18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Tests for Gumbel domain of attraction based on regression quantiles

J. Picek and J. Dienstbier

Department of Applied Mathematics, Technical University of Liberec, Czech Republic
Email: jan.picek@tul.cz

Abstract: If we are interested in such events as the extreme intensity of the wind, high flood levels of the rivers or extreme values of environmental indicators, or maximal or minimal performance of foreign exchange rates or share prices, we should take an interest in the tails of the underlying probability distribution rather than in its central part. Many authors were have dealt with an estimation of the tails of the distribution. However, besides the point and interval estimation, a typical and important part of statistical inference and modelling is the testing of hypotheses.

Many authors have developed methods for location model, i.e. they consider an i.i.d. sample, from an underlying distribution function with unknown shape, location and scale parameters, belonging to some max-domain of attraction. They tested the problem of Gumbel domain against Fréchet or Weibull domains. Neves, Picek and Alves (2006) based the testing decision on the ratio between the maximum and the mean of the top sample excesses above some random threshold.

The present paper deals with a linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$, where the errors are again from an underlying distribution function with unknown shape, location and scale parameters, belonging to some max-domain of attraction. We study a generalization of test as above based on the regression quantilesfor the null hypothesis that the distribution comes from the Gumbel domain of attraction.

The regression quantiles were introduced as a generalization of usual quantiles to linear regression model. The key idea in generalizing the quantiles is the fact that we can expressed the problem of finding the sample quantile as the solution to a simple optimization problem. This leads, naturally, to more general method of estimating of conditional quantiles fuctions. The optimization problem may be reformulated as a linear program and the simplex approach may be used to computing regression quantiles.

Dienstbier (2009) showed that location and scale invariant smooth functionals of the standardized intercepts of the highest order regression quantiles have the same asymptotic distribution as the same functionals based on the empirical tail quantile function of the underlying sample of errors. We generalize the tests on the basis of the exceedances over high quantile regression threshold. The type I error and power of the test are studied for finite sample sizes by simulation.

Keywords: *extreme value index, max-domain of attraction, quantile regression, statistical tail functional*

1 Introduction

Let $V_1, V_2, ..., V_n$ be independent and identically distributed random variables with common distribution function F with unknown shape, location and scale parameters, belonging to some max-domain of attraction. F is in the domain of attraction of an extreme-value distribution G_γ for some index $\gamma \in \mathbb{R}$ ($F \in \mathcal{D}(G_\gamma)$):

$$\exists_{b_n \in \mathbb{R}}^{a_n > 0} : F^n(a_n x + b_n) \underset{n \to \infty}{\longrightarrow} G_\gamma(x)$$

for all x, with

$$G_\gamma(x) := \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), & 1+\gamma x > 0 \quad \text{if} \quad \gamma \neq 0 \\ \exp(-\exp(-x)), & x \in \mathbb{R} \qquad\quad \text{if} \quad \gamma = 0 \end{cases}$$

the Generalized Extreme Value (GEV(γ)) distribution in the von Mises parameterization. Gnedenko (1943) has established that the class $\{G_\gamma\}_{\gamma \in \mathbb{R}}$ represents in an unified version all possible non-degenerate weak limits of the maximum $V_{n:n}$, up to location/scale parameters. GEV(γ) d.f. reduces to Weibull, Gumbel and Fréchet distributions, respectively, for $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$.

For positive γ, the behavior in the tail of the underlying distribution function F has important implications since it may suggest, for instance, the presence of infinite moments. All distribution functions belonging to $\mathcal{D}(G_\gamma)$ with $\gamma < 0$ are light tailed with finite right endpoint. The intermediate case $\gamma = 0$ is of particular interest in many applied fields where extremes are important, because an inference within the Gumbel domain G_0 is simple and also the great variety of distributions has an exponential tail.

Taking all into consideration, it has become clear the advantage of looking for the most appropriate type of tail when fitting empirical distributions at high quantiles. Effectively, separating statistical inference procedures according the most suitable domain of attraction for the underlying d.f. F has become an usual practice.

A test for Gumbel domain against Fréchet or Weibull max-domain has received the general designation of statistical choice of extreme domains of attraction (see e.g. Castillo et al. (1989), Hasofer and Wang (1992), Fraga Alves and Gomes (1996), Marohn (1998), Segers and Teugels (2001) and Neves, Picek and Alves (2006).

One of the challenging ideas of the recent advances in the field of statistical modeling of extreme events has been the development of models with time-dependent parameters or more generally models incorporating covariates. Consider the linear regression model

$$\mathbf{Y}_n = \mathbf{X}_n \boldsymbol{\beta} + \mathbf{E}_n, \tag{1}$$

where $\mathbf{Y}_n = \mathbf{Y} = (Y_1, \ldots, Y_n)'$ is a vector of observations, $\mathbf{X}_n = \mathbf{X}$ is an $(n \times p)$ known design matrix with the rows $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the $(p \times 1)$ unknown parameter ($p > 1$) and $\mathbf{E}_n = \mathbf{E} = (E_1, \ldots, E_n)'$ is an $(n \times 1)$ vector of i. i. d. errors with a distribution function $F \in \mathcal{D}(G_\gamma)$. We assume that the first column of $\mathbf{X}_n$ is $\mathbf{1}_n = (1, \ldots, 1)'$, i.e. the first component of $\boldsymbol{\beta}$ is an intercept.

The present paper deals with the two-sided problem of testing Gumbel domain against Fréchet or Weibull domains in the model (1), i.e.,

$$F \in \mathcal{D}(G_0) \quad \text{versus} \quad F \in \mathcal{D}(G_\gamma)_{\gamma \neq 0}. \tag{2}$$

2 Regression quantiles

Koenker and Basset (1978) introduced the regression quantile as a generalization of usual quantiles to linear regression model. The key idea in generalizing the quantiles is the fact that we can expressed the problem of finding the sample quantile as the solution to a simple optimization

problem. This leads, naturally, to more general method of estimating of conditional quantiles fuctions.

They defined the α-regression quantile $\hat{\boldsymbol{\beta}}(\alpha) = (\hat{\boldsymbol{\beta}}_1(\alpha), \ldots, \hat{\boldsymbol{\beta}}_p(\alpha))'$ $(0 < \alpha < 1)$ for the model (1) as any solution of the minimization

$$\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i'\mathbf{t}) := \min, \quad \mathbf{t} \in \mathbb{R}^p, \tag{3}$$

where

$$\rho_\alpha(x) = x\psi_\alpha(x),\ x \in \mathbb{R}^1 \text{ and } \psi_\alpha(x) = \alpha - I_{[x<0]},\ x \in \mathbb{R}^1. \tag{4}$$

The same authors characterized the α-regression quantile $\hat{\boldsymbol{\beta}}(\alpha)$ as the component $\boldsymbol{\beta}$ of the optimal solution $(\boldsymbol{\beta}, \mathbf{r}^+, \mathbf{r}^-)$ of the linear program

$$\alpha \mathbf{1}_n'\mathbf{r}^+ + (1-\alpha)\mathbf{1}_n'\mathbf{r}^- := \min$$

$$\mathbf{X}\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{Y} \tag{5}$$

$$\boldsymbol{\beta} \in \mathbb{R}^{p+1}, \mathbf{r}^+, \mathbf{r}^- \in \mathbb{R}^n_+ \quad 0 < \alpha < 1,$$

where $\mathbf{1}_n = (1, \ldots, 1)' \in \mathbb{R}^n$. This simplex approach may be used to computing regression quantiles. Implementation is contained for example in the software R.

One of the important properties of regression quantiles is their consistency, that is $\|\hat{\boldsymbol{\beta}}(\alpha) - \boldsymbol{\beta}(\alpha)\| = o_p(1)$, for each $\alpha \in (0,1)$ under some conditions of design matrix $\boldsymbol{X}$ and distribution function F, $\boldsymbol{\beta}(\alpha) = (\beta_1 + F^{-1}(\alpha), \beta_2, \ldots, \beta_p)$. For details, see Jurečková and Sen (1996).

That result can be generalized by a strong approximation of regression quantiles over the whole region of $\alpha \in [\alpha_n^*, 1-\alpha_n^*]$, where $\alpha_n^* \to 0$ with a selected order. We consider the following regularity conditions on the distribution function F and the design matrix $\boldsymbol{X}$.

(F.1) F has a derivative f that is positive and bounded on some left neighbourhood of the right endpoint x^*; f' is bounded and f'' exists on some left neigbourhood of x^*.

(F.2) the von Mises condition holds, i.e.

$$\lim_{t \to x^*} \frac{(1-F(t))f'(t)}{f^2(t)} = -1 - \gamma.$$

Fix b such that $0 < \delta \le b - |\gamma| \le |\gamma| + \delta$, for some $\delta > 0$.

(X.1) $x_{i1} = 1, \quad \mathrm{i} = 1, \ldots, \mathrm{n}.$

(X.2) $\lim_{n\to\infty} \boldsymbol{D}_n = \boldsymbol{D}$, where $\boldsymbol{D}_n = n^{-1}\boldsymbol{X}_n'\boldsymbol{X}_n$ and $\boldsymbol{D}$ is a positive definite $(p \times p)$ matrix.

(X.3) $n^{-1}\sum_{i=1}^n \|\mathbf{x}_i\|^4 = O(1)$ as $n \to \infty$.

(X.4) $\max_{1\le i\le n} \|\mathbf{x}_i\| = O(n^{\Delta})$ as $n \to \infty$, where

$$\Delta = \frac{b - |\gamma| - \delta}{1 + 2b} < \frac{1}{4}$$

It has been shown in Dienstbier (2009) using similar results as in Jurečková (1999), that under condition (F.1)-(F.2) and (X.1)-(X.4)

$$\sup_{\alpha_n^* \le \alpha \le 1-\alpha_n^*} \left\| \sigma_\alpha^{-1}(\hat{\boldsymbol{\beta}}(\alpha) - \boldsymbol{\beta}(\alpha)) \right\| = O_P(n^{-1/2}C_n), \tag{6}$$

where $C_n = C(\log\log n)^{1/2},\ 0 < C < \infty$ and

$$\begin{aligned} \alpha_n^* &:= n^{-\frac{1}{1+2b}} \\ \sigma_\alpha &:= \frac{(\alpha(1-\alpha))^{1/2}}{f(F^{-1}(\alpha))}, \qquad 0 < \alpha < 1. \end{aligned}$$

3 Tests based on regression quantiles

Dienstbier (2009) used (6) and showed the similarily of the tail quantile process in i.i.d. case and the regression quantiles process.

Theorem 3.1. *Consider the linear model (1) and conditions (F.1) – (F.2), (X.1) – (X.4.). Suppose also that the F satisfies the second order extreme value condition*

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - \frac{x^\gamma - 1}{\gamma}}{A(t)} = \frac{1}{\rho}\left(\frac{x^{\gamma+\rho}-1}{\gamma+\rho} - \frac{x^\gamma - 1}{\gamma}\right) \tag{7}$$

with tail quantile function $U(t) := \inf\{x : (\frac{1}{1-F})(x) \geq t\}$ *for all* $x > 0$ *with* ρ *the non-positive second order parameter,* $a > 0$ *and* A *a suitable positive or negative function. Then there exists a sequence of Wiener process* $\{W_n(s)\}_{s>0}$ *such that for each* $\varepsilon > 0$

$$\sup_{1/k_n \leq s \leq 1} s^{\gamma+1/2+\varepsilon} \left| \left(\frac{\hat{\beta}_1(1-\frac{k_n}{n}) - \beta_1 - F^{-1}(1-k_n/n)}{a(n/k_n)} - \frac{s^{-\gamma}-1}{\gamma} \right) - s^{-(\gamma+1)} W_n(s) + \sqrt{k} A(n/k_n) \Psi_{\gamma,\rho}(s^{-1}) \right| \xrightarrow[n\to\infty]{\mathrm{P}} 0$$

where $(k_n)_{n\in\mathbb{N}}$ *is an intermediate sequence such that* $k_n > n^{\frac{2b}{2b+1}}$ *and* $k_n/n \to 0$ *as* $n \to \infty$ *and* $\Psi_{\gamma,\rho}$ *is defined as in de Haan and Ferreira (2006).*

See Dienstbier (2009), Theorem 2.1.

According to the previously theorem, the asymptotic distribution of the high regression quantile can be well approximated by the tail quantile function of the Generalized Pareto distribution. Similarly as in Drees (1998), we can derive the asymptotic properties of the whole class of smooth and location and scale invariant functionals of the tail quantile function. We introduce this idea for the tail quantile function calculated from the exceedances over given threshold. Suppose to have a sample of observations $Y_1, \ldots, Y_n$ obtained from the linear model (1). Define a subsample of exceedances over some "high" regression quantile threshold

$$Z_i := \left(Y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}(\tau_{k_n})\right)^+, \tag{8}$$

i.e. for some $\tau_{k_n} = (1 - k_n/n)$ and the intermediate order sequence of k_n such that $k_n/n \to 0$ as $n \to \infty$ and $k_n > n^{\frac{2b}{2b+1}}$. Define also the empirical tail quantile function of this subsample as

$$Q_n^Z(t) := Z_{n-[k_n t]:n}$$

for $t \in [0,1]$, $Z_{i:n}$ denotes the i-th order statistics, $i = 1, \ldots, n$. Note, that the number of positive exceedances l depends on the exact form of regression matrix $\mathbf{X}$ and the vector of observations $\mathbf{Y}$. Denote the empirical tail quantile function of the unobservable errors of the model (1) as

$$Q_n^E(t) := E_{n-[k_n t]:t}$$

for $t \in [0,1]$. Let T is a suitable functional, then it follows from Theorem 3.1 and Theorem 2.1 in Drees (1998) that the distributions of $T(Q_n^E)$ and $T(Q_n^Z)$ coincide. If we introduce the concept of Hadamard differentiability according to Drees (1998) then obtain the same solution as in the location model for the test statistics of the various tests for Gumbel domain. Hence we can generalize these tests in the situation of the regression model on the basis of the exceedances over high quantile regression threshold in the following way. First we create subsample $\boldsymbol{Z}$, see (8). Then we plug ordered positive exceedances into the usual test statistics.

For example, the test statistic $T_{k,n}$ suggested by Neves, Picek and Alves (2006) has the form

$$T_{k,n} := \frac{V_{n:n} - V_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k}(V_{n-i+1:n} - V_{n-k:n})}$$

where $V_1, V_2, ..., V_n$ are i.i.d. random variables and $V_{1:n} \leq V_{2:n} \leq ... \leq V_{n:n}$ the order statistics after arranging the random sample in nondecreasing order, $k = k_n$ is a sequence of positive integers, $k_n \to \infty$ as $k_n/n \to 0$, as the sample size n tends to infinity.

We come back to the linear regression model (1)

$$\mathbf{Y}_n = \mathbf{X}_n\boldsymbol{\beta} + \mathbf{E}_n,$$

where the errors are from an underlying distribution function F with unknown shape, location and scale parameters, belonging to some max-domain of attraction $F \in \mathcal{D}(G_\gamma)$. If we are interested in the two-sided problem of testing Gumbel domain against Fréchet or Weibull domains

$$F \in \mathcal{D}(G_0) \quad \text{versus} \quad F \in \mathcal{D}(G_\gamma)_{\gamma \neq 0}.$$

then we suggest the following test statistics based on the largest regression quantiles

$$T_\tau := \frac{Z_{n:n} - Z_{n-l:n}}{\frac{1}{l}\sum_{i=1}^{l}(Z_{n-i+1:n} - Z_{n-l:n})} \tag{9}$$

where $Z_i := \left(Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}(\tau_{k_n})\right)^+, i = 1, \ldots, n$, and the l observations exceed the high regression threshold $\hat{\boldsymbol{\beta}}(\tau_{k_n})$ for some $\tau = \tau_{k_n} = (1 - k_n/n)$, where k_n is the intermediate order sequence $k_n \to \infty$ as $k_n/n \to 0$, as the sample size n tends to infinity.

We can prove on the basis of a result in Neves, Picek and Alves (2006) that T_τ under the null hypothesis converges to a random variable with the Gumbel distribution and the test is consistent

Theorem 3.2. *Suppose $F \in \mathcal{D}(G_0)$ and that second order property (7) holds for $\gamma = 0$. Let A is a suitable positive or negative function in (7) and let k_n be an intermediate sequence of integers such that $A(\frac{n}{k_n})\log^2 k \to 0$, as $n \to \infty$ ($\rho = 0$) or $A(\frac{n}{k_n})\log k \to 0$, as $n \to \infty$ (a negative ρ) and let $\tau = \tau_{k_n} = (1 - k_n/n)$,*

$$T_\tau \xrightarrow{d} G, \quad G \sim Gumbel$$

Theorem 3.3. *Suppose $F \in \mathcal{D}(G_\gamma)$ and that following condition holds for some $\gamma \in \mathbb{R}$.*

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}$$

for every $x > 0$ and some positive mensurable function a, with $U(t) := \inf\{x : (\frac{1}{1-F})(x) \geq t\}$. Let k_n be an intermediate sequence of integers such that $k_n \to \infty$ and $\frac{k_n}{n} \to 0$ as $n \to \infty$. Then, as $n \to \infty$,

(i) if $\gamma < 0$, $T_\tau \xrightarrow{P} -\infty$;

(ii) if $\gamma > 0$, $T_\tau \xrightarrow{P} +\infty$.

4 Numerical illustration

The performance of the test in the regression model

$$Y_i = \beta_0 + x_i\beta_1 + E_i, \quad i = 1, \ldots, n$$

is studied on the simulated values. The power of the test is illustrated by means of the frequency of rejections under various error distributions. The chosen values of the parameter $\boldsymbol{\beta}$ are $\beta_0 = 1$, $\beta_1 = 3$. The regressors $x_1, \ldots, x_n$ were simulated for $n = 1000$ from the uniform distribution, independently of the errors, which the distributions were generated from the Pareto, exponential and Student distributions.

1000 replications of linear regression model were simulated for each case, and the test statistics T_τ were computed for $\tau = 1 - k/n$, $k = 3, \ldots, 997$. Figures 1-3 show estimated type I error probability, respectively the empirical power.

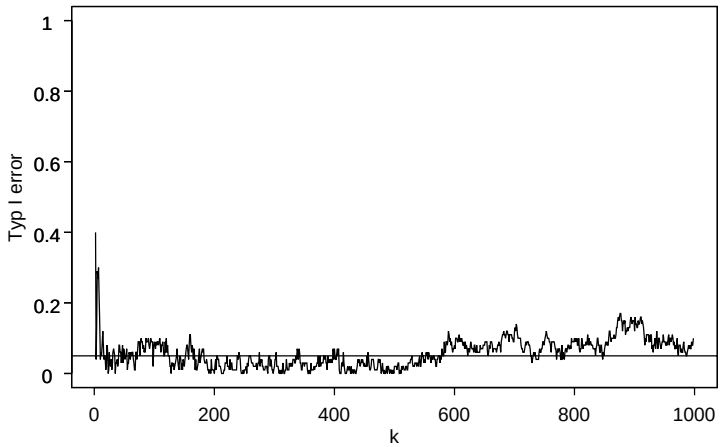

Figure 1. Estimated type I error probability of T_τ at a level $\alpha = 0.05$ for exponential distribution against $\tau = 1 - k/n$, $k = 3, \ldots, 997$.

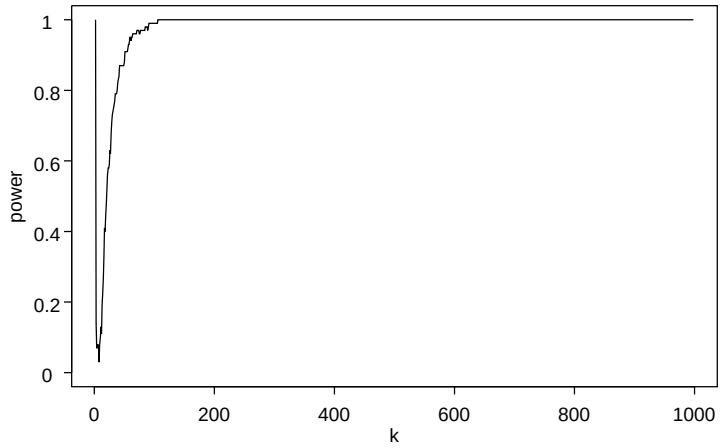

Figure 2. Empirical power of T_τ at a level $\alpha = 0.05$ for Pareto distribution ($\gamma = 1$) against $\tau = 1 - k/n$, $k = 3, \ldots, 997$.

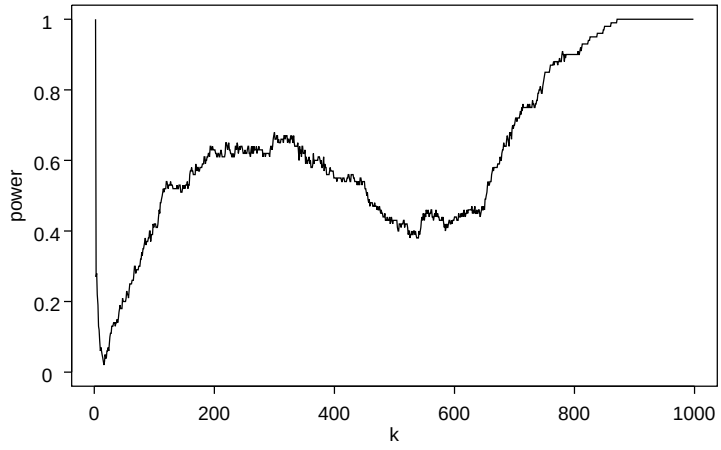

Figure 3. Empirical power of T_τ at a level $\alpha = 0.05$ for Student distribution (3 d.f.) against $\tau = 1 - k/n$, $k = 3, \ldots, 997$.

Acknowledgments

Research was supported by Research Project LC06024. The authors thank two referees for their careful reading and for their comments, which helped to improve the text.

References

[1] Castillo, E. Galambos, J. and Sarabia, J.M. (1989). The selection of the domain of attraction of an extreme value distribution from a set of data. In: *Extreme Value Theory*, (J. Hüsler and R.-D. Reiss eds) Lecture Notes in Statistics 51, Springer, Berlin-Heidelberg, 181-190.

[2] de Haan, L. and Ferreira, A. (2006), *Extreme Value Theory: An Introduction*, Springer Verlag.

[3] Dienstbier, J. (2009), Estimators of the extreme value index based on quantile regression. *RevStat*, to appear

[4] Drees, H. (1998), On smooth statistical tail functionals, *Scand. J. Statist.*, **25**, 187–210.

[5] Fraga Alves, M.I. and Gomes, M.I. (1996), Statistical Choice of Extreme Value domains of attraction - a comparative analysis. *Commun. Statist.-Theory Meth.*, **(25)4**, 789–811.

[6] Gnedenko, B.V. (1943), Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.*, **44**, 423–453.

[7] Hasofer, A.M. and Wang, Z. (1992), A test for extreme value domain of attraction. *JASA*, **87**, 171–177.

[8] Jurečková, J. (1999), Regression rank scores tests against heavy-tailed alternatives, *Bernoulli*, **5**, 659-676.

[9] Jurečková, J. and Sen, P.K. (1996), *Robust Statistical Procedures: Asymptotics and Interrelations* John Wiley & Sons., New York

[10] Koenker, R. and Bassett, G. (1978), Regression quantiles. *Econometrica* 46 : 33–50.

[11] Marohn, F. (1998), Testing the Gumbel hypothesis via the POT-method. *Extremes* **1:2**, 191–213.

[12] Neves, C., Picek, J. and Alves, F.M.I. (2006), The contribution of the maximum to the sum of excesses for testing max-domains of attractions. *J. Statist. Planning Infer.* **136** (4), 1281–1301.

[13] Segers, J. and Teugels, J. (2001), Testing the Gumbel hypothesis by Galton's ratio. *Extremes*, **3:3**, 291–303.