Regression with response distributions of Pareto-type

Jan Beirlant\textsuperscript{a}, Yuri Goegebeur\textsuperscript{b,*,1}

\textsuperscript{a}Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium
\textsuperscript{b}Department of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

Received 1 April 2001; received in revised form 1 March 2002

Abstract

The estimation of the Pareto index in presence of covariate information is discussed. The Pareto index is modelled as a function of the explanatory variables and hence measures the tail heaviness of the conditional distribution of the response variable given this covariate information. The original response data are transformed in order to obtain generalized residuals, possessing a common Pareto-type distribution. An exponential regression model will be developed for these generalized residuals. The parameters of this model are estimated using a profile likelihood method. The resulting maximum likelihood estimates of the regression coefficients can be used for the estimation of extreme quantiles of the conditional distribution of the dependent variable. The methods developed are illustrated with two practical examples. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Extreme-value index; Pareto quantile plot; Generalized residuals; Exponential regression model; Profile likelihood function

1. Introduction

We consider distributions of Pareto-type, i.e. distribution functions $F$ for which there exists a positive constant $\gamma$ such that

$$1 - F(y) = y^{-1/\gamma} \tilde{I}(y), \quad y > 0$$

(1)

\textsuperscript{*} Corresponding author. Tel.: +32-16-32-67-62; fax: +32-16-32-67-32.
\textit{E-mail addresses:} jan.beirlant@wis.kuleuven.ac.be (J. Beirlant), yuri.goegebeur@econ.kuleuven.ac.be (Y. Goegebeur).
\textsuperscript{1} Postdoctoral researcher, Fund for Scientific Research—Flanders.

0167-9473/03/$-see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-9473(02)00120-2
with $\tilde{l}$ denoting a slowly varying function at infinity:
\[
\frac{\tilde{l}(\lambda y)}{\tilde{l}(y)} \to 1 \quad \text{as} \quad y \to \infty \quad \text{for all} \quad \lambda > 0.
\] (2)

The parameter $\gamma$, called the extreme-value index, measures the tail heaviness of the distribution function $F$. Defining the tail quantile function $U$ as $U(x) = \inf \{y: F(y) \geq 1 - 1/x\}$, $x > 1$, model (1) can be stated equivalently as
\[
U(x) = x^\gamma l(x)
\] (3)

with $l$ again a slowly varying function at infinity.

The estimation of $\gamma$ based on an i.i.d. sample of size $n$ has received considerable attention in the extreme-value literature. In case $1 - F(y) \sim cy^{-\gamma}$ as $y \to \infty$, it is obvious that estimators for $\gamma$ in general should be based on the upper order statistics. For an up to date overview we refer to Csörgő and Viharos (1998), Beirlant et al. (1999), Feuerverger and Hall (1999), Feuerverger and Hall (1999) among others.

The fit of a statistical model to a random sample $Y_1, \ldots, Y_n$ can be visually assessed by inspection of a quantile–quantile or QQ plot. Since log-transformed Pareto distributed random variables (with $l(x) = 1$) are exponentially distributed, the coordinates of the points on a Pareto quantile plot are given by
\[
\left(\log \frac{n + 1}{j}, \log Y_{n-j+1,n}\right), \quad j = 1, \ldots, n
\]
with $Y_{1,n} \leq \cdots \leq Y_{n,n}$ denoting the order statistics associated with $Y_1, \ldots, Y_n$. In case of a good fit of the strict Pareto distribution to the data, the Pareto quantile plot should be approximately linear. Moreover, then the slope of a line fitted through the points on the QQ plot passing through the origin will estimate $\gamma$. In fact, one can show that this method corresponds to the maximum likelihood procedure in this specific case. For Pareto-type distributions $\log l(y)/\log y \to 0$ as $y \to \infty$, which implies that for Pareto-type distributed data the QQ plot will be ultimately linear. Again the slope of the linear part will approximate $\gamma$. Several of the better known estimators for $\gamma$ can be seen as estimators of the slope of the linear part of the Pareto quantile plot. Hill (1975) introduced the following popular estimator for $\gamma$:
\[
H_{k,n} = \frac{1}{k} \sum_{j=1}^{k} \log Y_{n-j+1,n} - \log Y_{n-k,n}, \quad k = 1, \ldots, n - 1.
\]

Clearly, the Hill estimator measures the average increase of the Pareto quantile plot to the right of an anchor point ($\log((n + 1)/(k + 1))$, $\log Y_{n-k,n}$), and hence can be interpreted as a slope estimator. Other well-known alternative estimators for $\gamma$ are the moment estimator (Dekkers et al., 1989) and the kernel estimators (Csörgő et al., 1985).

Next to the above introduced estimators of $\gamma$ based on Pareto quantile plots, the real-valued index $\gamma$ can also be estimated in a parametric way. These estimators follow
directly from the limit theorems that form the core of extreme-value theory. A first possibility is based on the following result of Fisher and Tippett (1928). If for a distribution function $F$ there exist sequences of constants $(a_n > 0)_{n}$ and $(b_n)_{n}$ such that

$$\lim_{n \to \infty} P \left( \frac{Y_{n,n} - b_n}{a_n} \leq y \right) = \lim_{n \to \infty} F^n(a_n y + b_n) = H(y)$$

(4)

at all continuity points of $H$, with $H$ a nondegenerate distribution, then $F$ is said to belong to the domain of (maximum) attraction of $H$. Moreover, it is known that if such a nondegenerate limit distribution $H$ exists, it should be of the form

$$H(y; \mu, \sigma, \gamma) = \begin{cases} \exp(-(1 + \gamma(y - \mu)/\sigma)^{-1/\gamma}), & 1 + \gamma[(y - \mu)/\sigma] > 0, \gamma \neq 0, \\ \exp(-(-\gamma)\mu), & y \in \mathbb{R}, \gamma = 0 \end{cases}$$

(5)

with $\mu \in \mathbb{R}$ and $\sigma > 0$. This limit distribution is the so-called generalized extreme-value distribution (GEV). In this paper, the domain of attraction of the GEV will be denoted by $\mathcal{D}$. According to the sign of $\gamma$ three subclasses of $\mathcal{D}$ can be distinguished: $\mathcal{D}_-, \mathcal{D}_0$ and $\mathcal{D}_+$. Based on this result, $\gamma$ can be estimated by fitting $H$ to sample maxima (Gumbel, 1960). The parameters of the GEV can be estimated with the maximum likelihood method (Prescott and Walden, 1980; Smith, 1985) or with the method of (probability-weighted) moments (Hosking et al., 1985). This classical approach has received considerable criticism, mainly because of its inefficient use of the available data. This problem was remedied by using several of the largest order statistics, i.e. by using the observations that exceed a certain specified threshold. This is the so-called peaks over threshold (POT) method. According to this approach the generalized Pareto distribution (GPD), given by

$$G(y; \sigma, \gamma) = \begin{cases} 1 - (1 + \gamma y/\sigma)^{-1/\gamma} & \text{if } \gamma \neq 0, \\ 1 - \exp(-\gamma) & \text{if } \gamma = 0 \end{cases}$$

(6)

with $\sigma > 0$ and with $y > 0$ if $\gamma > 0$, $0 < y < -\sigma/\gamma$ if $\gamma < 0$, is fitted to exceedances over a specified threshold (Pickands, 1975). Parameters can be estimated with the maximum likelihood method (Smith, 1985, 1987), the method of (probability-weighted) moments (Hosking and Wallis, 1987) or the elemental percentile method (Castillo and Hadi, 1997).

In this paper we discuss the estimation of the extreme value-index $\gamma$ (in case $\gamma > 0$) when covariate information is available. We assume that with each random variable $Y_i$, a $(p + 1)$-dimensional vector of explanatory variables $x_i$ is associated. The response variables are considered independent but not identically distributed. In generalized linear models dependence of the distribution of the response variable on the explanatory variables is modelled through the mean of the distribution of the dependent variable (McCullagh and Nelder, 1992). For heavy tailed distributions, the mean does not always exist. In this case univariate models are extended to regression models by allowing one or more of their parameters to vary with $x$, an approach that is common in the survival analysis literature (see for instance Lawless, 1982). For an actuarial application of a
procedure involving the Burr distribution (being a special case of model (1)) we refer to Beirlant et al. (1998). From an extreme-value point of view we are interested in the estimation of the tail index and extreme high quantiles of the conditional distributions, and hence it is natural to model the extreme-value index as a function of $x$. The regression case has received much less attention in the extreme-value literature. Attempts to estimate $\gamma$ in presence of covariate information have been constructed from a purely parametric point of view. One possible approach consists of extending the GEV (5) to a regression model by taking one or more of its parameters as a function of the independent variables and the fitting of this regression model to sample maxima. An alternative approach is based on fitting a GPD regression model (using (6)) to exceedances over a specified threshold. The basic reference here is Davison and Smith (1990).

The approach presented in this paper consists of an extension of the method developed in Beirlant et al. (1999) and Feuerverger and Hall (1999).

Applications of model (1) can be found in scientific disciplines such as finance, insurance, reliability theory, telecommunication, environmetrics, geology and climatology. Here we motivate the presented model with two practical case studies.

The first dataset contains claim data from a Norwegian fire insurance portfolio. Claim values (in Krones) are available for the period 1988–1991 (source: Beirlant et al., 1996a). These values have to be multiplied by 1000 to obtain their real value in Krones. The data are shown in Fig. 1. The ultimate goal of the analysis is to describe how the tail of the claim size distribution varies over time. Here, we give a visual indication about the appropriateness of model (1) for these data. In Fig. 2 we show the Pareto quantile plots of the claim sizes for the different years. Clearly, all quantile plots are approximately linear in the largest observations. Further, the slopes of these
linear parts show that the tail heaviness of the conditional claim size distributions decreases over time. Remark that the extremely large claim in 1988 does not really influence this inference. Deleting this value does not change the results to a large extent. Further, note that this example could also be treated using an ANOVA-type approach, but in order to gain degrees of freedom we prefer a continuous regression.

Our second example comes from the geostatistical context and concerns the valuation of diamonds. The dataset contains measurements on the variables size (in carat) and value (in USD) for 1914 diamonds obtained from a kimberlite deposit. For this example the aim is to examine how the tail of the distribution of the variable value depends on the variable size. The value versus log(size) scatterplot is given in Fig. 3. Here, we comment on the appropriateness of model (1) through Pareto quantile plots of the variable value. Unlike in the previous example, here the explanatory variable is continuous implying that for each observation on the variable size only one or at
Fig. 3. Diamond data: scatterplot of value versus log(size).

Fig. 4. Diamond data: Pareto quantile plots of the variable value for a possible partition of the variable size.

most a small number of observations are available for the dependent variable. Therefore, a certain grouping along the variable size is necessary, hereby ignoring possible differences in the conditional distributions of the variable value. The Pareto quantile plots for one possible partition of the design space are given in Fig. 4. Also here we see that the quantile plots become linear in the largest observations. Moreover the tail heaviness of the value distribution, conditional on size, increases with the size of the stone.
In Section 2 it is shown how the case specific part of each observation can be removed by a power transformation of the dependent variable. Then an exponential regression model for log-spacings of these residuals is proposed. The parameters of this survival regression model can be estimated with the maximum likelihood method. This is described in Section 3 where we will also give some results from simulation experiments. Section 4 describes the estimation of extreme quantiles of the conditional distribution of the dependent variable. In a final section the method proposed will be illustrated using the two practical examples.

2. Derivation of the exponential regression model

Consider independent positive random variables $Y_1, \ldots, Y_n$ and let $x_i$ denote a $(p+1)$-dimensional vector of explanatory variables associated with $Y_i$. We assume that the conditional distribution of $Y$ given $x$ is of Pareto-type, i.e. there exists a $\gamma(x) > 0$ such that

$$1 - F_{Y|x}(y) = y^{-\frac{1}{\gamma(x)}} \tilde{l}(y; x).$$

Clearly, the extreme-value index $\gamma$ and the slowly varying function $\tilde{l}$ may depend on $x$. Dependence of $\gamma$ on $x$ can be modelled by any positive function. In this article we will focus on an exponential link function

$$\gamma(x) = \exp(\beta' x)$$

with $\beta$ denoting a $(p+1)$-dimensional vector of regression coefficients. Of course, the method carries over to other link functions.

We denote with $\beta(0), x(0)$ the parameter vector, respectively, the vector of independent variables, without the intercept so that $\beta' = (\beta_0, \beta'(0))$ and $x' = (1, x'(0))$. The transformation

$$R = Y^{\exp(-\beta'(0)x(0))},$$

removes the case specific part from the extreme-value index:

$$1 - F_{R|x}(r) = r^{-1/\exp(\beta_0)} \tilde{l}(r^{\exp(\beta'(0)x(0))}; x).$$

In this paper, the class of distribution functions satisfying (7) will be restricted to the distributions $F$ for which

$$\tilde{l}(r^{\exp(\beta'(0)x(0))}; x) = \tilde{l}(r).$$

Assumption (11) is equivalent to considering these distributions for which transformation (9) removes the conditioning on $x$ completely. For the Hall (1982) class of Pareto-type distributions

$$1 - F_{Y|x}(y) = ay^{-\gamma} [1 + dy^{-\theta} + o(y^{-\theta})], \quad \gamma, a, \theta > 0, \quad d \in \mathbb{R},$$

(12)
or

\[ U_{\gamma|x}(y) = (ay)^{\gamma} \left\{ 1 + \frac{d\gamma}{d\theta} y^{-\theta} [1 + o(1)] \right\} \]

with \( \gamma = \exp(\beta' x) \) and \( a, d \) and \( \theta \) possibly depending on \( x \), the tail quantile function of the generalized residuals, \( U_{\gamma|x} \), can be approximated by

\[ U_{\gamma|x}(r) \sim (ra)^{\gamma_0} \left\{ 1 + \frac{d\gamma_0}{d\theta} r^{-\theta \gamma} \right\} \]

with \( \gamma_0 = \exp(\beta_0) \), so that condition (11) is essentially equivalent to requiring \( a, d \) and \( \theta \gamma \) constant. As an illustration of condition (11), consider the Burr(\( \eta, \tau, \lambda \)) distribution (Burr, 1942) with distribution function given by

\[ F(y) = 1 - \left( \frac{\eta}{\eta + y^\tau} \right)^{\lambda}, \quad y > 0, \quad \lambda, \eta, \tau > 0. \]

Clearly, the Burr(\( \eta, \tau, \lambda \)) distribution is of Pareto-type with \( \gamma = 1/(\lambda \tau) \) and \( \tilde{l}(y) = \eta^\tau (1 + \eta y^{-\tau})^{-\lambda}. \) For the parametrization \( \tau = \exp(-\beta' x) \) transformation (9) removes the conditioning on \( x \) completely and hence \( \tilde{l} \) satisfies (11):

\[ \tilde{l}(r, \exp(\beta_0' x_0); x) = \left( \frac{\eta}{1 + \eta/r^{\exp(-\beta_0)}} \right)^{\lambda}. \]

On the other hand, for the parametrization \( \lambda = \exp(-\beta' x) \) condition (11) is not satisfied:

\[ \tilde{l}(r, \exp(\beta_0' x_0); x) = \left( \frac{\eta}{1 + \eta/r^{\exp(\exp(-\beta_0))}} \right)^{\exp(-\beta' x)}. \]

The random variables \( R_1, \ldots, R_n \), obtained by applying (9) to \( Y_1, \ldots, Y_n \), are clearly independent (since the \( Y_i \) are independent) and identically distributed (\( \gamma \) and \( \tilde{l} \) no longer depend on \( x \)). Therefore, the \( R_i \) will be called generalized residuals. The order statistics associated with \( R_1, \ldots, R_n \) will be denoted by

\[ R_{1,n} \leq \cdots \leq R_{n,n}. \]

In the extreme-value literature one imposes the so-called slow variation with remainder condition (see Section 3.12.1 of Bingham et al., 1987) on the slowly varying function \( l \). This condition specifies the rate of convergence to the limit in (2).

**Assumption (R).** There exists a real constant \( \rho < 0 \) and a rate function \( b \) satisfying \( b(x) \to 0 \) as \( x \to \infty \), such that for all \( \lambda \geq 1 \), as \( x \to \infty \),

\[ \log \frac{l(\lambda x)}{l(x)} \sim b(x) k_\rho(\lambda) \]

with \( k_\rho(\lambda) = \int_1^\lambda e^{b(t)} \, dt \).
Note that this assumption is clearly satisfied for the Hall (1982) class of Pareto-type distributions with $\rho = -\gamma$.

Under ($R_l$) and using derivations similar to the ones in Beirlant et al. (1999), the following regression model can be proposed:

$$Z_j = \left( \gamma_0 + b_{n,k} \left( \frac{j}{k + 1} \right)^{-\rho} \right) F_j, \quad j = 1, \ldots, k$$

with $Z_j = j(\log R_{n-j-1,n} - \log R_{n-j,n})$, $b_{n,k} = b((n + 1)/(k + 1))$ and $F_1, \ldots, F_k$ denoting independent standard exponential random variables. Considering the reduced model

$$Z_j = \gamma_0 F_j, \quad j = 1, \ldots, k,$$

the Hill estimator $H_{k,n}$ follows as the maximum likelihood estimator of $\gamma_0$ under (14); in fact $H_{k,n} = (1/k) \sum_{j=1}^{k} Z_j$.

Model (13) is illustrated in Fig. 5 using a simulated dataset of size $n=1500$ from the Burr(1, 1, 1) distribution for which $\gamma_0 = 1$. In Fig. 5, we plot $Z_j$ versus $j/(k + 1)$, $j = 1, \ldots, 500$, for $k = 500$. The horizontal reference line represents the Hill estimator which is obtained as the sample mean of the $Z_j$, $j = 1, \ldots, 500$. Model (13) states that conditional on $j$, the distribution of $Z_j$ is approximately exponential. This is illustrated in Fig. 6 where we show the exponential quantile plots of the $Z_j$ values for which respectively $0.20 \leq j/(k + 1) \leq 0.25$ (left) and $0.70 \leq j/(k + 1) \leq 0.75$ (right). As is also clear from these figures, the exponential scale parameter $\gamma_0 + b_{n,k}(j/(k + 1))^{-\rho}$
increases with \( j \). This also explains the positive bias of the Hill estimator as \( b_{n,k} > 0 \) in this case.

Remark that in practice, a regression analysis of type (13) has to be applied for each \( k = 3, \ldots, n-1 \). The resulting different estimators of \( \gamma_0 \) are typically stable over \( k \), this in contrast to for instance the Hill estimator. Model (13) will be used in the remainder of this article. The parameters of regression model (13) can be estimated with the maximum likelihood method. This will be discussed in the next section.

3. Maximum likelihood estimation of \( \beta \)

In this section we discuss how the parameters of the exponential regression model (13) can be estimated with the maximum likelihood method. For independent realizations \( z_1, \ldots, z_k \) the log-likelihood function is given by

\[
\log L(\gamma_0, \beta_{(0)}, b_{n,k}, \rho | z_k) = -\sum_{j=1}^{k} \log \left( \gamma_0 + b_{n,k} \left( \frac{j}{k+1} \right)^{-\rho} \right) - \sum_{j=1}^{k} \frac{z_j}{\gamma_0 + b_{n,k}(j/(k+1))^{-\rho}}.
\]  

Note that the likelihood function depends on \( \beta_{(0)} \) through the ordered residuals. Maximum likelihood estimates are obtained by solving the following constrained nonlinear optimization problem:

\[
\max \log L(\gamma_0, \beta_{(0)}, b_{n,k}, \rho | z_k)
\]  

Fig. 6. Burr(1,1,1) data (\( n = 1500 \)): exponential quantile plots of \( Z_j \) values with \( 0.20 \leq j/(k+1) \leq 0.25 \) and \( Z_j \) values with \( 0.70 \leq j/(k+1) \leq 0.75 \).
subject to
\[ \gamma_0 > 0, \]
\[ \beta_{(0)} \in \mathbb{R}^p, \]
\[ b_{n,k} \in \mathbb{R}, \]
\[ \rho < 0, \]
\[ \gamma_0 + b_{n,k} \left( \frac{k}{k+1} \right)^{-\rho} > 0. \]  
(17)

The last constraint of (17) ensures that all exponential parameters are positive. Remark that if \( \rho = 0 \), model (13) is not identifiable. This implies that numerical instabilities can be expected in cases where \( \rho \) is too close to zero. The Fortran NAG library subroutine E04UCF (Numerical Algorithms Group, 1993) was used to solve the constrained nonlinear optimization problem defined by (16) and (17) for \( k = 3, \ldots, n - 1 \).

As an alternative procedure, the maximum likelihood estimates can also be obtained by maximizing the profile log-likelihood function (see Barndorff-Nielsen and Cox, 1994) of \( \beta_{(0)} \) defined as
\[ \log L_p(\beta_{(0)}) = \max_{\gamma_0, b_{n,k}, \rho} \log L(\gamma_0, \beta_{(0)}, b_{n,k}, \rho | z_k). \]  
(18)

A Fortran program is available that performs the required optimizations over a grid of \( \beta_{(0)} \) values and this for \( k = 3, \ldots, n - 1 \). The optimization involved in computing \( \log L_p(\beta_{(0)}) \), as defined by (18), is also performed by the nonlinear optimization subroutine E04UCF of the NAG library. Both procedures, the direct optimization based on (16) and (17) or the profile likelihood method based on (18), yield almost identical results on several simulated datasets. The profile log-likelihood approach will be followed in this paper. Both Fortran programs are available upon request from the authors.

The optimizations necessary to solve (16) and (17), or (18), can still be implemented in other optimization packages such as e.g. the GAMS nonlinear optimization subroutine of which a demo version can be downloaded from www.gams.com.

Inference about the regression coefficients can be drawn by using the profile log-likelihood ratio test statistic given by \( 2(\log L_p(\beta_{(0)}) - \log L_p(\beta_{(0)}^*)) \). This statistic equals the log-likelihood ratio statistic for testing the hypothesis \( \beta_{(0)} = \beta_{(0)}^* \). As discussed in the appendix, here the classical \( \chi^2 \) approximation to the null distribution of the test statistic is inappropriate. We therefore propose to simulate the reference distribution by using a parametric bootstrap procedure (Efron and Tibshirani, 1993) and to compute a \( p \)-value based on this simulated reference set. Bootstrap samples are generated from a strict Pareto distribution with parameters \( \beta_{(0)}^* \) and the maximum likelihood estimate of \( \gamma_0 \) that is obtained under \( \beta_{(0)} = \beta_{(0)}^* \). A further discussion about the scale of the log-likelihood function is given in the appendix where an approximate expression for the expected value of the log-likelihood is derived.

An Example: To illustrate the procedure, a dataset of size \( n = 1500 \) was generated from a Burr distribution with \( \eta = 1, \lambda = 1 \) and \( \tau = \exp(0.5 + x) \), so \( \gamma(x) = \exp(-0.5 - x) \).
and $\rho = -1$. The values for the explanatory variable $x$ are generated from the uniform $(-2, 2)$ distribution. In Fig. 7(a) a graphical presentation of the conditional distributions $F_{Y|x}$ is given. Remark the variety of distributions from which the data are generated. For the larger $x$-values the conditional density is clearly unimodal, while for the smaller $x$-values the conditional density is L-shaped. The Pareto quantile plot of the generalized residuals $r_i = y_i^{\exp(x_i)}$, $i = 1, \ldots, n$, is given in Fig. 7(b). The quantile plot becomes linear for the largest residuals indicating that the tail of the distribution of the residuals can be accurately described by a Pareto distribution. The maximum likelihood estimates of $\beta_0$ and $\beta_1$ for $k = 3, \ldots, 1000$ are shown in Fig. 7(c). The profile log-likelihood function of $\beta_1$, $\log L_\rho(\beta_1)$, is given in Fig. 7(d) for $k = 50, 100, 200, 300, 400, 500$. For each $k$ the maximum is attained around the true value of $\beta_1$. Note also that the profile log-likelihood functions are more peaked as $k$ increases.
Simulation results: The statistical properties of the proposed estimator will be investigated by a simulation study. Datasets of size $n=1500$ were generated from Burr$(1, \tau, \lambda)$ distributions with

- $\tau(x) = \exp(0.5 + x)$, $\lambda = 1$ giving $\gamma(x) = \exp(-0.5 - x)$ and $\rho = -1$;
- $\tau(x) = \exp(0.5 + x)$, $\lambda = 2$ giving $\gamma(x) = \exp(-1.2 - x)$ and $\rho = -0.5$;
- $\tau = 1$, $\lambda(x) = \exp(0.5 + x)$ giving $\gamma(x) = \exp(-0.5 - x)$ and $\rho(x) = -\exp(-0.5 - x)$.

Remark that in this final case, condition (11) is not satisfied. The values for the explanatory variables were drawn from the uniform $(-2, 2)$ distribution.

Here, the parameter $\rho$ is the most important design factor since it determines the speed at which the slowly varying nuisance part in (13) disappears (cf. assumption $(R_i)$). In general, tail index estimators are more biased as $|\rho|$ is smaller. Remark that for instance in the Hall class $b(x) \sim cx^\rho$ as $x \to \infty$.

Figs. 8–10 show the quartiles (based on 100 realizations) of $\hat{\beta}_0$ and $\hat{\beta}_1$ as a function of $k$. Next to the estimates obtained with the above described procedure (—), we
Fig. 9. Burr(1, exp(0.5 + x), 0.2)-data: quartiles of (a) $\hat{\beta}_0$ and (b) $\hat{\beta}_1$ as a function of $k$; solid line: maximum likelihood estimator based on exponential regression model, broken line: maximum likelihood estimator based on GPD regression model.

also show the maximum likelihood estimates obtained by fitting a GPD regression model with $\gamma(x) = \exp(\beta' x)$ in (6) to exceedances over a specified threshold (- - -). In this simulation study, we used a constant threshold (the $(k + 1)$-largest order statistic of the dependent variable). However, in a regression analysis a covariate dependent threshold is more appropriate since it takes the relative extremity of the observations into account. This issue was also noted by Coles and Tawn (1998) and by Davison and Smith (1990). Up to now, solutions seem to be more ad hoc and depending on the data set at hand. Separate research could be performed in order to develop a method to incorporate a covariate dependent threshold (constructed in such a way that exactly $k$ observations fall above it) in a fully automatic procedure.

From Figs. 8 and 9 it is clear that our estimator has a better overall performance than the GPD-based maximum likelihood estimator, both in terms of bias and interquartile range. Remark however that the GPD approach can be used in the more general setting with $\gamma \in \mathbb{R}$. Fig. 10 shows the results of the procedure when assumption (11) is not satisfied. In this case, the estimator performs only well at the smallest $k$-values since
the influence of the slowly varying function is minimal for the largest observations. In Fig. 10(c) and (d) we give the Pareto quantile plots of the generalized residuals computed by using the estimates of $\hat{\beta}_1$ at, respectively, $k = 50$ and 400. The Pareto quantile plot of the generalized residuals at $k=400$ does not become linear in the largest observations indicating that the tail of the residual distribution cannot be adequately described by the Pareto distribution.

The use of the parametric bootstrap as a device to test a linear hypothesis concerning the regression coefficients is illustrated in Fig. 11. Data were generated from
Burr(1, exp(0.5 + β1x), 1) distributions. The hypothesis of interest is H0: β1 = 0. Two cases are considered: (i) simulation under H0 (τ(x) = exp(0.5)) and (ii) simulation under H1 (τ(x) = exp(0.5 + x)). The bootstrap samples were generated from a strict Pareto distribution with β1 = 0 and with the maximum likelihood estimate of β0 given β1 = 0. The hypothesis was tested at k = 200. In Figs. 11(a) and (b) we show the histograms of 100 bootstrap realizations of the log-likelihood ratio test statistic, respectively, for case (i) (estimated p-value = 0.74) and case (ii) (estimated p-value = 0). The histogram of 100 realizations of the log-likelihood ratio statistic obtained from a Monte Carlo simulation under H0 is given in Fig. 11(c). On all histograms, we superimposed the χ²(1) density. The results obtained from the parametric bootstrap (case (i)) and from
the simulation study are compared by using the two sample quantile plot shown in Figure 11(d).

4. Estimation of large quantiles

In this section we discuss the estimation problem of extreme quantiles of the conditional distribution of the dependent variable: given a (small) exceedance probability \( p \) and a vector of explanatory variables \( x \) we consider the problem of estimating \( y_p \) such that \( 1 - F_{Y|x}(y_p) = p \). Situations where estimation of extreme quantiles is important are numerous. For instance an insurance company facing a fire insurance portfolio may be interested in the claim size that will be exceeded only once in 10000 claims given a sum insured and a certain type of building. In case of an i.i.d. sample, \( Y_1, \ldots, Y_n \), extreme quantiles can be estimated by extrapolation along a line through the anchorpoint \((\log((n+1)/(k+1)), \log Y_{n-k,n})\) with slope \( \hat{\gamma} \) on the Pareto quantile plot, resulting in the estimator (see e.g. Weissman, 1978)

\[
\log Q_{Y,k}(p) = \log Y_{n-k,n} + \hat{\gamma} \left[ \log \left( \frac{1}{p} \right) - \log \left( \frac{n+1}{k+1} \right) \right]
\]

or equivalently

\[
Q_{Y,k}(p) = Y_{n-k,n} \left[ \frac{k+1}{(n+1)p} \right]^{\hat{\gamma}}, \quad k = 1, \ldots, n-1.
\]

Here \( \hat{\gamma} \) is often taken to be equal to \( H_{k,n} \).

When covariate information is available, estimator (20) cannot be applied directly to the raw data since the observations are not longer identically distributed. In this situation, the observations will be first transformed to i.i.d. data by using (9). Next, (20) will be used in the extrapolation step, yielding an estimator of an extreme quantile of \( F_R \). Finally, the quantile estimator of the generalized residuals will be transformed back to the original observations by inverting (9). This results in the following estimator for the \((1-p)\)th quantile of \( F_{Y|x} \):

\[
Q_{Y,k}(p; x) = \left\{ \hat{\gamma}_{0,k} \exp(\hat{\beta}'_{0,1} x_{0,1}) \right\} \hat{R}_{n-k,n} \left[ \frac{k+1}{(n+1)p} \right]^{\hat{\gamma}_{0,k}}, \quad k = 3, \ldots, n-1,
\]

with \( \hat{\beta}'_{0,1} \) denoting the estimator of \( \beta_{(0)} \) obtained by using the \( k \) largest order statistics, \( \hat{\gamma}_{0,k} \) representing the \((k+1)\)th largest order statistic of the generalized residuals obtained by using \( \hat{\beta}'_{0,1} \) in (9), and with \( \hat{\gamma}_{0,k} = \exp(\hat{\beta}_{0,k}) \).

Simulation results: The estimates of \( \beta \) obtained in the simulation study described under Section 3 will now be used for the estimation of large quantiles. Here we will concentrate on the Burr(1, exp(0.5 + x), 1) data. In Fig. 12 the quartiles of \( Q_{Y,k}(0.001; x) \) and \( Q_{Y,k}(0.00001; x) \) are given as a function of \( k \) at \( x = -1, 0, 1 \). Note that for small
The accurate estimation of extreme quantiles is much more difficult than at larger x-values. This could be expected since the tail heaviness decreases with increasing x.

### 5. Practical examples

In this section we will illustrate the proposed method with the two practical examples introduced in Section 1.

As a first example, we fit regression model (13) to the Norwegian fire claim data. To further examine the possible influence of time on the claim size distribution, a regression model with time as explanatory variable and claim size as dependent variable was fitted to the data, so \( \gamma(\text{time}) = \exp(\beta_0 + \beta_1 \text{time}) \). Maximum likelihood estimates of \( \beta_0 \) and \( \beta_1 \) are shown in, respectively, Fig. 13(c) and (d) for \( k = 3, \ldots, 500 \) (solid line: maximum likelihood estimates for the exponential regression model presented in this paper, broken line: maximum likelihood estimates for GPD regression model).
The profile log-likelihood functions of $\beta_1$ for $k = 100, 200, 300, 400, 500$ are shown in Fig. 13(b). In Fig. 14(a) the histogram of 100 bootstrap realizations of the log-likelihood ratio statistic for the hypothesis test $H_0: \beta_1 = 0$ is given. The hypothesis is tested at $k = 250$. With an observed value of the test statistic of 72.79 (estimated $p$-value = 0), $H_0$ can be safely rejected. For insurance companies, estimates of claim sizes that will occur once in say 10,000 claims are of crucial importance. In Fig. 14(b) we show the estimate of the 0.9999 quantile for year 1990 obtained along formula (20). As can be expected with an extrapolation procedure as the one proposed here, the extreme quantile estimate is quite sensitive with respect to the estimation of $\beta$. The $\beta$-estimate is most stable for $k$ between 350 and 500 so preference could be given to a quantile estimate between 300,000 and 400,000.
Our second example concerns the valuation of precious stones. In a first attempt, trying to fit regression models over the whole range of the variable size, the application of model (7)–(8) with $Y = \text{value}$ and $x' = (1, \text{size})$ does not provide an appropriate fit: the Pareto quantile plot of the generalized residuals becomes horizontal for the largest observations i.e. $\hat{\gamma}_0 = 0$, see Fig. 15(f) where we show the Pareto quantile plot of the generalized residuals obtained by using the $\beta_1$-estimate at $k = 200$. Rather the extreme-value index is found to vary polynomially with size. The scatterplot of value versus log(size) is given in Fig. 15(a). In Fig. 15(b) we show the profile log-likelihood function of $\beta_1$ for $k = 200$, 250, 300, 350, 400. The maximum likelihood estimates of $\beta_0$ and $\beta_1$ are displayed in, respectively, Fig. 15(c) and (d) for $k = 3, \ldots, 500$. Again the estimator proposed in this paper—represented by the solid line—is compared with the estimator based on the GPD regression model—represented by the broken line. The large difference in $\beta_1$-estimates between the GPD regression model and our approach can be explained by the fact that the GPD threshold from $k = 200$ to 500 neglects data with log(size) $< -2$. Hence the GPD estimation is seriously influenced by the set of stones with size > 0.14. In Fig. 15(e) we show the Pareto quantile plot of the generalized residuals computed using the $\beta_1$-estimate obtained at $k = 200$ ($\hat{\gamma}_{0,200} = 0.12$). Finally, the 0.99 quantile conditional on log(size), obtained using $k = 102$, is superimposed on the value versus log(size) scatterplot shown in Fig. 15(a). The $k$-value used to compute the 0.99 quantiles is selected in an optimal way in the sense that this $k$ minimizes

$$
\min_{k} \left| \frac{1}{1914} \sum_{i=1}^{1914} I[\text{value}_i > Q_{Y,k}(0.01; \text{log(size)}_i)] - 0.01 \right|
$$
Fig. 15. Diamond data: (a) scatterplot of value versus log(size) with \( Q_{y,0.1}(0.01; \log(size)) \) superimposed, (b) profile log-likelihood function of \( \beta_1 \) for \( k = 200, 250, 300, 350, 400 \), (c) estimate of \( \beta_0 \) for \( k = 3, \ldots, 500 \), (d) estimate of \( \beta_1 \) for \( k = 3, \ldots, 500 \) (in (c) and (d) the maximum likelihood estimates for regression model (13) are given by the solid line, the GPD regression model based maximum likelihood estimates by the broken line), (e) Pareto quantile plot of the generalized residuals for the regression model with log(size) as explanatory variable and (f) Pareto quantile plot of the generalized residuals for the regression model with size as explanatory variable (in (e) and (f) the generalized residuals are computed using the \( \beta_1 \)-estimates at \( k = 200 \)).

6. Conclusion

In this paper a method is presented which allows the estimation of the positive extreme-value index when covariate information is available. The method is based on
the transformation of the dependent variables into generalized residuals and on an exponential regression model for these residuals. Parameters are estimated with the maximum likelihood method. We also proposed an estimator for extreme quantiles of the conditional distribution of the dependent variable.

In practice it sometimes appears to be difficult to find out an appropriate relationship between the extreme-value index $\gamma$ and the covariate information (see for instance the diamond example) as we still lack flexible diagnostic tools in comparison with classical regression techniques. Here, nonparametric procedures plotting univariate extreme-value index estimators applied locally against the covariates, might prove helpful in an initial exploratory step.

An AMSE criterium could be constructed to guide us in the selection of an optimal number of extremes to be used in the estimation. Such a procedure is clearly desirable in cases where assumption (11) is not satisfied. Another extension is towards the $\gamma \in \mathbb{R}$ case. Similarly to the approach followed here, a regression model for log-spacings of $UH$-statistics (Beirlant et al., 1996b) can be constructed.

Acknowledgements

The authors thank G. Dierckx, W. Gochet, G. Matthys for stimulating suggestions and discussions and J. Caers for providing the diamond data. The comments of the referee and the associate editor were very helpful in improving the presentation of the paper.

Appendix

In this section we further discuss the log-likelihood function. To simplify the derivations, the slowly varying part of (7) is ignored. Denote by $R^{(i)}_j$, $j=1, \ldots, n$, the residuals computed using $\beta^{(i)}_0$, $i=1, 2$, two different values of $\beta_0$ with $\beta^{(1)}_0$ being the true parameter vector. Moreover, let $Y^{(2)}_{n,j}, x^{(2)}_{n,j}$ denote the dependent variable respectively the vector of explanatory variables associated with $R^{(2)}_{j,n}$. The log-likelihood function evaluated at $\gamma_0$ and $\beta^{(2)}_0$ can be rewritten as

$$\log L(\gamma_0, \beta^{(2)}_0) = -k \log \gamma_0 - \sum_{j=1}^{k} \frac{j[\log R^{(2)}_{n-j+1,n} - \log R^{(2)}_{n-j,n}]}{\gamma_0}$$

$$= -k \log \gamma_0 - \sum_{j=1}^{k} \frac{j}{\gamma_0} \left\{ \exp \left[ -\beta^{(2)}_0' x^{(2)}_{n,j} \right] \log Y^{(2)}_{n,j+1,n} \right\} \log Y^{(2)}_{n,j,n}$$

$$- \exp \left[ -\beta^{(2)}_0' x^{(2)}_{n,j} \right] \log Y^{(2)}_{n,j,n} \right\}.$$
In (22), \( \log \left\{ Y_{R(2)}^{(1)} \right\} /\gamma_0, j = 1, \ldots, k + 1, \) are distributed as standard exponential order statistics. Using the Rényi representation of standard exponential order statistics (22) can be stated as

\[
\log L(\gamma_0, \beta^{(2)}(0)) = -k \log \gamma_0 - \sum_{j=1}^{k} \left\{ \frac{\exp \left[ -\beta^{(2)}(0) x_{(0), R^{(2)}_{n-j+1,n}} \right]}{\exp \left[ -\beta^{(1)}(0) x_{(0), R^{(2)}_{n-j+1,n}} \right]} \sum_{l=1}^{i(j)} \frac{\omega_l}{n - l + 1} \right\} - \frac{\exp \left[ -\beta^{(2)}(0) x_{(0), R^{(2)}_{n-j+1,n}} \right]}{\exp \left[ -\beta^{(1)}(0) x_{(0), R^{(2)}_{n-j+1,n}} \right]} \sum_{l=1}^{i(j)+1} \frac{\omega_l}{n - l + 1}
\]

with \( \omega_l, \ l = 1, \ldots, n, \) being independent standard exponential random variables and \( i(j) \) denoting the rank of the \( R^{(2)} \) residual that corresponds with \( R^{(2)}_{n-j+1,n} \). Note that in the above expression the number of terms in \( \sum_{l=1}^{i(j)} \omega_l/(n - l + 1) \) is random since this number depends on the relation between the ordered residuals \( R^{(1)} \) and \( R^{(2)} \). Conditioning on the relation between the two sets of residuals gives the following expression for the expected value of the log-likelihood:

\[
E[\log L(\gamma_0, \beta^{(2)}(0)) | I] = -k \log \gamma_0 - \sum_{j=1}^{k} \left\{ \frac{\exp \left[ -\beta^{(2)}(0) x_{(0), R^{(2)}_{n-j+1,n}} \right]}{\exp \left[ -\beta^{(1)}(0) x_{(0), R^{(2)}_{n-j+1,n}} \right]} \sum_{l=1}^{i(j)} \frac{1}{n - l + 1} \right\} - \frac{\exp \left[ -\beta^{(2)}(0) x_{(0), R^{(2)}_{n-j+1,n}} \right]}{\exp \left[ -\beta^{(1)}(0) x_{(0), R^{(2)}_{n-j+1,n}} \right]} \sum_{l=1}^{i(j)+1} \frac{1}{n - l + 1}
\]

In Fig. 16 we compare the sample means of (23), given by the full line, and (15), given by the broken line, in \( \beta^{(2)}_1 = -1.5 \) and computed over 100 simulated datasets.
of size $n = 1500$ from the Burr($1, \exp(0.5 + x), 1$) distribution for $k = 3, \ldots, 500$. Note that although (23) was obtained by ignoring the slowly varying nuisance part, both averages agree quite well. Further, Fig. 16 explains the scale of the log-likelihood function observed for instance in Fig. 7(d).

References


