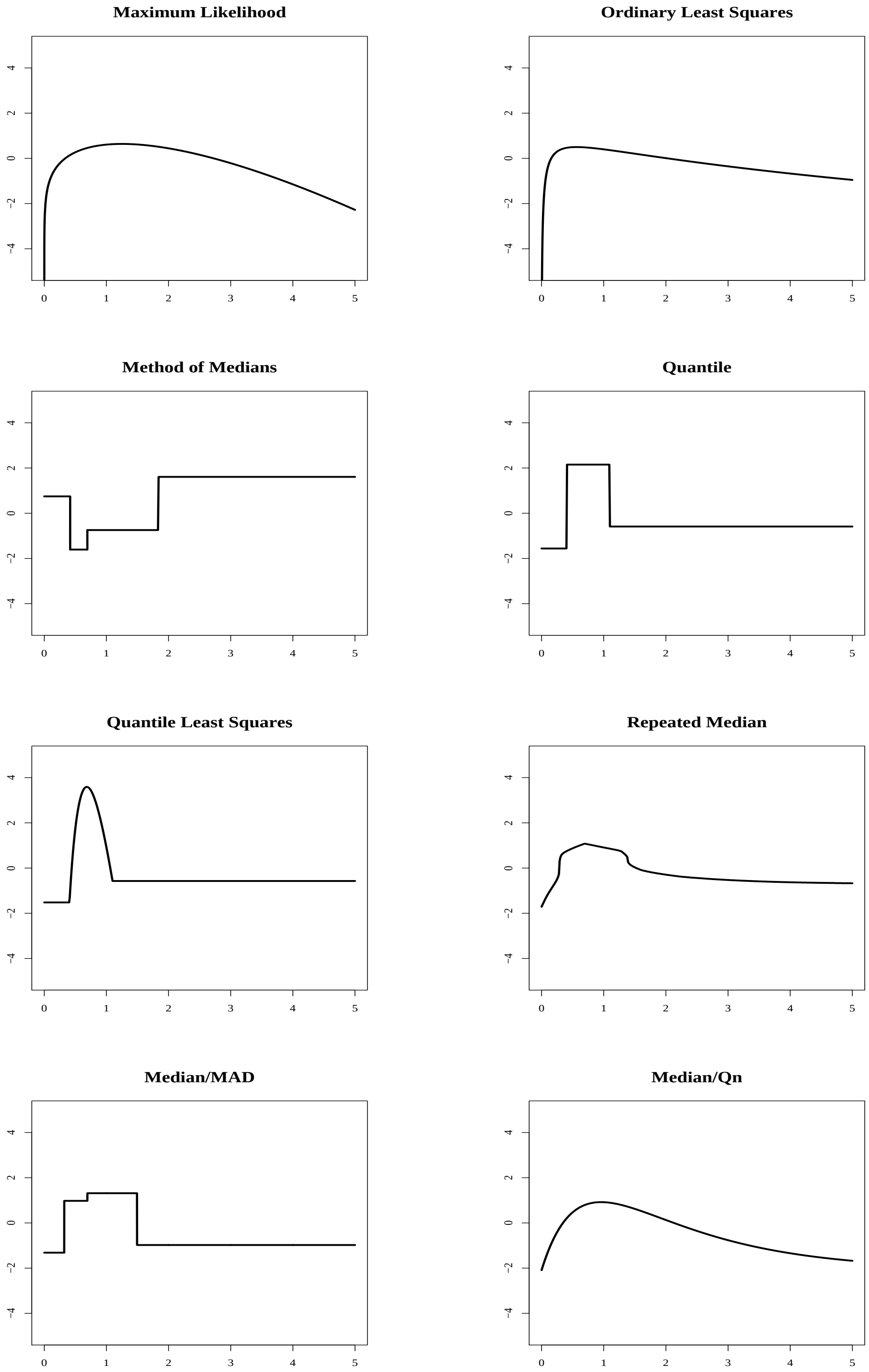


Figure 1: Influence function of the Weibull shape parameter estimators

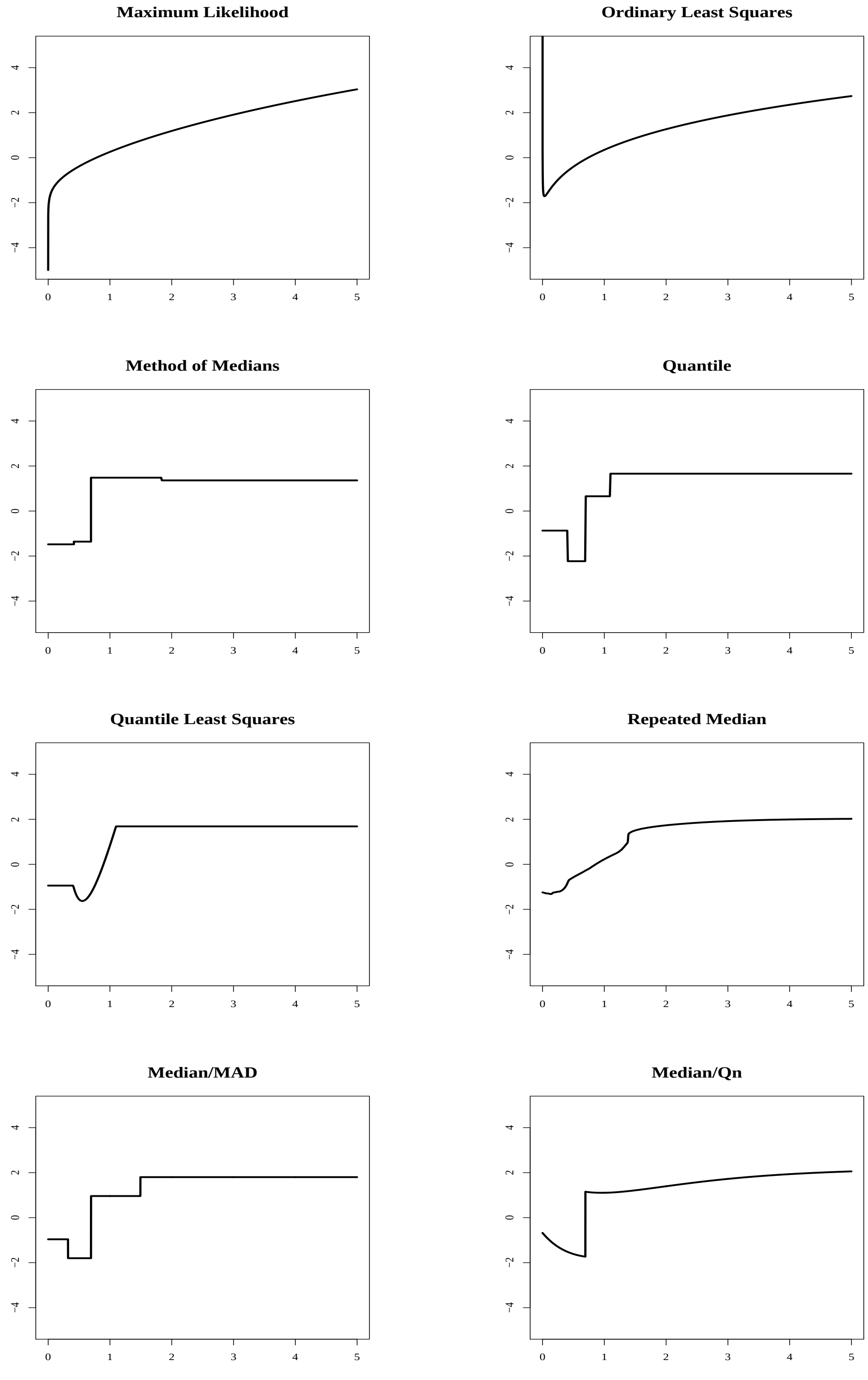


Figure 2: Influence function of the Weibull scale parameter estimators

The quantile and QLS estimators can be written as L-estimators, for which validity of (4.2) has been shown. Asymptotic normality of the the median, MAD and Q_n is well established (see e.g. Hampel et al. 1986; Rousseeuw and Croux 1993). For the method of medians we use the result of He and Fung (1999). For the repeated median estimators we claim that the limiting distribution is not normal and thus (4.2) cannot be used. To compute the asymptotic variance for these estimators, we use the well known result that

$$\sqrt{n}g(G^{-1}(\alpha))(G_n^{-1}(\alpha) - G^{-1}(\alpha)) \xrightarrow{d} B(\alpha),$$

where $G_n^{-1}(\alpha)$ is the empirical quantile process of the log-Weibull observations $\log x_1, \ldots, \log x_n$ and $\{B(\alpha); 0 \leq \alpha \leq 1\}$ is the standard Brownian bridge on $[0, 1]$. For the repeated median slope estimator, we thus have that $\sqrt{n}(\hat{b}_1 - 1)$ converges in distribution to

$$\mathrm{med}_{\alpha_1}\mathrm{med}_{\alpha_2}\frac{1}{G^{-1}(\alpha_1) - G^{-1}(\alpha_2)}\left(\frac{B(\alpha_1)}{g(G^{-1}(\alpha_1))} - \frac{B(\alpha_2)}{g(G^{-1}(\alpha_2))}\right).$$

A similar expression holds for the shape. Using these limiting distributions, that turn out to be very close to a normal distribution, we obtain by numerical methods that the asymptotic variance of the repeated median shape and scale estimators equal approximately 0.85 and 1.41, respectively. In Table 1 we report the asymptotic variance ($n = \infty$) of the estimators as well as their finite-sample counterparts for $n = 20, 100$ and 500, for $\lambda = \beta = 1$. This is without loss of generality, since

$$\mathrm{ASV}_{F_{\lambda,\beta}}(\hat{\beta}) = \beta^2\mathrm{ASV}_F(\hat{\beta}) \quad \text{and} \quad \mathrm{ASV}_{F_{\lambda,\beta}}(\hat{\lambda}) = (\lambda/\beta)^2\mathrm{ASV}_F(\hat{\lambda}),$$

for any $\lambda, \beta > 0$. The finite-sample variances are obtained by simulation (10,000 replications) and are multiplied by the sample size n. As can be seen from Table 1, the (standardized) finite sample variances converge quite well to their asymptotic counterpart.

For the estimation of the shape parameter, we find that the maximum likelihood estimator has, as expected, the lowest variance for all sample sizes, but the proposed repeated median and median/Q_n estimators are a good second best. Their asymptotic efficiency, with respect to the ML estimator, is 71.5% and 82.2%, respectively. This is significantly higher than the 55.3% of the least squares estimator and the 42.2% of the method of medians. The median/Q_n estimator is almost twice as efficient as the median/MAD estimator. The quantile and quantile least squares estimators have the lowest efficiency (around 20%).

For the scale estimation, we find that the least squares estimator is almost as efficient as the maximum likelihood estimator. In contrast with the median/Q_n shape estimator, the median/Q_n scale estimator has a rather low efficiency. For all sample sizes, the repeated median estimator is the most efficient of all robust estimators.

5 Simulation study

In this Section we evaluate the effect of outliers on the accuracy of the conventional and proposed robust estimators by means of a Monte Carlo simulation. The reference distribution is the Weibull distribution with parameters $(\lambda_0 = 1, \beta_0 = 1)$. Like He and Fung (1999) we consider the case of no outliers, the case of 10% replacement outliers coming from another Weibull distribution with either a different scale parameter $(\lambda_1 = 0.2)$ or a different shape parameter $(\beta_1 = 0.5)$ and the case of 10%

Table 1: Finite sample and asymptotic ($n = \infty$) variance of several Weibull shape and scale parameter estimators: ML, OLS, Method of Medians (MoM), Quantile (Quan), Repeated Median (RM), median/MAD (MAD) and median/Q_n (Q_n) estimators.

n	ML	OLS	MoM	Quan	QLS	RM	MAD	Q_n
Shape								
20	0.83	0.88	2.16	4.13	4.20	1.25	2.33	1.07
100	0.65	0.96	1.55	2.74	3.22	0.92	1.64	0.75
500	0.61	1.03	1.45	2.52	3.04	0.89	1.53	0.70
∞	0.61	1.10	1.44	2.47	2.97	0.85	1.38	0.74
Scale								
20	1.11	1.17	1.64	2.12	2.11	1.28	1.70	1.79
100	1.12	1.18	1.76	2.11	1.84	1.30	1.80	1.84
500	1.11	1.16	1.75	2.08	1.75	1.28	1.79	1.83
∞	1.11	1.17	1.76	2.07	1.73	1.41	1.97	1.79

replacement outliers from a uniform distribution on $[0, 20]$. Hence we allow that some observations come from a different Weibull population and, in the last model, we allow for the occurrence of errors. We generate samples of size $n = 100$ according to different sampling schemes and compute for each sample the scale estimate $\hat{\lambda}_i$ and shape estimate $\hat{\beta}_i$, for $i = 1, \ldots, n$. For each simulation setting and each type of estimator, we compute the root mean squared error

$$RMSE_\beta = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_i - \beta_0)^2}\ ,\ RMSE_\lambda = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{\lambda}_i - \lambda_0)^2}.$$

The results are reported in Table 2. The conclusions from the study are as follows.

(i) When there is no contamination, the ML estimator performs the best, as expected. The median/Q_n and repeated median are more efficient than the OLS estimator and the method of medians.

(ii) Outlier contamination causes a large increase in the RMSE of the ML and OLS estimators and a much smaller increase in the RMSE of the robust alternatives.

Table 2: Root mean squared error of Weibull shape and scale parameters estimators for samples of size $n = 100$ for different sampling schemes. The same estimators as in Table 1 are considered.

n	ML	OLS	MoM	Quan	QLS	RM	MAD	Q_n
No contamination								
β	0.08	0.10	0.13	0.17	0.18	0.10	0.13	0.09
λ	0.11	0.11	0.13	0.15	0.14	0.11	0.13	0.14
10% Contamination from Weibull($\lambda_1 = 0.2$,$\beta_1 = \beta_0$)								
β	0.17	0.13	0.12	0.16	0.17	0.11	0.12	0.13
λ	0.28	0.26	0.21	0.22	0.21	0.21	0.21	0.24
10% Contamination from Weibull($\lambda_1 = 1$,$\beta_1 = 0.5$)								
β	0.14	0.17	0.12	0.16	0.17	0.11	0.13	0.12
λ	0.12	0.13	0.14	0.15	0.14	0.12	0.14	0.15
10% Contamination from U(0,20)								
β	0.28	0.17	0.14	0.16	0.17	0.13	0.13	0.18
λ	0.51	0.45	0.27	0.26	0.25	0.27	0.26	0.31

(iii) For the shape estimation, the quantile and quantile least squares estimators have the worst performance of all considered robust estimators in all scenarios.

(iv) For the scale estimation, only the quantile estimator has a higher RMSE than the median/Q_n estimator in the case of no outliers. In the presence of outliers, the median/Q_n scale estimator has the highest RMSE of all robust estimators considered.

(v) For both the estimation of shape and scale, the repeated median estimator has a lower RMSE than the method of medians in all cases considered.

On the basis of its 50% breakdown point, bounded influence function and high efficiency, and also because of its high robustness to outliers as shown in this simulation study, we recommend the repeated median estimator. This estimator has the best robustness/efficiency trade-off of all robust estimators considered.

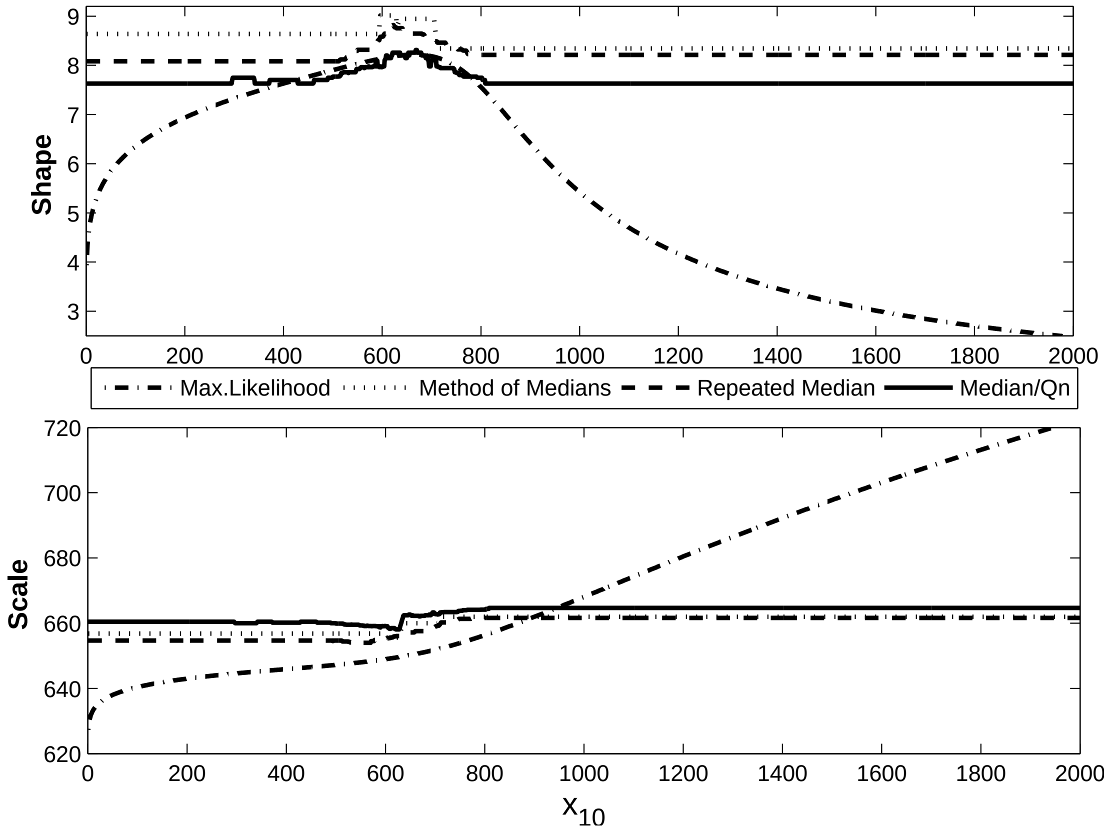
6 Empirical application

Here we illustrate the sensitivity of the maximum likelihood, method of medians, repeated median and median/Q_n estimators to the value of one single observation. We consider a sample of 38 lifetime observations of male mice who had received a radiation dose of 300 rads at age 5-6 weeks: 317,318, 399, 495, 525, 536, 549, 552, 554, **337**, 558, 571, 586, 594, 596, 605, 612, 621, 628, 631, 636, 643, 647, 648, 649,661, 663, 666, 670, 695, 697, 700, 705, 712, 713, 738, 748, 753 days. These data were originally reported in Hoel (1972) and republished in Kalbfleisch and Prentice (1980). He and Fung (1999) discovered a recording error for the 10th observation in the sample: it should be a lifetime of 557 instead of 337 days. In Figure 3 we plot the estimated shape and scale parameters for the same sample but where we replace the tenth observations x_{10} by a range of values between 1 and 2000. Since we know that x_{10} is a recording error, it is desirable that for all values of x_{10} the estimated shape and scale are close to the value obtained for $x_{10} = 337$. We see that changing the value of the single observation x_{10} has little influence on the method of median, repeated median and median/Q_n estimators, but induces a large variation in the maximum likelihood estimates. This sensitivity analysis illustrates that robust methods have a built-in protection against recording errors.

7 Conclusion

In this paper we propose explicit and highly robust estimators of the Weibull parameters and derive their breakdown point, influence function and asymptotic variance. Of all considered estimators, the repeated median and the median/Q_n estimator perform best, yielding a good trade-off between robustness and efficiency. We have a preference for the repeated median since it has a 50% breakdown point, a bounded

Figure 3: Estimate of shape and scale for the Hoel data, where the tenth observation is replaced by $x_{10} = 0, \ldots, 2000$



and continuous influence function, high efficiency, and, as shown in the simulation study, it remains very accurate in the presence of outliers, for both the estimation of shape and scale.

The quantile and quantile least squares estimator are less attractive from an efficiency/robustness trade-off point of view, but they have the appealing property of producing the same estimate in the absence and presence of up to 33% of left and right censoring. Censoring is a typical feature of lifetime data. A topic for further research is to compare the performance of the proposed estimators in presence of censoring with the maximum likelihood estimators of e.g. Cohen (1965) and Muralidharan and Lathika (2006).

A Asymptotic breakdown point

Proof of Proposition 1. The shape estimator $\hat{\beta}_{\mathrm{Q}}$ in (2.3) explodes for $\hat{q}_{\alpha_2} = \hat{q}_{\alpha_1}$. This occurs if a proportion $(\alpha_2 - \alpha_1)$ of the observations is placed to the same position as $\hat{q}_{\alpha_1}$. The estimator $\hat{\beta}_{\mathrm{Q}}$ implodes if $\hat{q}_{\alpha_2} \to \infty$ and $\hat{q}_{\alpha_1}$ remains bounded or if $\hat{q}_{\alpha_1} \to 0$ and $\hat{q}_{\alpha_2}$ remains bounded. For this, it suffices to place $1-\alpha_2$ observations to ∞ or α_1 observations to zero. The breakdown point of $\hat{\beta}_{\mathrm{Q}}$ is thus $\varepsilon(\alpha_1, \alpha_2) \equiv \min(\alpha_2 - \alpha_1, 1 - \alpha_2, \alpha_1)$. Given α_1, the highest value of $\varepsilon(\alpha_1, \alpha_2)$ is obtained at the intersection of the lines $\varepsilon = \alpha_2 - \alpha_1$ and $\varepsilon = 1 - \alpha_2$, i.e. for $\alpha_2 = (1 + \alpha_1)/2$. We further have that the maximum of $\varepsilon(\alpha_1, (1 + \alpha_1)/2)$ is $1/3$ for $\alpha_1 = 1/3$. Since given α_1, the highest breakdown point is obtained for $\alpha_2 = (1 + \alpha_1)/2$, the quantile estimator of shape has thus maximum breakdown point for $\alpha_1 = 1/3$ and $\alpha_2 = 2/3$.

Proof of Proposition 2. We need to distinguish three cases. First assume that $0 < \alpha < 1 - e^{-1}$. Then we have that $-\log(1 - \alpha) < 1$ and the denominator of the scale estimator $\hat{\lambda}_{\mathrm{Q}}$ in (2.4) will be finite for every possible value of $\hat{\beta}_{\mathrm{Q}}$. Implosion of scale is then only possible if one replaces more than a proportion α of the data by zero. Explosion of $\hat{\lambda}_{\mathrm{Q}}$ can arise when more than a fraction $(1 - \alpha)$ of the data are placed to infinity or if $\hat{\beta}_{\mathrm{Q}}$ becomes zero. It follows from the proof of Proposition 1 that this only occurs if more than a proportion $\min(\alpha_1, 1-\alpha_2)$ is replaced. We thus have that for $\alpha < e^{-1}$, the asymptotic breakdown point equals

$$\min(\alpha, 1 - \alpha, \alpha_1, 1 - \alpha_2). \tag{A.1}$$

By means of a similar reasoning, but reverting the role of the explosion and implosion scenarios, gives that (A.1) is also the asymptotic breakdown point for $1 - e^{-1} < \alpha < 1$. Finally, note that $\alpha = 1 - e^{-1}$ corresponds with $\hat{\lambda} = \hat{q}_\alpha$, for which the result is immediate.

B Influence function

Maximum Likelihood. The parameter of interest is $\theta = (\lambda, \beta)'$. The IF of the ML estimator equals the product between the inverse of the Fisher information matrix in (4.1) and the score function (Hampel et al., 1986), i.e.

$$IF(x_0; \theta_{\text{ML}}, F_{\lambda,\beta}) = I_\theta^{-1} \psi_{\lambda,\beta}(x_0),$$

where $\psi_{\lambda,\beta}(x)$ is the score function

$$\psi_{\lambda,\beta}(x) = \begin{pmatrix} \frac{\partial \log f_{\lambda,\beta}(x)}{\partial \lambda} \\ \frac{\partial \log f_{\lambda,\beta}(x)}{\partial \beta} \end{pmatrix} = \begin{pmatrix} \frac{\beta}{\lambda}[(x/\lambda)^\beta - 1] \\ \frac{1}{\beta}[1 + (\beta \log(x/\lambda))(1 - (x/\lambda)^\beta)] \end{pmatrix}.$$

Method of medians. He and Fung (1999) show that the IF of the method of medians estimator is given by

$$|\nu_{11}\nu_{22} - \nu_{12}\nu_{21}|^{-1} \begin{pmatrix} \nu_{22}\lambda/\beta & -\nu_{21}\lambda/\beta \\ -\nu_{12}\beta & \nu_{11}\beta \end{pmatrix} \begin{pmatrix} \text{sgn}(\log 2 - (\lambda x_0)^\beta) \\ \text{sgn}((1 - (\lambda x_0)^\beta) \log(\lambda x_0)^\beta - c) \end{pmatrix}.$$

Let Y be exponentially distribution with mean one. Then the constants $\nu_{11} = -E[\text{sgn}(\log 2 - Y)(1 - Y)] \approx -0.6931$, $\nu_{12} = -E[\text{sgn}(\log 2 - Y)(1 + \log Y - Y \log Y)] \approx 0.2541$, $\nu_{21} = -E[\text{sgn}(\log Y - Y \log Y - c)(1 - Y)] \approx -0.0354$, $\nu_{22} = -E[\text{sgn}(\log Y - Y \log Y - c)(1 + \log Y - Y \log Y)] \approx -0.8369$ and $c = \text{median}((1 - Y) \log Y) \approx -0.51$.

Quantile. Denote $Q_\alpha(\cdot)$ the functional returning the α-quantile of the distribution in its argument. The functional corresponding to the shape parameter in (2.3) is given by

$$\beta_{\text{Q}}(F) = \frac{G^{-1}(\alpha_2) - G^{-1}(\alpha_1)}{\log[Q_{\alpha_2}(F)/Q_{\alpha_1}(F)]}.$$

Its influence function equals

$$IF(x_0;\beta_{\mathrm{Q}},F)=\frac{-[G^{-1}(\alpha_2)-G^{-1}(\alpha_1)]}{\{\log[Q_{\alpha_2}(F)/Q_{\alpha_1}(F)]\}^2}\left(\frac{IF(x_0,Q_{\alpha_2},F)}{Q_{\alpha_2}(F)}-\frac{IF(x,Q_{\alpha_1},F)}{Q_{\alpha_1}(F)}\right),$$

where the influence function of the quantile functional Q_α is given by

$$IF(x_0;Q_\alpha,F)=\frac{\alpha-I[x_0<Q_\alpha(F)]}{f(Q_\alpha(F))}.$$

The statistical functional corresponding with λ is given by

$$\lambda_{\mathrm{Q}}(F)=\frac{Q_\alpha(F)}{[-\log(1-\alpha)]^{\frac{1}{\beta_{\mathrm{Q}}(F)}}}.$$

Its influence function equals

$$IF(x_0;\lambda_{\mathrm{Q}},F_{\lambda,\beta})=[-\log(1-\alpha)]^{-2/\beta}\Big(IF(x_0;Q_\alpha,F)[-\log(1-\alpha)]^{\frac{1}{\beta}}$$
$$+\beta^{-2}Q_\alpha(F)[\log(-\log(1-\alpha))][-\log(1-\alpha)]^{1/\beta}IF(x,\beta_{\mathrm{Q}},F)\Big).$$

Quantile least squares. We first derive the influence function of the QLS intercept and slope parameters. Let α be a uniformly distributed random variable on $[\overline{\alpha},1-\overline{\alpha}]$ and u the associated density function. For $\overline{\alpha}=0$, we get the results for OLS as a special case. Denote $Q_\alpha(F)=F^{-1}(\alpha)$, $g_\alpha=G^{-1}(\alpha)=\log Q_\alpha(F)$, $c_1=E(g_\alpha)$ and $c_2=\mathrm{Var}(g_\alpha)$. For $\lambda=\beta=1$, we have that for $\overline{\alpha}=0$, $c_1\approx-0.5772$ and $c_2\approx 1.6449$ and for $\overline{\alpha}=1/3$, $c_1\approx-0.3788$ and $c_2\approx 0.0806$. The functional of the QLS slope parameter equals the covariance between $\log Q_\alpha(F)$ and g_α, divided by the variance of g_α. This can be rewritten as

$$b_1(F)=c_2^{-1}\left\{E[g_\alpha\log Q_\alpha(F)]-c_1E[\log Q_\alpha(F)]\right\}.$$

The functional corresponding to the intercept is given by

$$b_0 = E[\log Q_\alpha(F)] - c_1 b_1(F).$$

The influence functions of $b_1(F)$ and $b_0(F)$ are

$$\begin{aligned} IF(x_0; b_1, F) &= c_2^{-1} \int_{\overline{\alpha}}^{1-\overline{\alpha}} \frac{1}{Q_\alpha(F)} (g_\alpha - c_1) IF(x; Q_\alpha, F) u(\alpha) d\alpha \\ IF(x_0; b_0, F) &= \int_{\overline{\alpha}}^{1-\overline{\alpha}} \frac{1}{Q_\alpha(F)} IF(x_0; Q_\alpha, F) u(\alpha) d\alpha - c_1 IF(x_0; b_1, F). \end{aligned}$$

Let $\psi(x) = (\log x - c_1)^2$. By partial integration, one can further show that

$$\begin{aligned} IF(x_0; b_1, F) &= \frac{1}{2c_2(1-2\overline{\alpha})} [\psi(\max(x_0, Q_{\overline{\alpha}}(F))) - \psi(\max(x_0, Q_{1-\overline{\alpha}}(F))) \\ &\quad + (1-\overline{\alpha})\psi(Q_{1-\overline{\alpha}}(F)) - \overline{\alpha}\psi(Q_{\overline{\alpha}}(F)) - c_2(1-2\overline{\alpha})] \\ IF(x_0; b_0, F) &= \frac{1}{1-2\overline{\alpha}} [\log(\max(x_0, Q_{\overline{\alpha}}(F))) - \log(\max(x_0, Q_{1-\overline{\alpha}}(F))) \\ &\quad + (1-\overline{\alpha}) \log(Q_{1-\overline{\alpha}}(F)) - \overline{\alpha} \log(Q_{\overline{\alpha}}(F)) - c_1(1-2\overline{\alpha})] - c_1 IF(x_0; b_1, F). \end{aligned}$$

Since $\beta_{\text{QLS}}(F) = 1/b_1(F)$ and $\lambda_{\text{QLS}}(F) = \exp(b_0(F))$, the influence functions of the QLS shape and scale parameter estimators are

$$IF(x_0; \beta_{\text{QLS}}, F_{\lambda,\beta}) = -\beta^2 IF(x_0; b_1, F) \quad \text{and} \quad IF(x_0; \lambda_{\text{QLS}}, F_{\lambda,\beta}) = \lambda IF(x_0; b_0, F).$$

For the OLS estimator ($\overline{\alpha} = 0$), this reduces to

$$\begin{aligned} IF(x_0; \beta_{\text{OLS}}, F_{\lambda,\beta}) &= -0.5\beta^2[(\log(x_0) - c_1)^2/c_2 - 1] \\ IF(x_0; \lambda_{\text{OLS}}, F_{\lambda,\beta}) &= \lambda[\log(x_0) - c_1 + c_1 IF(x_0; \beta_{\text{OLS}}, F)]. \end{aligned}$$

Repeated median. We first derive the influence function of the RM intercept and slope parameter estimators. Let $\alpha_1, \alpha_2 \sim U[0,1]$ and recall that $g_\alpha = \log Q_\alpha(F)$. The functional of the RM slope and intercept parameters equals

$$b_1(F) = \text{med}_{\alpha_1} H_F(g_{\alpha_1}, \log Q_{\alpha_1}(F)) \quad \text{and} \quad b_0(F) = \text{med}_{\alpha_1} K_F(g_{\alpha_1}, \log Q_{\alpha_1}(F)),$$

where $H_F(x,y) = \text{med}_{\alpha_2}\left(\frac{y - \log Q_{\alpha_2}(F)}{x - g_{\alpha_2}}\right)$ and $K_F(x,y) = \text{med}_{\alpha_2}\left(\frac{g_{\alpha_2} y - x \log Q_{\alpha_2}(F)}{g_{\alpha_2} - x}\right)$. Note that, since $\log Q_\alpha(F) = b_0 + b_1 h(\alpha)$, $b_0(F) = b_0$ and $b_1(F) = b_1$. The first order Taylor expansion of $H_{F_\varepsilon}(g_{\alpha_1}, \log Q_{\alpha_1}(F_\varepsilon))$ yields

$$\begin{aligned} &H_{F_\varepsilon}(g_{\alpha_1}, \log Q_{\alpha_1}(F_\varepsilon)) = H_F(g_{\alpha_1}, \log Q_{\alpha_1}(F)) \\ &\quad + \varepsilon\left(\left.\frac{\partial H_{F_\varepsilon}(g_{\alpha_1}, g_{\alpha_1})}{\partial \varepsilon}\right|_{\varepsilon=0} + \left.\frac{\partial H_F(g_{\alpha_1}, y)}{\partial y}\right|_{y=g_{\alpha_1}} \frac{IF(x_0; Q_{\alpha_1}, F)}{Q_{\alpha_1}(F)} \right) + \mathrm{O}(\varepsilon^2), \end{aligned}$$

where $\partial H_F(x,y)/\partial y = \text{med}_{\alpha_2} 1/(x - g_{\alpha_2})$, since

$$H_F(x,y) = \text{med}_{\alpha_2}\left(\frac{y - \log Q_{\alpha_2}(F) + x - x}{x - g_{\alpha_2}}\right) = 1 + (y - x)\text{med}_{\alpha_2}\frac{1}{x - g_{\alpha_2}}.$$

From the Taylor expansion

$$\log Q_{\alpha_2}(F_\varepsilon) = \log Q_{\alpha_2}(F) + \varepsilon \mathrm{IF}(x_0; Q_{\alpha_2}, F)/Q_{\alpha_2}(F) + \mathrm{O}(\varepsilon^2),$$

it further follows that

$$H_{F_\varepsilon}(x,y) = \text{med}_{\alpha_2}\left(\frac{y - \log Q_{\alpha_2}(F)}{x - g_{\alpha_2}} - \varepsilon\frac{IF(x_0; Q_{\alpha_2}, F)}{Q_{\alpha_2}(F)(x - g_{\alpha_2})}\right) + \mathrm{O}(\varepsilon^2).$$

We thus obtain

$$\left.\frac{\partial H_{F_\varepsilon}(g_{\alpha_1}, g_{\alpha_1})}{\partial \varepsilon}\right|_{\varepsilon=0} = -\text{med}_{\alpha_2}\left(\frac{IF(x_0; Q_{\alpha_2}, F)}{Q_{\alpha_2}(F)(g_{\alpha_1} - g_{\alpha_2})}\right).$$

Combining all these results yields the following approximation for the RM slope parameter

$$b_1(F_\varepsilon) = 1+\varepsilon\cdot\mathrm{med}_{\alpha_1}\left(\mathrm{med}_{\alpha_2}\frac{IF(x_0;Q_{\alpha_1},F)}{(g_{\alpha_1}-g_{\alpha_2})Q_{\alpha_1}(F)}-\mathrm{med}_{\alpha_2}\frac{IF(x_0;Q_{\alpha_2},F)}{(g_{\alpha_1}-g_{\alpha_2})Q_{\alpha_2}(F)}\right)+\mathrm{O}(\varepsilon^2).$$

The influence function of the RM slope estimator equals

$$IF(x_0;b_1,F) = \mathrm{med}_{\alpha_1}\left(\mathrm{med}_{\alpha_2}\frac{IF(x_0;Q_{\alpha_1},F)}{(g_{\alpha_1}-g_{\alpha_2})Q_{\alpha_1}(F)} - \mathrm{med}_{\alpha_2}\frac{IF(x_0;Q_{\alpha_2},F)}{(g_{\alpha_1}-g_{\alpha_2})Q_{\alpha_2}(F)}\right).$$

Analogously, one can show that the influence function of the RM intercept estimate is

$$IF(x_0;b_0,F) = \mathrm{med}_{\alpha_1}\left(\mathrm{med}_{\alpha_2}\frac{g_{\alpha_1}IF(x_0;Q_{\alpha_2},F)}{(g_{\alpha_1}-g_{\alpha_2})Q_{\alpha_2}(F)} - \mathrm{med}_{\alpha_2}\frac{g_{\alpha_2}IF(x_0;Q_{\alpha_1},F)}{(g_{\alpha_1}-g_{\alpha_2})Q_{\alpha_1}(F)}\right).$$

Since $\beta_{\mathrm{RM}}(F) = 1/b_1(F) = \beta$ and $\lambda_{\mathrm{RM}}(F) = \exp(b_0(F)) = \lambda$, the influence functions of the RM shape and scale parameters are

$$IF(x_0;\beta_{\mathrm{RM}},F_{\lambda,\beta}) = -\beta^2 IF(x_0;b_1,F) \quad \text{and} \quad IF(x_0;\lambda_{\mathrm{RM}},F_{\lambda,\beta}) = \lambda IF(x;b_0,F).$$

Median/MAD. The influence functions of the median/MAD shape and scale estimators equal

$$\begin{aligned} IF(x_0;\beta_{\mathrm{med/MAD}},F_{\lambda,\beta}) &= -1.3037\beta^2 IF(\log x_0;MAD,G) \\ IF(x_0;\lambda_{\mathrm{med/MAD}},F_{\lambda,\beta}) &= \lambda[IF(\log x_0;Q_{0.5},G) + 0.4778 IF(\log x_0;MAD,G)], \end{aligned}$$

where $IF(\log x_0;Q_{0.5},G) = IF(x_0;Q_{0.5},F)/Q_\alpha(F)$. The function $IF(\log x_0;MAD,G)$

is the influence function of the MAD at the log-Weibull distribution

$$\frac{\operatorname{sgn}(|\log x_0 - g_{0.5}| - MAD) - (g(g_{0.5} + MAD) - g(g_{0.5} - MAD))IF(\log x_0; Q_{0.5}, G)}{2(g(g_{0.5} + MAD) + g(g_{0.5} - MAD))}$$

(see Huber 1981).

Median/Q_n. The influence functions of the median/Q_n shape and scale estimators equal

$$\begin{aligned} IF(x_0; \beta_{\text{med}/Q_n}, F_{\lambda,\beta}) &= -\beta^2 IF(\log x_0; Q_n, G) \\ IF(x_0; \lambda_{\text{med}/Q_n}, F_{\lambda,\beta}) &= \lambda[IF(\log x_0; Q_{0.5}, G) - IF(\log x_0; Q_n, G) \log\log 2], \end{aligned}$$

where $d = 1.9577$ and $IF(y_0; Q_n, G)$ is the influence function of the Q_n estimator given by

$$IF(y_0; Q_n, G) = d\frac{\frac{1}{4} - G(y_0 + d^{-1}) + G(y_0 - d^{-1})}{\int g(z + d^{-1})g(z)dz}$$

(see Rousseeuw and Croux 1993).

References

Adatia, A. and L. Chan (1982). Robust procedures for estimating the scale parameter and predicting future order statistics of the Weibull distribution. *IEEE Transactions on Reliability 31*, 491–198.

Cohen, A. C. (1965). Maximum likelihood estimation in the weibull distribution based on complete and on censored samples. *Technometrics 7*, 579–588.

Dixit, U. J. (1994). Bayesian approach to prediction in the presence of outliers for Weibull distribution. *Metrika 41*, 127–136.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust Statistics. The Approach based on Influence Functions*. Wiley, New York.

He, X. and W. K. Fung (1999). Method of medians for lifetime data with Weibull models. *Statistics in Medicine 18*, 1993–2009.

Hoel, D. (1972). A representation of mortality data by competing risks. *Biometrics 28*, 475–488.

Huber, P. (1981). *Robust Statistics*. Wiley, New York.

Kalbfleisch, J. and R. Prentice (1980). *The statistical analysis of failure time data*. Wiley, New York.

Li, Y.-M. (1994). A general linear-regression analysis applied to the 3-parameter Weibull distribution. *IEEE transactions on reliability 43*, 255–263.

Lingappaiah, G. (1976). Effect of outliers on the estimation of parameters. *Metrika 23*, 27–30.

Marks, N. (2005). Estimation of Weibull parameters from common percentiles. *Journal of Applied Statistics 32*, 17–24.

Muralidharan, K. and P. Lathika (2006). Analysis of instantaneous and early failures in Weibull distribution. *Metrika 64*, 305–316.

Olive, D. (2006). Robust estimators for transformed location scale families. *Mimeo*.

Rousseeuw, P. and C. Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association 88*, 1273–1283.

Seki, T. and S.-I. Yokoyama (1996). Robust parameter-estimation using the bootstrap method for the 2-parameter Weibull distribution. *IEEE Transactions on Reliability 45*, 34–41.

Shier, D. and K. Lawrence (1984). A comparison of robust regression techniques for the estimation of Weibull parameters. *Communications in Statistics B-Simulation and Computation 13*, 743–750.

Siegel, A. F. (1982). Robust regression using repeated medians. *Biometrika 69*, 242–244.