

Adaptively truncated maximum likelihood regression with asymmetric errors

Alfio Marazzi and Victor J. Yohai
University of Lausanne and University of Buenos Aires and CONICET

November 2002

Running title. Truncated regression.

Key words. Maximum breakdown point, efficient estimation, adaptive estimation, adaptive cut-off, asymmetrically distributed errors.

AMS classification. Primary 62J05; secondary 62F35.

Abstract

We consider robust estimators for the linear regression model with asymmetric (or symmetric) error distribution. We assume that the error model belongs to a location-scale family of distributions. Since in the asymmetric case the mean response is very often the parameter of interest and scale is a main component of mean, we do not assume that scale is a nuisance parameter. First, we show how to convert an ordinary robust estimate for the usual model with symmetric errors to an estimate for the more general model with asymmetric errors. Then, in order to improve efficiency, we introduce the truncated maximum likelihood or TML-estimator. A TML-estimate is computed in three steps: first, an initial high breakdown point estimate is computed; then, observations that are unlikely under the estimated model are rejected; finally, the maximum likelihood estimate is computed with the retained observations. The rejection rule used in the second step is based on a cut-off parameter that can be tuned to attain the desired efficiency while maintaining the breakdownpoint of the initial estimator (e.g., 50%). Optionally, one can use a new adaptive cut-off that, asymptotically, does not reject any observation when the data are generated according to the model. Under the model, the influence function of this adaptive TML-estimator (or ATML-estimator) coincides with the influence function of the maximum likelihood estimator. The ATML-estimator is, therefore, fully efficient at the model; nevertheless, its breakdown point is not smaller than the breakdown point of the initial estimator. We evaluate the TML- and ATML-estimators for finite sample sizes with the help of simulations and discuss an example with real data.

1 Introduction

Positive random variables with asymmetric distributions arise in many applications (e.g., analysis of income and expenditures, failure times, output of biological systems). Often the population mean (e.g., a mean expenditure in a budgeting problem) is the parameter of interest and depends upon a number of covariates. Unfortunately, the data may contain outliers and the mean is a difficult parameter to estimate well in this case.

The main approaches to regression with asymmetric errors are: generalized linear models (McCullagh and Nelder, 1989), transformation and weighting (Carroll and Ruppert, 1988), and non-parametric regression based on ranks (Hettmansperger and McKean, 1998). Some special response models have also been used (e.g., Williams, 1997). Both generalized linear models and transformation and weighting allow modeling of the mean with the help of covariates; however, they use maximum likelihood, quasi likelihood, and generalized least squares methods of estimation which are very sensitive to outliers. Rank-based non-parametric regression is naturally outlier resistant; however, it is not intended for modeling the response mean and is less efficient than parametric methods when a model is adequate.

Recent research in parametric robust estimation (Hampel et al., 1986) pays attention to univariate asymmetric models (e.g., Victoria-Feser and Ronchetti, 1994, 1997; Marazzi and Ruffieux, 1996, 1999) as well as to asymmetric response regression (e.g., Cantoni and Ronchetti, 2001). Most of the proposed regression methods can attain high efficiency levels; none of them can however simultaneously attain both high efficiency and maximum (50%) breakdown point.

In this paper we follow the parametric approach and propose high efficiency and high breakdown point estimators for a class of regression models with asymmetric (or symmetric) error distribution. We assume that the error model belongs to a location-scale family of distributions. Examples are the Log-Weibull and the Gaussian distributions. Since in the asymmetric case the mean response is very often the parameter of interest and scale is a main component of mean, we do not assume, as usual, that scale is a nuisance parameter. The mean of the estimated model will then be interpreted as a robust estimate of the population mean after removal of the extreme observations.

First, we show how to convert an ordinary robust estimate for the usual regression model with symmetric errors to an estimate for the more general model with asymmetric errors. In general, the transformed estimator keeps the same robustness properties than the original estimate, e.g., the same breakdown point; however it is not necessarily efficient. In order to improve efficiency, we introduce a new class of estimators that we call the truncated maximum likelihood estimators or TML-estimators. The computation of a TML-estimate is very transparent: in a first step, an initial high breakdown point but inefficient S-estimate is computed; in a second step, observations that are unlikely under the estimated model are rejected; in a third step, the maximum likelihood estimate is computed with the retained observations; finally, some corrections are introduced to suppress bias at the model. This proposal generalizes the sugges-

tion of Rousseeuw and Leroy (1987) to compute – in the symmetric error case – a weighted least squares estimate, skipping those observations with absolute standardized residuals (with respect to the least median of squares estimate) greater than some fixed cut-off.

The rejection rule used in the second step is based on a cut-off parameter that can be tuned to attain the desired efficiency while maintaining the breakdownpoint of the initial estimator. However, under the model, a certain fraction of data is systematically rejected and the estimate suffers a small efficiency loss. Optionally, an adaptive rejection rule may then be used. This rule (which improves to the one proposed in Gervini and Yohai, 2002) is defined as follows: first, the empirical distribution of the observed likelihoods with respect to the initial estimate is obtained and compared with the theoretical one; second, observations with the smallest likelihoods are rejected, so that the empirical distribution of the remaining observed likelihoods is stochastically smaller than the theoretical one. No observation is therefore (asymptotically) rejected when the data agree with the model and we show that the influence function of the ATML-estimator coincides with the influence function of the maximum likelihood estimator. This result strongly suggests that the ATML-estimator is fully efficient.

The idea of adaptively trimming observations which are least likely to occur as indicated by the likelihood has also been investigated by Bednarski and Clarke (1993) and Clarke (2000) in the contaminated Gaussian error setting. Their estimator and rejection rule are simultaneously defined without the help of an initial robust estimator to identify the outliers. In addition, they choose the trimming proportion that minimizes an estimate of the asymptotic variance of the estimates. Field and Smith (1994) also propose a class of weighted maximum likelihood estimates where outliers are downweighted using two different probability scales. Markatou et al. (1998) develop an approach to robust and efficient estimation, where a weighted least squares estimate starting with an initial robust estimate is proposed; the weights are based on a measure of disparity between the density of the errors under the initial model and the smoothed empirical density of the residuals. The methods we consider in this paper are computationally simpler and theoretically more tractable than the approaches mentioned above.

In Section 2, we introduce the models. In Section 3, we define the estimators and the adaptive cut-off value. Section 4 states that the breakdownpoint of the new estimators is not smaller than the breakdown point of the initial estimator. In addition, the influence functions of the TML- and the ATML-estimators are derived. In Section 5, we provide empirical results on the performance of the new estimators for finite sample sizes. In Section 6, we discuss an example concerning costs of hospital stays. Proofs are collected in the Appendix.

2 Models and notations

We consider a random sample $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$, where \mathbf{x}_i is a vector of p explanatory variables and y_i is a real response variable. We assume that they are linked by the linear relationship

$$y_i = \mathbf{x}_i^T \boldsymbol{\theta} + \sigma e_i,$$

where $\boldsymbol{\theta} \in R^p$ is a parameter vector and the first component θ_1 of $\boldsymbol{\theta}$ is an intercept term. The carriers \mathbf{x}_i are distributed according to a cdf G with density g with respect to the Lebesgue measure; the unobservable errors e_i are i.i.d. random variables with cdf $F(\cdot/\sigma)$, where σ is an unknown scale parameter; e_i is independent of \mathbf{x}_i . The joint cdf of (\mathbf{x}_i, y_i) is denoted by H and its density by h . In practice, the actual distribution of the errors is not known, so we have to use a hypothetical model cdf F_0 instead of F . We assume that $F_0(\cdot) = F_{0,1}(\cdot)$, where $F_{0,1}$ is the standard member of a parametric family of asymmetric or symmetric distributions with location parameter λ , scale parameter σ , density $f_{\lambda,\sigma}$ and cdf $F_{\lambda,\sigma}(z) = F_{0,1}((z - \lambda)/\sigma)$. Thus, the cdf of y_i is $F_{\lambda_i,\sigma}$ with $\lambda_i = \mathbf{x}_i^T \boldsymbol{\theta}$. The joint model cdf of (\mathbf{x}_i, y_i) is denoted by H_0 and its density by $h_0(x, y) = \sigma^{-1} f_0((y - x^T \boldsymbol{\theta})/\sigma) g(x)$. We want to estimate the parameter vector $(\boldsymbol{\theta}, \sigma)$, without assuming, as often is the case, that σ is a nuisance parameter.

The log-likelihood function will be denoted by $\rho_{\lambda,\sigma}(z) = -\ln f_{\lambda,\sigma}(z)$ and the score functions by $s_{1,\lambda,\sigma}(z) = \partial \ln f_{\lambda,\sigma}(z) / \partial \lambda$, $s_{2,\lambda,\sigma}(z) = \partial \ln f_{\lambda,\sigma}(z) / \partial \sigma$. We assume that $\rho_{0,1}$ is convex and that $\int_0^\infty \rho_{0,1}(z) f_0(z) dz < \infty$. We will use the abbreviations

$$\rho(z) = \rho_{0,1}(z), \quad s_1(z) = s_{1,0,1}(z), \quad s_2(z) = s_{2,0,1}(z) + 1.$$

Examples of location-scale error models are the *Gaussian model* with density

$$f_{\lambda,\sigma}(z) = \phi((z - \lambda)/\sigma), \quad -\infty < z < \infty,$$

and *Log-Weibull model* with density

$$f_{\lambda,\sigma}(z) = \frac{1}{\sigma} \exp \left[\left(\frac{z - \lambda}{\sigma} \right) - \exp \left(\frac{z - \lambda}{\sigma} \right) \right], \quad -\infty < z < \infty.$$

3 Definition of the estimators

First, we observe that it is in general possible to convert a regression estimate for the usual model with symmetric errors to an estimate for the more general model introduced in Section 2.

3.1 General robust estimators for asymmetric errors. Suppose that $\hat{F}_\theta(z) = (1/n) \sum I(y_i - \mathbf{x}_i^T \boldsymbol{\theta} \leq z)$ for given $\boldsymbol{\theta}$ (where $I(\cdot)$ denotes the indicator function) and let U denote a scale equivariant functional defined on the set of the distribution functions on R^1 . Consider the estimate of $\boldsymbol{\theta}$ defined by

$$\mathbf{T}^* = \arg \min_{\boldsymbol{\theta}} U(\hat{F}_\theta)$$

and note that many robust estimates (e.g., S-, LMS-, LTS-, τ -estimators) can be defined in this way. In addition, let

$$S^* = \inf_{\theta} U(\hat{F}_{\theta}).$$

Finally, let $a = \arg \min_{\lambda} U(F_{\lambda,1})$, suppose that this minimum be unique, and let $b = U(F_{a,1})$. Define the *corrected estimates* S and \mathbf{T} by

$$S = S^*/b, \quad \mathbf{T} = (T_1^* - aS, T_2^*, \dots, T_p^*).$$

Then, (\mathbf{T}, S) is a consistent estimate of (θ, σ) when the underlying error distribution is $F_{0,1}$.

As a specific example, we consider the S-estimate (\mathbf{T}^*, S^*) of (θ, σ) defined by $\mathbf{T}^* = \arg \min_{\theta} S_{k_0}(\theta)$, where $S_{k_0}(\theta)$ is the solution of

$$\frac{1}{n-p} \sum_{i=1}^n \chi_{k_0}((y_i - \mathbf{x}_i^T \theta)/S) = 0.5$$

with respect to S , for given θ , and $S^* = S_{k_0}(\mathbf{T}^*)$. We assume that the function χ_k and the constants a_0 and k_0 are defined by

$$\chi_k(z) = \begin{cases} 3(z/k)^2 - 3(z/k)^4 + (z/k)^6 & \text{if } |z| \leq k, \\ 1 & \text{if } |z| > k, \end{cases}$$

$$a_0 = \arg \min_a \int \chi_{k_0}(z-a) f_0(z) dz \quad \text{and} \quad \int \chi_{k_0}(z-a_0) f_0(z) dz = 0.5,$$

In the Gaussian case, one obtains $a_0 = 0$ and $k_0 = 1.548$; in the Log-Weibull case, $a_0 = -0.135$ and $k_0 = 1.718$. The *corrected S-estimate* is given by $T_1 = T_1^* - S^* a_0$, $T_j = T_j^*$ for $j = 2, \dots, p$, and $S = S^*$.

In general, the estimators \mathbf{T} and S keep the same robustness properties as the original estimates, e.g., the same breakdown point. However they are not necessarily efficient. One way of increasing the efficiency is to use \mathbf{T} and S as starting points to detect and reject those observations which may appear as outliers and then compute the maximum likelihood estimator with the remaining observations. This proposal is made precise in the next two subsections.

3.2 Rejection rules. We suppose that $(T^{(0)}, S^{(0)})$ is an initial high breakdown point estimator, such as the S-estimator defined above. In order to define cut-off values for outlier rejection, we consider the standardized residuals $r_i = (y_i - \mathbf{x}_i^T \mathbf{T}^{(0)})/S^{(0)}$ with respect to $(\mathbf{T}^{(0)}, S^{(0)})$ and compute the empirical negative log-likelihoods $\rho_i = \rho(r_i)$ for $i = 1, \dots, n$. A large ρ_i corresponds to an observation (\mathbf{x}_i, y_i) with a small likelihood under the model F_0 and suggests that this observation is an outlier. To be precise, let F_0^+ denote the cdf of $\rho(e)$ under the model and η be some large quantile of F_0^+ . We then define

$$w_i = I(\rho_i < \eta)$$

and reject observations (\mathbf{x}_i, y_i) such that $w_i = 0$. A *fixed upper cut-off value* u and a *fixed lower cut-off value* l are obtained by solving $\rho(z) = \eta$.

Following an idea of Gervini and Yohai (2002), we may also define adaptive cut-off values that depend on the degree of contamination. For this purpose, we compare the empirical cdf F_n^+ of ρ_1, \dots, ρ_n with F_0^+ ; observations with the largest ρ_i are rejected, so that the empirical distribution of the remaining ρ_i -s is stochastically smaller than the theoretical one. Specifically, let \tilde{F}_n^+ denote the empirical cdf of ρ_1, \dots, ρ_n truncated at t , i.e.,

$$\tilde{F}_n^+(z) = \begin{cases} F_n^+(z)/F_n^+(t) & \text{if } z \leq t, \\ 1 & \text{otherwise.} \end{cases}$$

We look for the largest $t > 0$ such that $\tilde{F}_n^+(z) \geq F_0^+(z)$ for all $z \geq \eta$, where η is some large quantile of F_0^+ . More precisely, let

$$\begin{aligned} \bar{t}_n &= \sup(t \mid F_n^+(z)/F_n^+(t) \geq F_0^+(z) \text{ for all } z \geq \eta) \\ &= \sup(z \mid F_n^+(z) \leq \alpha_n), \end{aligned}$$

where

$$\alpha_n = \min[\inf_{z \geq \eta} F_n^+(z)/F_0^+(z), 1].$$

Since there may be an entire interval such that $F_n^+(z) = \alpha_n$, the lower bound being $\underline{t}_n = \inf(z \mid F_n^+(z) \geq \alpha_n)$, we define the *adaptive cut-off value* on the likelihood scale as

$$t_n = \max[F_n^{+-1}(\alpha_n), \eta],$$

where F_n^{+-1} is given by some interpolation rule between \underline{t}_n and \bar{t}_n . (Note that $\bar{t}_n \geq \eta$, but that $F_n^{+-1}(\alpha_n)$ may be lower than η .) Finally, we define weights

$$w_i = I(\rho_i < t_n)$$

and reject an observation (\mathbf{x}_i, y_i) if $w_i = 0$. An *upper adaptive cut-off value* u_n and a *lower adaptive cut-off value* l_n for the residuals are obtained by solving $\rho(z) = t_n$. Intuitively, the empirical distribution of the remaining residuals has tails that are comparable with those of F_0 .

Remark 1. Gervini and Yohai (2002) propose an adaptive cut-off procedure based on the difference $F_0^+(z) - F_n^+(z)$ to measure the proportion of outliers. The rule proposed in this paper seems more natural and facilitates the analysis.

Remark 2. An attempt to define \bar{t}_n as the largest $t > 0$ such that $\tilde{F}_n^+(z) \geq F_0^+(z)$ for all $z \geq 0$ did not provide satisfactory results. Simulations showed that the cut-off value obtained in this way is often too low for small “clean” samples. For this reason, to ensure high efficiency for small samples, we introduced the lower bound η into the definition of \bar{t}_n .

Remark 3. In order to keep the analysis simple, we consider here only hard rejection weights (with values 0 or 1). This means that we are looking for protection against outliers that are larger than a certain “rejection point” and

no protection against lower contaminating observations. A smoother weight system (e.g., biweight) would be more appropriate if a mild and more scattered contamination is expected.

3.3 The truncated maximum likelihood estimator. Finally, we specify a complete procedure based on adaptive or non-adaptive outlier rejection.

Step 1. Compute a corrected S-estimate $(T^{(0)}, S^{(0)})$.

Step 2. Using the standardized residuals with respect to $(T^{(0)}, S^{(0)})$, compute the fixed cut-off values l, u or the adaptive cut-off values l_n, u_n , and the weights w_i . Set $\tilde{n} = \sum w_i$.

Step 3. Compute the maximum likelihood estimator after rejection of the observations such that $w_i = 0$, i.e., solve for \mathbf{T} and S :

$$\frac{1}{\tilde{n}} \sum_{i=1}^n w_i s_1((y_i - \mathbf{x}_i^T \mathbf{T})/S) \mathbf{x}_i = 0, \quad (1)$$

$$\frac{1}{\tilde{n}} \sum_{i=1}^n w_i s_2((y_i - \mathbf{x}_i^T \mathbf{T})/S) = 1. \quad (2)$$

Step 4. Correct the scale estimate: in the non-adaptive case, replace S with S/b where b is the solution of

$$\frac{1}{F_0(u) - F_0(l)} \int_l^u s_2\left(\frac{z}{b}\right) f_0(z) dz = 1; \quad (3)$$

in the adaptive case, replace l, u with l_n, u_n in (3).

In the non-adaptive case, we call the estimator defined in Steps 3 and 4 the *truncated maximum likelihood estimator* or *TML-estimator*; in the adaptive case, the estimator will be called the *adaptively truncated maximum likelihood estimator* or *ATML-estimator*. In general, we will use the notation (\mathbf{T}, S) to denote both the TML- and the ATML-estimators.

Remark 4. Often, it is more convenient to use the following equivalent definition of the TML-estimator:

$$(\mathbf{T}, S) = \arg \min_{\boldsymbol{\theta}, \sigma} \left[\frac{1}{\tilde{n}} \sum_1^n w_i \rho((y_i - \mathbf{x}_i^T \boldsymbol{\theta})/\sigma) + \beta \ln \sigma \right], \quad (4)$$

$$\beta = \frac{1}{F_0(u) - F_0(l)} \int_l^u s_2(z) f_0(z) dz. \quad (5)$$

The limits l and u must be replaced with l_n and u_n in the adaptive case. In addition, \tilde{n} can be replaced with $\tilde{n} - p$.

Remark 5. By definition, the cut-off values satisfy $f_0(u) = f_0(l)$ and $f_0(u_n) = f_0(l_n)$; therefore, $\int_l^u s_1(z) f_{0,1}(z) dz = \int_{l_n}^{u_n} s_1(z) f_{0,1}(z) dz = 0$. It follows that \mathbf{T} is asymptotically unbiased under the hypothetical model with error cdf F_0 . Consistency of S is ensured thanks to the correction in Step 4 or the use of β in equation (4). Moreover, the hard rejection ATML-estimator is asymptotically

unbiased under point contaminated distributions of the form $F = (1 - \epsilon)F_0 + \epsilon\Delta_{z_0}$, where $0 \leq \epsilon \leq 1$, Δ_{z_0} denotes the cdf of a point mass at z_0 , and z_0 belongs to the rejection domain.

4 Robustness properties

In this section we show that the breakdown point of the TML- and ATML-estimators are not lower than the breakdown point of the initial estimator; moreover, we provide the influence functions of these estimators at the model.

4.1 Functional definitions. First, we establish the functional definition of the ATML-estimator. We denote by $\mathbf{T}^{(0)}(H)$, $S^{(0)}(H)$ the initial estimator functionals, let (\mathbf{x}, y) denote a generic observation and let

$$r_H(\mathbf{x}, y) = \frac{y - \mathbf{x}^T \mathbf{T}^{(0)}(H)}{S^{(0)}(H)}$$

be a generic standardized residual with respect to $\mathbf{T}^{(0)}(H)$. We define F_H^+ as the distribution of $\rho(r_H(\mathbf{x}, y))$ and let F_0^+ denote the distribution of $\rho(e_i)$ under the model. Also, let

$$\begin{aligned} \alpha(H) &= \min[\inf_{z \geq \eta} F_H^+(z)/F_0^+(z), 1], \\ t(H) &= \max[F_H^{+^{-1}}(\alpha(H)), \eta], \end{aligned}$$

where $F_H^{+^{-1}}(\alpha(H))$ is given by some interpolation rule between

$$\underline{t}(H) = \inf\{z \mid F_H^+(z) \geq \alpha(H)\} \text{ and } \bar{t}(H) = \sup\{z \mid F_H^+(z) \leq \alpha(H)\}.$$

Moreover, let

$$\tilde{F}_H^+(z) = \begin{cases} F_H^+(z)/\alpha(H) & \text{if } z \leq t(H), \\ 1 & \text{if } z > t(H), \end{cases}$$

be the distribution F_H^+ truncated at $t(H)$. We set

$$w_H(\mathbf{x}, y) = I(\rho(r_H(\mathbf{x}, y)) < t(H)),$$

and define

$$L(H, \boldsymbol{\theta}, \sigma) = \frac{1}{\alpha(H)} E_H[w_H(\mathbf{x}, y)\rho((y - \mathbf{x}^T \boldsymbol{\theta})/\sigma)] + \beta(H) \ln(\sigma),$$

where

$$\beta(H) = \frac{1}{F_0(u(t(H))) - F_0(l(t(H)))} \int w(\rho(z)/t(H)) s_2(z) f_0(z) dz,$$

and $l(t)$, $u(t)$ denote the lower and upper solutions of $\rho(z) = t$. The final ATML-estimator functionals $\mathbf{T}(H)$ and $S(H)$ are then defined by

$$(\mathbf{T}(H), S(H)) = \arg \min_{(\boldsymbol{\theta}, \sigma)} L(H, \boldsymbol{\theta}, \sigma). \quad (6)$$

The TML-estimator is a special case of (6) with $t(H) = \eta$, $\alpha(H) = F_H^+(\eta)$, $\beta(H) = \beta$, $l(t(H)) = l$, and $u(t(H)) = u$, l , and u being the lower and upper solutions of $\rho(z) = \eta$.

4.2 Breakdown point. We now consider the neighborhood of H_0 given by

$$\mathcal{H}_\varepsilon = \{H : H = (1 - \varepsilon)H_0 + \varepsilon H^* \text{ and } H^* \text{ is a distribution on } R^{p+1}\},$$

where $\varepsilon \in [0, 0.5)$. We assume that the breakdown point of the initial estimators is $\varepsilon_0^* > 0$. In particular, we assume that, for any $\varepsilon < \varepsilon_0^*$, there exists $\sigma_1(\varepsilon) > 0$ and $\sigma_2(\varepsilon) < \infty$, such that $\sigma_1(\varepsilon) < S^{(0)}(H) < \sigma_2(\varepsilon)$ for any $H \in \mathcal{H}_\varepsilon$.

Theorem 1. If $P(\mathbf{x}^T \boldsymbol{\theta} \neq 0) > 0$ for all $\boldsymbol{\theta}$ and F_0^+ has a finite mean, then the breakdown point of $(\mathbf{T}(H), S(H))$ is larger or equal than ε_0^* .

4.3 Influence functions. In order to derive the influence functions, we observe that the functional $(\mathbf{T}(H), S(H))$ can also be defined as a solution of the system

$$E_H \left[w \left(\frac{1}{t(H)} \rho \left(\frac{y - \mathbf{x}^T \mathbf{T}^{(0)}(H)}{S^{(0)}(H)} \right) \right) s_1 \left(\frac{y - \mathbf{x}^T \mathbf{T}(H)}{S(H)} \right) \mathbf{x} \right] = 0, \quad (7)$$

$$E_H \left[w \left(\frac{1}{t(H)} \rho \left(\frac{y - \mathbf{x}^T \mathbf{T}^{(0)}(H)}{S^{(0)}(H)} \right) \right) s_2 \left(\frac{y - \mathbf{x}^T \mathbf{T}(H)}{S(H)} \right) \right] = \alpha(H)\beta(H), \quad (8)$$

where $w(z) = I(|z| < 1)$. We denote by $IF((\mathbf{x}_0, y_0), Z, H)$ the influence function at (\mathbf{x}_0, y_0) of a functional $Z(H)$, when the data are distributed according to H . At the model $H_0(\mathbf{x}, y)$, we use the abbreviation $z_0 = (y_0 - \mathbf{x}_0^T \boldsymbol{\theta})/\sigma$ and note that $I_{\mathbf{T}}$ and I_S depend on (\mathbf{x}_0, y_0) via (\mathbf{x}_0, z_0) . Therefore, we use the abbreviations $I_{\mathbf{T}}^0(\mathbf{x}_0, z_0)$, $I_S^0(\mathbf{x}_0, z_0)$ (or $I_{\mathbf{T}}^0$, I_S^0) in place of $IF((\mathbf{x}_0, y_0), \mathbf{T}, H_0)$, $IF((\mathbf{x}_0, y_0), S, H_0)$.

Theorem 2. The influence function at the model of the non-adaptive TML-estimator (\mathbf{T}, S) is given by

$$\begin{pmatrix} I_{\mathbf{T}}^0(\mathbf{x}_0, z_0) \\ I_S^0(\mathbf{x}_0, z_0) \end{pmatrix} = M \mathbf{q}(\mathbf{x}_0, z_0), \quad (9)$$

where

$$M = \begin{pmatrix} a_1 E_G[\mathbf{x}\mathbf{x}^T] & b_1 E_G[\mathbf{x}] \\ a_2 E_G[\mathbf{x}^T] & b_2 \end{pmatrix}^{-1},$$

$$a_1 = \frac{1}{\sigma} \int_l^u s_1'(z) f_0(z) dz, \quad a_2 = \frac{1}{\sigma} \int_l^u s_2'(z) f_0(z) dz, \quad (10)$$

$$b_1 = \frac{1}{\sigma} \int_l^u s_1'(z) z f_0(z) dz, \quad b_2 = \frac{1}{\sigma} \int_l^u s_2'(z) z f_0(z) dz, \quad (11)$$

$$\mathbf{q}(\mathbf{x}_0, z_0) = \begin{pmatrix} \mathbf{c}_1(\mathbf{x}_0, z_0) + \mathbf{d}_1(\mathbf{x}_0, z_0) \\ c_2(z_0) - \alpha\beta + d_2(\mathbf{x}_0, z_0) - \beta I_{\alpha}^0(\mathbf{x}_0, z_0) \end{pmatrix},$$

$$\begin{aligned} \mathbf{c}_1(\mathbf{x}_0, z_0) &= s_1(z_0)I(l < z_0 < u)\mathbf{x}_0, \\ c_2(z_0) &= s_2(z_0)I(l < z_0 < u) - \alpha\beta, \end{aligned}$$

$$\begin{aligned} \mathbf{d}_1(\mathbf{x}_0, z_0) &= \frac{f_0(u)}{\sigma} ([s_1(u) - s_1(l)] E_G[\mathbf{xx}^T] I_{\mathbf{T}^{(0)}}^0 + [us_1(u) - ls_1(l)] E_G[\mathbf{x}] I_{S^{(0)}}^0), \\ d_2(\mathbf{x}_0, z_0) &= \frac{f_0(u)}{\sigma} ([s_2(u) - s_2(l)] E_G[\mathbf{x}]^T I_{\mathbf{T}^{(0)}}^0 + [us_2(u) - ls_2(l)] I_{S^{(0)}}^0). \end{aligned}$$

$$\alpha = F_0(u) - F_0(l), \quad \beta = \frac{1}{F_0(u) - F_0(l)} \int_l^u s_2(z) f_0(z) dz,$$

$$I_\alpha^0(\mathbf{x}_0, z_0) = \frac{1}{\sigma} [uf_0(u) - lf_0(l)] I_{S^{(0)}} + \Delta_{\rho(z_0)}(\eta) - \alpha,$$

and $(I_{\mathbf{T}^{(0)}}^0, I_{S^{(0)}})$ denotes the influence function of the initial estimator $(\mathbf{T}^{(0)}, S^{(0)})$.

Remark 1. In practice, G and σ must be estimated. We propose to use S in place of σ and to estimate $E_G[\mathbf{xx}^T]$ and $E_G[\mathbf{x}]$ with

$$\hat{E}_G[\mathbf{xx}^T] = \frac{\sum w_i \mathbf{x}_i \mathbf{x}_i^T}{\sum w_i}, \quad \hat{E}_G[\mathbf{x}] = \frac{\sum w_i \mathbf{x}_i}{\sum w_i}.$$

Theorem 3. If $F_0(u)$ has a continuously differentiable density function f_0 , such that $f_0^+(z) > 0$ for all $z > 0$, the influence function of the ATML-estimator at the model H_0 is given by (9), where $\mathbf{q}(\mathbf{x}_0, z_0) = (s_1(z_0)\mathbf{x}_0, s_2(z_0) - 1)^T$, and a_1, a_2, b_1, b_2 are given by (10)-(11) with $u = -l = \infty$.

Remark 2. According to Theorem 3, the influence function of the ATML-estimator coincides with the influence function of the maximum likelihood estimator. This result strongly suggests that the ATML-estimator is fully efficient at the model H_0 . A rigorous proof of this theorem is still lacking but we conjecture that it can be given using the methods of Gervini and Yohai (2002). (These authors prove asymptotic normality of a truncated least squares coefficient estimator assuming a symmetric error distribution with σ as a nuisance parameter; they also show that the rate of convergence of the initial estimator may affect the rate of convergence of the final estimator. For this reason, the TML-estimator starts with an S-estimator which is asymptotically normal at rate $n^{-1/2}$). On the other side, Theorem 3 is not useful to estimate the finite sample variance of the ATML parameter estimates, which is clearly larger than the variance of the maximum likelihood estimate. Empirical computations show however that a reasonable estimate of the covariance matrix of the ATML-estimate is

$$\text{Cov}((\mathbf{T}, S)) = \frac{1}{n} M \left[\frac{1}{\sum w_i} \sum_{i=1}^n w_i \int \mathbf{q}(\mathbf{x}_i, z) \mathbf{q}(\mathbf{x}_i, z)^T f_0(z) dz \right] M^T, \quad (12)$$

where M and \mathbf{q} are computed according to Theorem 2 but the cut-off values are set to u_n and l_n .

5 Empirical results

The performance of the estimators defined in Section 3 was evaluated with the help of Monte Carlo simulation. Bivariate observations (x_i, y_i) were generated according to the nominal model

$$\begin{aligned} y_i &= \theta_0 + x_i\theta_1 + \sigma e_i, \quad i = 1, \dots, n, \\ x_i &\sim N(0, 1), \quad e_i \sim F_0, \end{aligned}$$

with $\theta_0 = 0$, $\theta_1 = 1$, $\sigma = 1$. Both the standard Gaussian and the standard Log-Weibull error distributions (F_0) were considered. In the Gaussian case, the fixed cut-off value u was set to 2.5 so that the nominal fraction of retained observations was $F_0(u) - F_0(l) = 0.9876$. In the Log-Weibull case, the fixed cut-off values were $u = 1.8554$ and $l = -4.5277$ so that $f_0(u) = f_0(l)$ and $F_0(u) - F_0(l) = 0.9876$. The preliminary cut-off for the ATML-estimator was $\eta = \rho(u)$. Since the results for the Gaussian and the Log-Weibull case were very similar, only the latter are reported here. (Gervini and Yohai, 2002, report results for the Gaussian case and their adaptively truncated least squares estimator.)

Simulations at the nominal model. Table 1 gives simulated variances (multiplied by the sample size n) of intercept, slope, and scale estimates. As expected, both the TML- and ATML-estimates attain a much higher efficiency than the initial S-estimate and the variances of the adaptive estimates get close to the maximum likelihood values when n increases. However, the efficiency gain provided by adaptive versus fixed truncation does not appear to be appreciable unless n is large ($n \geq 200$).

Estimator : n	20	50	100	200	500	1000	∞	
intercept S	S	3.29	3.36	3.26	3.12	3.34	3.39	2.76
	TML	1.86	1.45	1.30	1.27	1.22	1.26	1.20
	ATML	1.71	1.45	1.28	1.25	1.17	1.17	1.11
	ML	1.24	1.17	1.16	1.17	1.13	1.16	1.11
slope S	S	4.15	3.69	3.68	3.60	3.24	3.22	2.84
	TML	2.20	1.55	1.34	1.24	1.10	1.11	1.10
	ATML	2.20	1.55	1.30	1.19	1.03	1.03	1.00
	ML	1.35	1.12	1.07	1.07	0.97	1.00	1.00
scale S	S	1.09	1.13	1.06	1.09	1.05	0.95	1.04
	TML	0.96	0.94	0.84	0.88	0.82	0.78	0.81
	ATML	0.96	0.94	0.83	0.84	0.71	0.67	0.61
	ML	0.69	0.65	0.61	0.66	0.58	0.61	0.61

Table 1. Simulated variances multiplied by sample size n at the nominal Log-Weibull model (2000 samples).

In addition, standard normal 95%-confidence intervals for intercept and slope were computed using the influence function estimates of variance (12). Table 2 gives the observed coverages: unless the sample size is very low ($n \leq 50$), the

coverages provided by the TML- and ATML-estimates are close to the nominal level.

n	S	TML	ATML	ML
20	88.7 / 84.4	84.4 / 76.8	84.5 / 74.9	91.0 / 82.1
50	90.4 / 89.1	91.0 / 86.6	91.0 / 86.5	94.3 / 89.5
100	91.7 / 89.9	91.9 / 89.9	92.2 / 90.2	93.7 / 91.0
200	92.7 / 91.2	94.4 / 92.9	94.2 / 92.9	94.8 / 93.0
500	92.9 / 93.7	95.0 / 94.7	95.1 / 94.9	95.0 / 95.0
1000	92.1 / 93.8	94.4 / 94.4	94.8 / 94.5	94.7 / 94.3

Table 2. Coverages (%) for intercept / slope at the nominal Log-Weibull model (2000 samples).

Simulations under point contamination. In order to assess the behaviour of the three estimators in the presence of outliers, we used contaminated simulated samples, where a fraction ϵ of pairs (x_i, y_i) was replaced with a fixed point (x_0, y_0) . In addition, we used the mean squared error of the fitted values as a criterion of goodness of fit. More precisely, for each simulated sample, we computed the estimates $\nu_i^{(k)} = \ln E(\exp(y)|x_i)^{(k)}$ of the conditional log-expectations $\nu_i = \ln E(\exp(y)|x_i)$ ($i = 1, \dots, n$) as well as the mean squared errors $MSE^{(k)} = (1/n) \sum_{i=1}^n (\nu_i^{(k)} - \nu_i)^2$ (where k indicates the estimator type). Table 3 gives the average mean squared errors based on 1000 samples of size 100 with $\epsilon = 10\%$, $x_0 = 1$, and various values of y_0 . The MSE-s of the TML- and the ATML-estimators are very close over the entire range of y_0 values.

y_0	S	TML	ATML	ML
0.0	0.062	0.051	0.050	0.050
1.0	0.056	0.025	0.024	0.019
2.0	0.148	0.040	0.039	0.031
3.0	0.242	0.080	0.078	0.067
3.4	0.272	0.102	0.100	0.088
4.0	0.247	0.129	0.128	0.118
5.0	0.203	0.182	0.182	0.177
6.0	0.159	0.219	0.224	0.235
7.0	0.118	0.244	0.255	0.303
8.0	0.093	0.222	0.240	0.362
9.0	0.075	0.210	0.233	0.425
10.0	0.071	0.167	0.189	0.485
15.0	0.062	0.049	0.055	0.799

Table 3. Average mean squared errors under point contamination at (x_0, y_0) ; $\epsilon = 10\%$, $x_0 = 1$, $n = 100$, 1000 samples.

The maximum average mean square error (maxMSE) of each estimator was estimated using a grid search. We observe that the maxMSE (indicated with

bold face characters in Table 3) of both the TML-and the ATML-estimators is smaller than the maxMSE of the initial S-estimator.

6 Example

We consider a sample of 100 patients hospitalized in a Swiss hospital during 1999 for “medical back problems”. We study the relationship between the cost of stay (Cost, in Swiss francs) and some explanatory variables that are available on administrative files: length of stay (LOS, in days), admission type (0 = planned, 1 = emergency), insurance type (0 = regular, 1 = private), age (years), sex (0 = female, 1 = male), discharge destination (1 = home, 0 = another health institution). This relationship is often used as a basis for reimbursement. We use the abbreviations $y = \log(\text{Cost})$, $x_1 = \log(\text{LOS})$, $x_2 = \text{admission type}$, $x_3 = \text{insurance type}$, $x_4 = \text{age}$, $x_5 = \text{sex}$, and $x_6 = \text{discharge destination}$.

The $\log(\text{LOS})/\log(\text{Cost})$ -plot of Figure 1, panel (a), suggests the tentative model

$$y = \mathbf{x}^T \boldsymbol{\theta} + e_i,$$

where the errors e_i are distributed according to a Log-Weibull distribution, $\mathbf{x}^T = (1, x_1, \dots, x_6)$, and $\boldsymbol{\theta}^T = (\theta_0, \theta_1, \dots, \theta_6)$. We observe a few mild outliers, but no leverage points. Table 4 shows the results provided by a TML-estimate of $\boldsymbol{\theta}$ (with fixed cut-off values $u = 1.8554$ and $l = -4.5277$), the maximum likelihood (ML) estimate, as well as the classical least squares (LS) estimate (as the most easily available procedure). The robust and classical coefficient estimates seem quite similar, but the classical estimates of scale are much larger than the robust estimate.

	TML			ML			LS		
	$\hat{\theta}_i$	st.err.	t	$\hat{\theta}_i$	st.err.	t	$\hat{\theta}_i$	st.err.	t
1	7.10	0.086	82.59	7.19	0.137	52.30	7.25	0.162	44.77
x_1	0.89	0.017	53.00	0.81	0.026	30.93	0.82	0.031	26.75
x_2	0.31	0.030	10.30	0.17	0.047	3.58	0.24	0.055	4.28
x_3	-0.06	0.049	-1.17	0.13	0.074	1.81	0.08	0.087	0.95
x_4	-0.00	0.001	-1.21	0.00	0.001	1.04	-0.00	0.001	-0.85
x_5	0.03	0.030	1.17	0.18	0.047	3.79	0.07	0.055	1.21
x_6	-0.07	0.040	-1.69	-0.06	0.065	-0.96	-0.11	0.076	-1.50
	scale estimate: 0.136			scale estimate: 0.208			scale estimate: 0.246		

Table 4. Full model: coefficient estimates, standard errors, and t -values provided by the TML-estimate (Log-Weibull errors), the maximum likelihood (ML) estimate and the least squares (LS) estimate.

The model can obviously be simplified by removing the non-significant effects

of variables x_3 , x_4 , x_5 , and x_6 . We obtain the results reported in Table 5.

	TML			ML			LS		
	$\hat{\theta}_i$	st.err.	t	$\hat{\theta}_i$	st.err.	t	$\hat{\theta}_i$	st.err.	t
1	7.00	0.049	144.02	7.35	0.077	94.91	7.12	0.081	88.32
x_1	0.89	0.017	51.49	0.80	0.029	27.72	0.82	0.030	27.28
x_2	0.35	0.029	11.96	0.15	0.049	3.14	0.26	0.051	5.17
	scale estimate: 0.139			scale estimate: 0.236			scale estimate: 0.247		

Table 5. Reduced model: coefficient estimates, standard errors, and t -values provided by the TML-estimate (Log-Weibull errors), the maximum likelihood (ML) estimate, and the least squares (LS) estimate.

The residual qq-plot of Figure 1, panel (b) indicates that the Log-Weibull model is an adequate description of the error distribution. (A similar analysis based on a Gaussian error model provides an inferior fit). We may therefore use the reduced model to estimate $\mu_{\mathbf{x}} = E(\exp(y) \mid \mathbf{x}) = \exp(\mathbf{x}^T \boldsymbol{\theta}) \Gamma(1 + \sigma)$, i.e., the expected cost for given $\mathbf{x} = (1, x_1, x_2)^T$. The estimate is

$$\hat{\mu}_{\mathbf{x}} = \exp(\mathbf{x}^T \hat{\boldsymbol{\theta}}) \Gamma(1 + \hat{\sigma}) \quad (13)$$

and a standard confidence interval for $\mu_{\mathbf{x}}$ can be computed using the following estimate of variance,

$$\text{Var}(\hat{\mu}_{\mathbf{x}}) \approx \hat{\mu}_{\mathbf{x}}^2 \mathbf{x}^T \text{Cov}(\mathbf{T}) \mathbf{x} + \hat{\mu}_{\mathbf{x}}^2 \dot{\Gamma}(1 + \hat{\sigma})^2 \text{Var}(S) + 2\hat{\mu}_{\mathbf{x}}^2 \dot{\Gamma}(1 + \hat{\sigma}) \text{Cov}((\mathbf{T}, S)),$$

which is obtained from the influence function of $\hat{\mu}_{\mathbf{x}}$. (The notations $\Gamma(\cdot)$ and $\dot{\Gamma}(\cdot)$ indicate the gamma and the digamma functions). Estimates and confidence intervals are reported in Figure 1, panel (c), for different values of LOS and admission type. Figure 1, panel (d) shows analogous estimates and confidence intervals based on the classical maximum likelihood estimate.

Figure 1 about here

In this example, an important effect of the outliers is to inflate the classical scale estimates. This effect is due to the few outliers marked with crosses in Figure 1, panel (a). If we remove these observations from the sample, the classical scale estimate becomes 0.134, which is very close to the robust estimate 0.139. Observe that the scale is not a nuisance parameter but, according to (13), is a main component of the conditional mean. In addition, the robust estimate of θ_2 (0.35) is markedly larger than the classical estimate (0.15). As a consequence, the robust conditional cost estimates of emergency cases are considerably higher than the classical ones, especially for large LOS values (Figure 1, panel (c) and panel (d)).

Appendix

Lemma 1. If $H \in \mathcal{H}_\varepsilon$, then $1 \geq \alpha(H) \geq 1 - \varepsilon$, and $\tilde{F}_H^+(z) \geq F_0^+(z)$ for $z \geq \eta$.

Proof. The lemma follows from the definition of $t(H)$, since $t(H) \leq \bar{t}(H)$.

Lemma 2. Let $\varepsilon < \varepsilon_0^*$ and assume that F_0^+ has a finite mean. Then

$$\sup_{H \in \mathcal{H}_\varepsilon} L(H, \mathbf{T}^{(0)}(H), S^{(0)}(H)) < \infty.$$

Proof. For any $H \in \mathcal{H}_\varepsilon$,

$$L(H, \mathbf{T}^{(0)}(H), S^{(0)}(H)) = \int_0^{t(H)} z d\tilde{F}_H^+(z) + \beta(H) \ln(S^{(0)}(H)).$$

Since $t(H) \geq \eta$, using Lemma 1:

$$\begin{aligned} L(H, \mathbf{T}^{(0)}(H), S^{(0)}(H)) &= \int_0^\eta z d\tilde{F}_H^+(z) + \int_\eta^{t(H)} z d\tilde{F}_H^+(z) + \beta(H) \ln(S^{(0)}(H)) \\ &\leq \eta + \int_\eta^\infty z dF_0^+(z) + \beta(H) \ln(\sigma_2(\varepsilon)) < \infty. \end{aligned}$$

Proof of Theorem 1. Without loss of generality, we assume that $\boldsymbol{\theta} = 0$ and $\sigma = 1$. First, we prove that

$$\sup_{H \in \mathcal{H}_\varepsilon} S(H) < \infty.$$

In fact, if this was not true, there would exist a sequence $H_n \in \mathcal{H}_\varepsilon$ such that $\lim_{n \rightarrow \infty} S(H_n) = \infty$. Since $t(H_n) > \eta$, it follows that $\beta(H_n) > 0$ and, therefore,

$$\lim_{n \rightarrow \infty} L(H_n, \mathbf{T}^{(0)}(H_n), S^{(0)}(H_n)) \geq \beta(H_n) \ln(S(H_n)) = \infty.$$

This, together with the Lemma 2, would contradict the definition of the initial estimates. We now show that

$$\sup_{H \in \mathcal{H}_\varepsilon} \mathbf{T}(H) < \infty.$$

Suppose this is not true; then, there exists a sequence $H_n \in \mathcal{H}_\varepsilon$ such that $\lim_{n \rightarrow \infty} \mathbf{T}(H_n) = \infty$. Without loss of generality, we assume that $\mathbf{t}_{0n} = \mathbf{T}^{(0)}(H_n) \rightarrow \mathbf{t}_0$, $\mathbf{t}_n = \mathbf{T}(H_n)/c_n \rightarrow \mathbf{t}$ with $c_n = \|\mathbf{T}^{(0)}(H_n)\|$, and that $z_{0n} = S^{(0)}(H_n) \rightarrow z_0$, $z_n = S(H_n) \rightarrow z_1$, $t_n = t(H_n) \rightarrow t_0$ (t_0 may be ∞), $\alpha_n = \alpha(H_n) \rightarrow \alpha_0$, $\beta_n = \beta(H_n) \rightarrow \beta_0$. Note that $z_0 > 0$ thanks to the assumptions. Then,

$$E_{H_n}[w_{H_n}(\mathbf{x}, y) \rho((y - \mathbf{x}^T \mathbf{T}(H_n))/S(H_n))] \geq (1 - \varepsilon) E_{H_0}[R_n(\mathbf{x}, y)], \quad (14)$$

where

$$R_n(\mathbf{x}, y) = \rho((y - c_n \mathbf{x}^T \mathbf{t}_n)/z_n) I[\rho((y - \mathbf{x}^T \mathbf{t}_{0n})/z_{0n}) < t_n].$$

Let

$$V_n(\mathbf{x}) = E_{H_0}[R_n(\mathbf{x}, y) \mid \mathbf{x}]; \quad (15)$$

we show that, if $\mathbf{x}^T \mathbf{t}_0 \neq 0$,

$$\lim_{n \rightarrow \infty} \frac{1 - \varepsilon}{\alpha_n} V_n(\mathbf{x}) + \beta_n \ln(S(H_n)) \rightarrow \infty. \quad (16)$$

Consider a fix \mathbf{x} such that $\mathbf{x}^T \mathbf{t}_0 \neq 0$ and let

$$a_n = P_{H_0}[\rho((y - \mathbf{x}^T \mathbf{t}_{0n})/z_{0n}) < t_n \mid \mathbf{x}].$$

Then

$$\liminf_{n \rightarrow \infty} a_n \geq a = P_{H_0}[\rho((y - \mathbf{x}^T \mathbf{t}_0)/z_0) < t_0 \mid \mathbf{x}] > 0.$$

We first consider the case $z_1 > 0$. Let $K > 0$ be arbitrarily large and k_1 such that $|z| \geq k_1$ implies $\rho(z) > 2K/a$ and let k_2 such that $P_{F_0}(y < k_2) \geq 1 - a/4$. Let n_0 be such that $|c_n \mathbf{x}^T \mathbf{t}_n| \geq k_1 z_n + k_2$, and $a_n \geq (3/4)a$. Then, for $n \geq n_0$,

$$R_n(\mathbf{x}, y) \geq (2K/a) I[\{\rho((y - \mathbf{x}^T \mathbf{t}_{0n})/z_{0n}) < t_n\} \cap \{y < k_2\}].$$

Since $P_{H_0}[\{\rho((y - \mathbf{x}^T \mathbf{t}_{0n})/z_{0n}) < t_n\} \cap \{y < k_2\} \mid \mathbf{x}] \geq a/2$, we get $V_n(\mathbf{x}) \geq K$. But then, $\lim_{n \rightarrow \infty} V_n(\mathbf{x}) = \infty$ and (16) follows.

We now consider the case $z_1 = 0$. Let $k_1 > 0$ and $\gamma > 0$ be such that $\rho(z) \geq \gamma|z|$ for $|z| \geq k_1$ and let k_2 be such that $P_{F_0}(y < k_2) \geq 1 - a/4$. Let n_0 be such that $|c_n \mathbf{x}^T \mathbf{t}_n| \geq k_1 + k_2$, $z_n \leq 1$ and $a_n \geq (3/4)a$. Then, for $n \geq n_0$,

$$V_n \geq \frac{\gamma k_1}{z_n} P_{H_0}[\{\rho((y - \mathbf{x}^T \mathbf{t}_{0n})/z_{0n}) < t_n\} \cap \{y < k_2\} \mid \mathbf{x}].$$

Since $P_{H_0}[\{\rho((y - \mathbf{x}^T \mathbf{t}_{0n})/z_{0n}) < t_n\} \cap \{y < k_2\} \mid \mathbf{x}] \geq a/2$, we get that, for $n \geq n_0$,

$$\frac{(1 - \varepsilon)}{\alpha_n} V_n(\mathbf{x}) + \beta_n \log S(H_n) \geq \frac{\delta}{z_n} + \beta_n \log(z_n) = \frac{\delta + z_n \beta_n \log(z_n)}{z_n}, \quad (17)$$

where $\delta = (1 - \varepsilon)k_1 \gamma a/2$. By the l'Hopital rule, $\lim_{z \rightarrow 0} z \ln(z) = 0$, and then, using (17), we get (16). Since $P(\mathbf{x}^T \mathbf{t}_0 \neq 0) > 0$, by (14), (15), and (16) we have

$$\begin{aligned} \lim_{n \rightarrow \infty} L(H_n, \mathbf{T}(H_n), S(H_n)) &\geq \\ \lim_{n \rightarrow \infty} E_{H_n} \left[\frac{(1 - \varepsilon)}{\alpha_n} V_n(\mathbf{x}) + \beta_n \log(S(H_n)) \right] &= \infty. \end{aligned}$$

Therefore, there exists an integer m such that $L(H_m, \mathbf{T}(H_m), S(H_m)) > L(H_m, \mathbf{T}^{(0)}(H_m), S^{(0)}(H_m))$, contradicting the definition of (\mathbf{T}, S) .

Lemma 3. The influence function of $\alpha(H)$ at the model H_0 is given by

$$I_\alpha^0 = \min \left[\inf_{r \geq \eta} \left(\frac{1}{\sigma} \frac{\gamma(r)}{F_0^+(r)} I_{S^{(0)}}^0 + \frac{\Delta \rho_0(r)}{F_0^+(r)} - 1 \right), 0 \right],$$

where $\rho_0 = \rho((y_0 - \mathbf{x}_0^T \boldsymbol{\theta})/\sigma)$ and $\gamma(r) = u(r)f_0(u(r)) - l(r)f_0(l(r))$.

Proof. Let $H_\epsilon = (1 - \epsilon)H_0 + \Delta_{\mathbf{x}_0, y_0}$, where $\Delta_{\mathbf{x}_0, y_0}$ denotes the cdf of a point mass at (\mathbf{x}_0, y_0) . We have

$$F_{H_\epsilon}^+(r) = (1 - \epsilon)P_{H_0}(\mathcal{A}) + \epsilon\Delta_{\mathbf{x}_0, y_0}(\mathcal{A}),$$

where

$$\mathcal{A} = \{l(r)S^{(0)}(H_\epsilon)/\sigma + \mathbf{x}^T(\mathbf{T}^{(0)}(H_\epsilon) - \boldsymbol{\theta})/\sigma \leq z \leq u(r)S^{(0)}(H_\epsilon)/\sigma + \mathbf{x}^T(\mathbf{T}^{(0)}(H_\epsilon) - \boldsymbol{\theta})/\sigma\}$$

and $z = (y - \mathbf{x}^T \boldsymbol{\theta})/\sigma$. Since $f_0(u(t)) = f_0(l(t))$, we obtain

$$P_{H_0}(\mathcal{A}) = F_0(u(r)) - F_0(l(r)) + \frac{\epsilon}{\sigma} [u(r)f_0(u(r)) - l(r)f_0(l(r))] I_{S^{(0)}}^0 + o(\epsilon^2).$$

It follows that

$$\frac{F_{H_\epsilon}^+(r)}{F_0^+(r)} = 1 + \epsilon \left[\frac{1}{\sigma} \frac{\gamma(r)}{F_0^+(r)} I_{S^{(0)}}^0 + \epsilon \frac{\Delta_{\rho_0}(r)}{F_0^+(r)} - 1 \right] + o(\epsilon^2),$$

where $\gamma(r) = u(r)f_0(u(r)) - l(r)f_0(l(r))$ and $\Delta_{\rho_0}(r) = \Delta_{x_0, y_0}(\mathcal{A})$ with $\rho_0 = \rho((y_0 - \mathbf{x}_0^T \boldsymbol{\theta})/\sigma)$. The lemma easily follows.

Proof of theorem 3. We first suppose that $\tau(H)$ is any cut-off value depending on H , that $\alpha(H) = F_H^+(\tau(H))$ and that $f_H^+(\tau(H)) > 0$. Inserting $H_\epsilon = (1 - \epsilon)H + \epsilon\Delta_{\mathbf{x}_0, y_0}$ for H into the defining equations of (\mathbf{T}, S) and taking the derivative with respect to ϵ at $\epsilon = 0$, we obtain the system

$$A_1(H)I_{\mathbf{T}} + \mathbf{b}_1(H)I_S = \mathbf{c}_1(H) + \mathbf{d}_1(H), \quad (18)$$

$$\mathbf{a}_2^T(H)I_{\mathbf{T}} + b_2(H)I_S = c_2(H) + d_2(H) - \alpha(H)I_\beta - \beta(H)I_\alpha, \quad (19)$$

where

$$A_1(H) = E_H [\omega(\mathbf{x}, y, H) s_1'(z) \mathbf{x} \mathbf{x}^T] / S(H), \quad (20)$$

$$\mathbf{a}_2(H) = E_H [\omega(\mathbf{x}, y, H) s_2'(z) \mathbf{x}] / S(H), \quad (21)$$

$$\mathbf{b}_1(H) = E_H [\omega(\mathbf{x}, y, H) s_1'(z) z \mathbf{x}] / S(H), \quad (22)$$

$$b_2(H) = E_H [\omega(\mathbf{x}, y, H) s_2'(z) z] / S(H), \quad (23)$$

$$\mathbf{c}_1(H) = w(\rho(z_0) / \tau(H)) s_1(z_0) \mathbf{x}_0, \quad (24)$$

$$c_2(H) = w(\rho(z_0) / \tau(H)) s_2(z_0) - \alpha(H)\beta(H) \quad (25)$$

$$\mathbf{d}_1(H) = \left[\frac{\partial}{\partial \epsilon} E_H [\omega(\mathbf{x}, y, H_\epsilon) s_1(z) \mathbf{x}] \right]_{\epsilon=0},$$

$$d_2(H) = \left[\frac{\partial}{\partial \epsilon} E_H [\omega(\mathbf{x}, y, H_\epsilon) s_2(z)] \right]_{\epsilon=0},$$

z and z_0 are abbreviations of $(y - \mathbf{x}^T \mathbf{T}(H))/S(H)$ and $(y_0 - \mathbf{x}_0^T \mathbf{T}(H))/S(H)$, respectively, and

$$\omega(\mathbf{x}, y, H) = w \left(\frac{1}{\tau(H)} \rho \left(\frac{y - \mathbf{x}^T \mathbf{T}^{(0)}(H)}{S^{(0)}(H)} \right) \right).$$

Since $w(z) = I(|z| < 1)$, and assuming that differentiation and integration with respect to G can be interchanged, we have:

$$\begin{aligned} \mathbf{d}_1(H) &= E_G \left[\left[\frac{\partial}{\partial \epsilon} \int_{L(H_\epsilon)}^{U(H_\epsilon)} s_1 \left(\frac{y - \mathbf{x}^T \mathbf{T}(H)}{S(H)} \right) dF(y|\mathbf{x}) \right]_{\epsilon=0} \mathbf{x} \right], \\ d_2(H) &= E_G \left[\left[\frac{\partial}{\partial \epsilon} \int_{L(H_\epsilon)}^{U(H_\epsilon)} s_2 \left(\frac{y - \mathbf{x}^T \mathbf{T}(H)}{S(H)} \right) dF(y|\mathbf{x}) \right]_{\epsilon=0} \right], \end{aligned}$$

where $F(y|\mathbf{x})$ denotes the conditional distribution of y for given \mathbf{x} (with density $f(y|\mathbf{x})$) and

$$\begin{aligned} U(H) &= \mathbf{x}^T \mathbf{T}^{(0)}(H) + u(\tau(H))S^{(0)}(H), \\ L(H) &= \mathbf{x}^T \mathbf{T}^{(0)}(H) + l(\tau(H))S^{(0)}(H). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbf{d}_1(H) &= E_G [(s_1(z_U) f(U(H)|\mathbf{x}) I_U - s_1(z_L) f(L(H)|\mathbf{x}) I_L) \mathbf{x}], \\ d_2(H) &= E_G [s_2(z_U) f(U(H)|\mathbf{x}) I_U - s_2(z_L) f(L(H)|\mathbf{x}) I_L], \end{aligned}$$

where $z_U = [U(H) - \mathbf{x}^T \mathbf{T}(H)] / S(H)$, $z_L = [L(H) - \mathbf{x}^T \mathbf{T}(H)] / S(H)$. Moreover,

$$\begin{aligned} I_U &= \mathbf{x}^T I_{\mathbf{T}^{(0)}} + u(\tau(H)) I_{S^{(0)}} + u'(\tau(H)) S^{(0)}(H) I_\tau, \\ I_L &= \mathbf{x}^T I_{\mathbf{T}^{(0)}} + l(\tau(H)) I_{S^{(0)}} + l'(\tau(H)) S^{(0)}(H) I_\tau, \end{aligned}$$

and $I_\tau = IF(y_0, \tau, H)$ is easily obtained (Huber, 1981, p. 56).

Suppose now that $h = f((y - \mathbf{x}^T \boldsymbol{\theta} / \sigma) g(\mathbf{x}))$, that $\mathbf{T}^{(0)}(H) = \mathbf{T}(H) = \boldsymbol{\theta}$, $S^{(0)}(H) = S(H) = \sigma$ and that $f^+(\tau(H)) > 0$. We use the following abbreviations: $\tau^* = \tau(H)$, $u^* = u(\tau^*)$, $l^* = l(\tau^*)$, $\hat{u} = u'(\tau^*)$, $\hat{l} = l'(\tau^*)$. We have $z_U = u^*$, $z_L = l^*$, $U(H) = \mathbf{x}^T \boldsymbol{\theta} + u^* \sigma$, $L(H) = \mathbf{x}^T \boldsymbol{\theta} + l^* \sigma$, $I_U = \mathbf{x}^T I_{\mathbf{T}^{(0)}} + u^* I_{S^{(0)}} + \hat{u} \sigma I_\tau$, $I_L = \mathbf{x}^T I_{\mathbf{T}^{(0)}} + l^* I_{S^{(0)}} + \hat{l} \sigma I_\tau$ and obtain

$$\mathbf{d}_1(H) = E_G [(k_{1,1}(\tau^*) + k_{2,1}(\tau^*) + k_{3,1}(\tau^*)) \mathbf{x}], \quad (26)$$

$$d_2(H) = E_G [k_{1,2}(\tau^*) + k_{2,2}(\tau^*) + k_{3,2}(\tau^*)], \quad (27)$$

where, for $j = 1, 2$,

$$k_{1,j}(\tau^*) = \frac{1}{\sigma} f(u^*) [s_j(u^*) - s_j(l^*)] \mathbf{x}^T I_{\mathbf{T}^{(0)}}, \quad (28)$$

$$k_{2,j}(\tau^*) = \frac{1}{\sigma} f(u^*) [u^* s_j(u^*) - l^* s_j(l^*)] I_{S^{(0)}}, \quad (29)$$

$$k_{3,j}(\tau^*) = f(u^*) [\hat{u} s_j(u^*) - \hat{l} s_j(l^*)] I_\tau. \quad (30)$$

Finally, we assume that $h(\mathbf{x}, y) = h_b(\mathbf{x}, y) = f_{0,b}((y - \mathbf{x}^T \boldsymbol{\theta})/\sigma)g(\mathbf{x})$, where $f_{0,b}$ denotes the error distribution model truncated at $l(b)$ and $u(b)$ and b is sufficiently large to have $\mathbf{T}^{(0)}(H_b) = \boldsymbol{\theta}$ and $S^{(0)}(H_b) = \sigma$. It follows that $\mathbf{T}(H_b) = \boldsymbol{\theta}$ and $S(H_b) = \sigma$. Moreover, we consider the cut-off value $\tau(H) = t(H) - a$, where $t(H)$ is defined in Section 5 and a is an arbitrarily small positive number. Thus, $\tau^* = \tau(H_b) = b - a$, $f_{0,b}(z) = f_0(z)$ for $z \in [l(b - a), u(b - a)]$, and $f_0^+(\tau^*) > 0$. Therefore, the assumptions of the preceding paragraph are satisfied and I_τ is well defined for $H = H_b$. We obtain

$$\begin{aligned} A_{1,a}(H_b)I_{\mathbf{T}} + \mathbf{b}_{1,a}(H_b)I_S &= \mathbf{c}_{1,a}(H_b) + \\ &E_G [(k_{1,1}(b - a) + k_{2,1}(b - a) + k_{3,1}(b - a))\mathbf{x}], \\ \mathbf{a}_{2,a}^T(H_b)I_{\mathbf{T}} + b_{2,a}(H_b)I_S &= c_{2,a}(H_b) + \\ &E_G [k_{1,2}(b - a) + k_{2,2}(b - a) + k_{3,2}(b - a)] \\ &- \alpha_a(H_b)I_\beta - \beta_a(H_b)I_\alpha. \end{aligned}$$

where the suffix a indicates quantities that depend on a , the suffix b indicates quantities that depend on $f_{0,b}$, $f = f_0$ in $k_{i,j}$, and the influence functions depend on $H = H_b$. Deriving $f_0(u(z)) = f_0(l(z))$ at $z = b - a$, we get $u'(z)s_1(z) - l'(z)s_1(z) = 0$ and thus $k_{3,1}(b - a) = 0$. Moreover, $\alpha_a(H_b) = \beta_{2,a}(H_b)$. Therefore,

$$\begin{aligned} A_{1,a}(H_b)I_{\mathbf{T}} + \mathbf{b}_{1,a}(H_b)I_S &= \mathbf{c}_{1,a}(H_b) + E_G [(k_{1,1}(b - a) + k_{2,1}(b - a))\mathbf{x}], \\ \mathbf{a}_{2,a}^T(H_b)I_{\mathbf{T}} + b_{2,a}(H_b)I_S &= c_{2,a}(H_b) + E_G [k_{1,2}(b - a) + k_{2,2}(b - a)] + \\ &\frac{\beta_{1,a}(H_b)}{\beta_{2,a}(H_b)} [I_{\alpha_a} + \alpha_a(H_b) - \Delta_{y_0}(b - a)] \\ &- \beta_a(H_b)I_{\alpha_a}. \end{aligned}$$

Taking the limit for $a \rightarrow 0$, we obtain a system for $I_{\mathbf{T}}$ and I_S when $H = H_b$ and $\tau(H) = t(H)$. Letting $b \rightarrow \infty$, we have $k_{1,1}(b) \rightarrow 0$, $k_{2,1}(b) \rightarrow 0$, $k_{1,2}(b) \rightarrow 0$, $k_{2,2}(b) \rightarrow 0$, $\beta_{1,0}(H_b) \rightarrow 1$, $\beta_{2,0}(H_b) \rightarrow 1$, $\beta_0(H_b) \rightarrow 1$, $\alpha_0(H_b) \rightarrow 1$, $\Delta_{y_0}(b) \rightarrow 1$; in the limit, the system defines the influence functions of the maximum likelihood estimates.

Proof of theorem 2. The influence function of the TML-estimator at the model H_0 can be derived by specializing the proof of Theorem 3. More precisely, it is defined by equations (18)-(19), where $I_\beta = 0$ and A_1 , \mathbf{a}_2 , \mathbf{b}_1 , b_2 , \mathbf{c}_1 , c_2 are given by equations (20)-(25) with $t(H) = \eta$. In addition, \mathbf{d}_1 and d_2 are given by (26)-(30) with $k_{3,1} = k_{3,2} = 0$. Finally, the influence function of $\alpha(H) = F_H^+(\eta)$ can be easily computed; we find

$$IF((\mathbf{x}_0, y_0), F_H^+(\eta), H_0) = \frac{1}{\sigma} \gamma(\eta) I_{S^{(0)}}^0 + \Delta_{\rho_0}(\eta) - F_0^+(\eta),$$

where $\rho_0 = \rho((y_0 - \mathbf{x}_0^T \boldsymbol{\theta})/\sigma)$ and $\gamma(r) = u(r)f_0(u(r)) - l(r)f_0(l(r))$.

Acknowledgement. This work was supported in part by grant 21-54146.98 from the Swiss National Science Foundation.

References

- Bednarski T., Clarke B.R., 1993. Trimmed likelihood estimation of location and scale of the normal distribution. *Australian Journal of Statistics*, 35(2), 141-153.
- Carroll R.J., Ruppert D., 1988. Transformation and weighting in regression. Chapman and Hall, New York.
- Cantoni E., Ronchetti E., 2001. Robust inference for generalized linear models. *Journal of the American Statistical Association*. (In press).
- Clarke B.R., 2000. An adaptive method of estimation and outlier detection in regression applicable for small to moderate sample sizes. *Probability and Statistics*, 20, 25-50.
- Field C., Smith B., 1994. Robust estimation – A weighted maximum likelihood approach. *International Statistical Review*, 62(3), 405-424.
- Gervini D., Yohai V.J., 2002. A class of robust and fully efficient regression estimates. *The Annals of Statistics*, 30(2), 583-616.
- Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A., 1986. Robust statistics: An approach based on the influence function. Wiley, New York.
- Huber P., 1981. Robust Statistics. Wiley, New York.
- Hettmansperger T.P., McKean J.W., 1998. Robust nonparametric statistical methods. Arnold, London.
- Marazzi A., Ruffieux C., 1996. Implementing M-estimators of the gamma distribution. In: H. Rieder (Ed.), Robust Statistics, Data Analysis, and Computer Intensive Methods, In Honor of Peter Huber's 60th Birthday, Lecture Notes in Statistics, 109, Springer Verlag, Heidelberg.
- Marazzi A., Ruffieux C., 1999. The truncated mean of an asymmetric distribution. *Computational Statistics and Data Analysis*, 32(1), 79-100.
- Markatou, M. Basu, A., Lindsay, B.G., 1998. Weighted likelihood equations with bootstrap search. *Journal of the American Statistical Association*, 93, 740-750.
- McCullagh P., Nelder J.A., 1989. Generalized linear models. Second Edition. Chapman and Hall, New York.
- Rousseeuw P. J., Leroy A.M, 1987. Robust regression and outlier detection. Wiley, New York.
- Victoria-Feser M.P., Ronchetti E., 1994. Robust methods for personal-income distribution models. *The Canadian Journal of Statistics*, 22, 247-258.
- Victoria-Feser M.P., Ronchetti E., 1997. Robust estimation for grouped data. *Journal of the American Statistical Association*, 92(437), 333-340.
- Williams M.S., 1997. A regression technique accounting for heteroscedastic and asymmetric errors. *Journal of Agricultural, Biological, and Environmental Statistics*, 2 (1), 108-129.

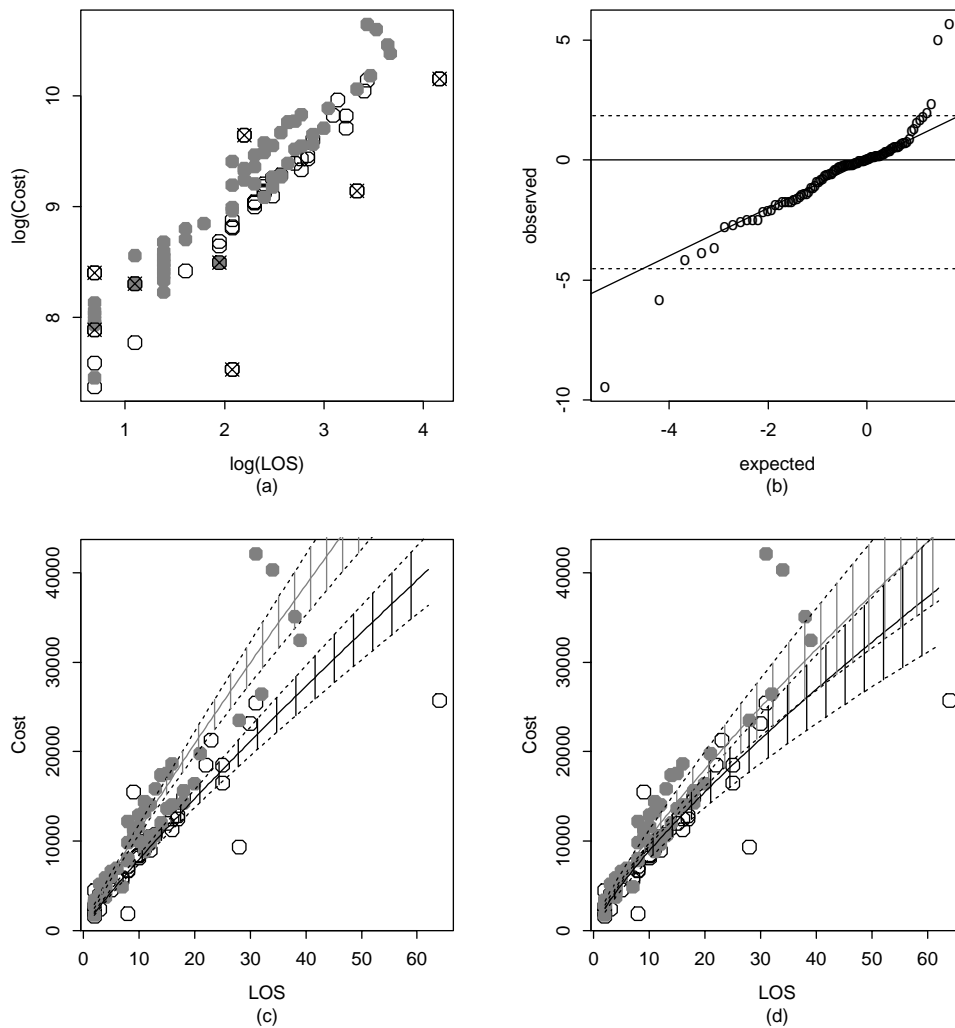


Figure 1. Panel (a): $\log(\text{LOS})$ and $\log(\text{Cost})$ of 100 patients of a Swiss hospital; circles denote planned admissions, bullets denote emergency admissions, crosses denote outliers such that $w_i = 0$. Panel (b): residual qq-plot provided by a TML-estimate for Log-Weibull errors; the dotted horizontal lines correspond to the cut-off values. Panel (c): robust estimates and confidence intervals for the expected cost of planned admissions (full line) and emergency admissions (broken line). Panel (d): maximum likelihood estimates and 99%-confidence intervals for the expected cost of planned admissions (full line) and emergency admissions (broken line).