Novi Sad J. Math.
Vol. 38, No. 3, 2008, 97-104

INCOMPLETE SAMPLES AND TAIL ESTIMATION FOR STATIONARY SEQUENCES [1]

Ivana Ilić[2], Pavle Mladenović[3]

Abstract. Let (X_n) be a strictly stationary sequence with a marginal distribution function F such that $1 - F(x) = x^{-\alpha}L(x)$, $x > 0$, where $\alpha > 0$ and $L(x)$ is a slowly varying function. We assume that only observations of (X_n) are available at certain points. Under assumption of weak dependency we proved the consistency of Hill's estimator of the tail index α based on an incomplete sample from $\{X_1, X_2 \ldots, X_n\}$. This is an extension of the results of Hsing [15] and Mladenović and Piterbarg [19].

AMS Mathematics Subject Classification (2000): Primary 60G70; Secondary 60G10

Key words and phrases: Hill's estimator, index of regular variation, stationary sequences, weak dependence, missing observations

1. Introduction

Let F be a distribution function with regularly varying uper tail, that is

$$1 - F(x) = x^{-\alpha}L(x), \quad x > 0, \tag{1.1}$$

where $\alpha > 0$ and L is slowly varying at infinity. Without loss of generality we may assume that $F(0) = 0$. The problem of estimating the tail index α has attracted a great attention among statisticians. There is a huge number of papers concerning this problem in i.i.d. settings, i.e. when the estimator is defined using a sample of independent random variables $X_1, \ldots, X_n$ distributed according to F. Probably, the most popular is the Hill estimator defined as follows [see Hill (1975)]: Let $X_{(1)} \geqslant X_{(2)} \geqslant \ldots \geqslant X_{(n)}$ be a sequence of order statistics. Based on $k+1$ largest of them, Hill's estimator is

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \ln X_{(i)} - \ln X_{(k+1)}. \tag{1.2}$$

Asymptotic behavior of Hill's estimator was studied by many authors under different conditions. Here the number $k = k(n)$ should also increase together

[1]This work was supported by the Joint Research Project financed by The Ministry of Education and Science of the Republic of Macedonia (MESRM) (Project No.17-1383/1) and The Scientific and Technical Research Council of Turkey (TUBITAK) (Project No. TBGA-U-105T056).

[2]University of Niš, Medical Faculty, Dept. of mathematics and informatics, Bul. dr Zorana Djindjića 82, 18000 Nish, Serbia, e-mail: ivana@medfak.ni.ac.yu

[3]University of Belgrade, Faculty of Mathematics, Studentski trg 16, 11000 Belgrade, Serbia, e-mail: paja@matf.bg.ac.yu

with n. Mason [18] (1982) proved weak consistency under conditions $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. Deheuvels, Haeusler and Mason [7] (1988) proved strong consistency for any sequence $k = k(n)$ such that $k/\ln \ln n \to \infty$ and $k/n \to 0$ as $n \to \infty$. For results concerning asymptotic normality of Hill's estimator see, for example, Davis and Resnick [4] (1984), Csörgó and Mason [3] (1985), Haeusler and Teugels [12] (1985), Goldie and Smith [9] (1987), Hall [13] (1982), Hill [14](1975) and Beirlant and Teugels [1] (1989). Dekkers, Einmahl and de Haan [5] (1989) extended Hill's estimator for the index of regular variation to an estimate for the index of an extreme-value distribution. Also see Dekkers and de Haan [6] (1989), Gnedenko [10](1943) and de Haan [11] (1970).

Some other estimators for extreme value index in i.i.d. settings were also proposed and studied: see Pickands [20] (1975), Chörgó, Deheuvels and Mason [3] (1985), Dekkers and de Haan [5] (1989) and Drees [8](1998). There are also relatively small number of papers devoted to estimation of the tail index using dependent data, see Hsing (1991) [15], Resnick and Starica [22, 23, 24] (1995, 1997, 1998), where asymptotic behavior of Hill's estimator was considered. Also see Seneta [25](1976) and Smith [26](1987).

2. Some preliminaries and notation

Let $(X_n)_{n\geqslant 1}$ be a strictly stationary sequence of random variables with "short range" dependence, that is to say that the finite dimensional distributions of (X_n) are invariant under shifts and the dependence between observations from (X_n) becomes weaker as time separation becomes larger. Moreover, we assume that only observations at certain points are available. Denote observed random variables among $\{X_1, \ldots, X_n\}$ by $\widetilde{X}_1$, ..., $\widetilde{X}_{S_n}$. Here the random variable S_n represents the number of observed rv's among the first n terms of the sequence (S_n). Incomplete sample can be obtained, for example, if every term of (X_n) is observed with probability p, independently of other terms, and in this case S_n is binomial random variable. But we shall assume that observed random variables are determined by a general point process, and only conditions on S_n will be imposed. See Leadbetter, Lindgren and Rootzén [17] (1983) and Resnick [21] (1987).

Assumption. $X_1, X_2, \ldots$ does not depend on S_n and there exists a sequence of real numbers (γ_n) such that:

$$\frac{S_n}{\gamma_n} \to_p c_0 > 0 \qquad \text{as} \quad n \to +\infty$$

and

$$\lim_{n\to+\infty} \gamma_n = \infty.$$

Suppose β_n is a sequence of real numbers such that

$$\lim_{n\to\infty} \beta_n = \infty$$

and

$$\lim_{n\to\infty} \frac{\beta_n}{\gamma_n} = 0$$

Let

$$M_n = \left[\frac{S_n}{\beta_n}\right] \quad \text{and} \quad B_n = \begin{cases} 0, & S_n = 0 \\ \frac{M_n}{S_n}, & S_n \geq 1 \end{cases}$$

We are interested in estimation of α, using some portion of a sample. Let $X_{1,S_n} \geq X_{2,S_n} \geq ... \geq X_{S_n,S_n}$ be the order statistics defined by S_n observed variables. Define

$$F^{-1}(y) = \inf\{x : F(x) \geq y\}, \qquad 0 < y < 1.$$

Let us also denote, for every $x \in R$, x_+ is $max(x, 0)$. Hill's estimator is given by:

$$H_{S_n} = \begin{cases} \frac{1}{M_n} \sum_{j=1}^{M_n} \ln X_{j,S_n} - \ln X_{M_{n+1},S_n}, & S_n \geq \beta_n \\ 0, & S_n < \beta_n \end{cases}$$

Let us also define:

$$\widetilde{H}_{S_n} = \begin{cases} \frac{1}{M_n} \sum_{j=1}^{M_n} \ln X_{j,S_n} - \ln F^{-1}(1 - B_n), & S_n \geq \beta_n, \\ 0, & S_n < \beta_n, \end{cases}$$

$$H^+_{S_n} = \begin{cases} \frac{1}{M_n} \sum_{j=1}^{S_n} (\ln \widetilde{X}_j - \ln F^{-1}(1 - B_n))_+, & S_n \geq \beta_n, \\ 0, & S_n < \beta_n. \end{cases}$$

Suppose $\widetilde{Y}_i$ (Y_i) is a functional of $\widetilde{X}_i$ (X_i), for example, $\widetilde{Y}_i$ may be:

$$(\ln \widetilde{X}_i - \ln F^{-1}(1 - B_n))_+,$$

or:

$$I\{\ln \widetilde{X}_i > lnF^{-1}(1 - B_n) + \epsilon\}.$$

Let $F_a^b\{Y_i\}$ be the σ-field; $\sigma\{Y_i : a \leq i \leq b\}$ and for $1 \leq l \leq n-1$ let:

$$\beta(l, \{Y_i\}) = \sup\{|P(A \cap B) - P(A)P(B)| : \\ A \in F_1^k\{Y_i\}, B \in F_{k+l}^n\{Y_i\}, \ 1 \leq k \leq n - l\}.$$

3. Results

Theorem 3.1. *Suppose (r_n) is a sequence of positive integers and $\frac{r_n}{\gamma_n} \to 0$, when $n \to \infty$. Let $\widetilde{S}_{nk}$ be a random variable measurable with respect to $F^{kr_n}_{(k-1)r_{n+1}}\{\widetilde{Y}_i\}$, where $\widetilde{Y}_i$ is a functional of $\widetilde{X}_i$ and $1 \leq k \leq K_n$, where $K_n = [\frac{S_n}{r_n}]$. Assume that:*

$$(a) \qquad \frac{n}{r_n}\beta(r_n, \{Y_i\}) \to 0,$$

$$(b) \qquad I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{k=1}^{K_n} E|\widetilde{S}_{nk}|I\{|\widetilde{S}_{nk}| > M_n\} \to_p 0,$$

$$(c) \qquad I\{S_n \geq \beta_n\}\frac{1}{M_n^2}\sum_{k=1}^{K_n} E(\widetilde{S}_{nk})^2 I\{|\widetilde{S}_{nk}| \leq M_n\} \to_p 0.$$

Then:

$$I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{k=1}^{K_n}(\widetilde{S}_{nk} - E\widetilde{S}_{nk}) \to_p 0.$$

Proof. Write:

$$I\{S_n \geq \beta_n\}\tfrac{1}{M_n}\sum_{k=1}^{K_n}(\widetilde{S}_{nk} - E\widetilde{S}_{nk}) = I\{S_n \geq \beta_n\}\tfrac{1}{M_n}\sum_{k=1,\ k\ odd}^{K_n}(\widetilde{S}_{nk} - E\widetilde{S}_{nk})$$
$$+I\{S_n \geq \beta_n\}\tfrac{1}{M_n}\sum_{k=2,\ k\ even}^{K_n}(\widetilde{S}_{nk} - E\widetilde{S}_{nk}).$$

Let us denote the set of all odd numbers in $\{1, 2, ..., K_n\}$ by O_{Sn}. It was proven in Hsing [15], using results of Ibragimov and Linnik [16], and condition (a) that, in the case when all the data are present, the variables $\widetilde{S}_{nk}$ for the set of all odd k $\in \{1, 2....n\}$ were treated as independent.

From above, we can proceed by assuming that $\widetilde{S}_{nk}$ are independent for every $k \in O_{Sn}$. Define:

$$S^*_{nk} = \widetilde{S}_{nk}I(|\widetilde{S}_{nk}| \leq M_n), \quad 1 \leq k \leq K_n.$$

For every $\epsilon > 0$,

$$\begin{aligned}
&I\{S_n \geq \beta_n\}P\left\{\frac{1}{M_n}\left|\sum_{k\in O_{s_n}}(\widetilde{S}_{nk}-ES^*_{nk})\right|>\epsilon\right\}\\
&\leq I\{S_n \geq \beta_n\}P\{\widetilde{S}_{nk}\neq S^*_{nk},\ \text{for some } k\in O_{S_n}\}\\
&+I\{S_n \geq \beta_n\}P\left\{\frac{1}{M_n}\left|\sum_{k\in O_{S_n}}(S^*_{nk}-ES^*_{nk})\right|>\epsilon\right\}\\
&\leq I\{S_n \geq \beta_n\}\sum_{k\in O_{S_n}}P\{|\widetilde{S}_{nk}|>M_n\}+I\{S_n \geq \beta_n\}\frac{1}{M_n^2}\frac{1}{\epsilon^2}Var(\sum_{k\in O_{S_n}}S^*_{nk}))\\
&\leq I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{k\in O_{S_n}}E|\widetilde{S}_{nk}|+I\{S_n \geq \beta_n\}\frac{1}{M_n^2}\frac{1}{\epsilon^2}\sum_{k\in O_{S_n}}Var(S^*_{nk})\\
&\leq I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{k\in O_{S_n}}E|\widetilde{S}_{nk}|+I\{S_n \geq \beta_n\}\frac{1}{M_n^2}\frac{1}{\epsilon^2}\sum_{k\in O_{S_n}}E(S^*_{nk})^2\\
&\leq I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{k\in O_{S_n}}E|\widetilde{S}_{nk}|\\
&\qquad +I\{S_n \geq \beta_n\}\frac{1}{M_n^2}\frac{1}{\epsilon^2}\sum_{k\in O_{S_n}}E(\widetilde{S}_{nk})^2 I(|\widetilde{S}_{nk}|\leq M_n)\to_p 0.
\end{aligned}$$

We used the fact that $I^2\{|\widetilde{S}_{nk}|\leq M_n\}=I\{|\widetilde{S}_{nk}|\leq M_n\}$.

Since the following equality holds

$$\begin{aligned}
\widetilde{S}_{nk} &= \widetilde{S}_{nk}I\{|\widetilde{S}_{nk}|>M_n\}+\widetilde{S}_{nk}I\{|\widetilde{S}_{nk}|\leq M_n\}\\
&= \widetilde{S}_{nk}I\{|\widetilde{S}_{nk}|>M_n\}+S^*_{nk}
\end{aligned}$$

we have that

$$\begin{aligned}
&I\{S_n \geq \beta_n\}\frac{1}{M_n}\left|\sum_{k\in O_{S_n}}(E\widetilde{S}_{nk}-ES^*_{nk})\right|\\
&= I\{S_n \geq \beta_n\}\frac{1}{M_n}\left|\sum_{k\in O_{S_n}}E\widetilde{S}_{nk}I\{|\widetilde{S}_{nk}|>M_n\}\right|\\
&\leq I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{k\in O_{S_n}}E|\widetilde{S}_{nk}|I\{|\widetilde{S}_{nk}|>M_n\}\to_p 0.
\end{aligned}$$

Finally, we have that

$$\begin{aligned}
&I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{k\in O_{S_n}}(\widetilde{S}_{nk}-ES^*_{nk}-(E\widetilde{S}_{nk}-ES^*_{nk}))\\
&= I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{k\in O_{S_n}}(\widetilde{S}_{nk}-E\widetilde{S}_{nk})\to_p 0.
\end{aligned}$$

We have similar deduction for even numbers k from the set $\{1, 2, ..., K_n\}$. $\square$

Theorem 3.2. *All three quantities* H_{S_n}, $H^+_{S_n}$, *and* $\widetilde{H}_{S_n}$ *converge to* α^{-1} *in probability under the following conditions.*

(i) There exists a sequence (r_n) *of positive integers such that* $\frac{r_n}{\gamma_n} \to 0$, *and that* $\frac{n}{r_n}\beta(r_n, \{\widetilde{Y}_i\}) \to 0$. *Let us denote* $\widetilde{Y}_i = (\ln \widetilde{X}_i - \ln F^{-1}(1 - B_n))_+$ *and suppose that (b) and (c) from Theorem 3.1 hold for* $\widetilde{S}_{nk} = \sum_{i=(k-1)r_n+1}^{kr_n} \widetilde{Y}_i$.

(ii) For each $\varepsilon \in R$ *and* ρ *in some interval containing 1 there exists a sequence* (r_n) *of positive constants for which* $\frac{r_n}{\gamma_n} \to 0$, *such that* $\frac{n}{r_n}\beta(r_n, \{\widetilde{I}_i\}) \to 0$, *where* $\widetilde{I}_i = I\{\ln \widetilde{X}_i > \ln F^{-1}(1 - \rho B_n) + \varepsilon\}$. *Suppose that (b) and (c) from Theorem 3.1 hold for* $\widetilde{S}_{nk} = \sum_{i=(k-1)r_n+1}^{kr_n} \widetilde{I}_i$, *where* $K_n = \left[\frac{S_n}{r_n}\right]$.

(iii) $r_n E\left(I\{S_n \geq \beta\}\frac{1}{M_n}\right) \to 0$.

Proof. It follows from Theorem 3.1 and the condition (i) that the following relations hold:

$$I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{k=1}^{K_n}\sum_{i=(k-1)r_n+1}^{kr_n}(\widetilde{Y}_i - E\widetilde{Y}_i) \to_p 0,$$

$$I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{i=(k-1)r_n+1}^{K_n r_n}(\widetilde{Y}_i - E\widetilde{Y}_i) \to_p 0.$$

It follows from (iii) that the positive quantity

$$I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{i=K_n r_n+1}^{S_n}\widetilde{Y}_i$$

has the expectation tending to 0. Consequently we conclude that:

$$I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{i=1}^{S_n}(\widetilde{Y}_i - E\widetilde{Y}_i) \to_p 0.$$

Condition (ii) implies that:

$$I\{S_n \geq \beta_n\}\frac{1}{M_n}\sum_{i=1}^{S_n}(I_i - EI_i) \to_p 0.$$

Finally, the conclusion of the theorem is a consequence of Theorem 1 from Mladenović and Piterbarg [19]. $\square$

Acknowledgement

The work is supported by the Ministry of Science and Environmental Protection of the Republic of Serbia, Grant No. 144032 and Grant No. 149041.

References

[1] Beirlant, J., Teugels, J.L., Asymptotic normality of Hill's estimator. In: Hüsler, J. and Reiss, R.D. (Eds.), Lecture Notes in Statistics, vol. 51., Berlin: Springer, pp. 148-155, (1989).

[2] Bingham, N., Goldie, C., Teugels, J., Regular Variation. Encyclopedia of Mathematics and its Applications, vol. 27., Cambridge, UK: Cambridge University Press, (1987).

[3] Chörgó, S., Deheuvels, P., Mason, D.M.,: Kernel estimates of the tail index of a distribution. Ann. Statist. 13 (1985), 1050-1077.

[4] Davis, R.A., Resnick, S.T., Tail estimates motivated by extreme value theory. Ann. Statist. 12 (1984), 1467-1487.

[5] Dekkers, A.L.M., Einmahl J.H.J., de Haan, L., A moment estimator for the index of an extreme-value distribution. Ann. Statist. 17 (1989), 1833-1855.

[6] Dekkers, A.L.M., de Haan, L., On the estimation of the extreme value index and large quantile estimation. Ann. Statist. 17 (1989), 1795-1832.

[7] Deheuvels, P., Haeusler, E., Mason, D.M., Almost sure convergence for the Hill estimator. Math. Proc. Cambridge Philos. Soc. 104, (1988), 371-381.

[8] Drees, H. (1998). A general class of estimators of the extreme value index. Journal of Statistical Planning and Inference **66**, 95-112.

[9] Goldie, C.M. and Smith, R.L. Slow variation with remainder: Theory and applications. Quart. J. Math. Oxford Ser. (2) **38** (1987) 45-71.

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

[11] de Haan, L., On Regular Variation and its Application to the Weak Convergence of Sample Extremes. Mathematical Centre Tracts 32, Amsterdam, 1970.

[12] Haeusler, E., Teugels, J.L., On asymptotic normality of Hill's estimator for the exponent of regular variation. Ann. Statist. 13 (1985), 743-756.

[13] Hall, P., On some simple estimates of an exponent of regular variation. J. Roy. Statist. Soc. Ser. B 44 (1982), 37-42.

[14] Hill, B.M., A simple general approach to inference about the tail of a distribution. Ann. Statist. 3 (1975), 1163-1174.

[15] Hsing, T., On tail index estimation using dependent data. Ann. Statist. 19 (1991), 1547-1569.

[16] Ibragimov, I.A. and Linnik, Y.V. Independent and Stationary Sequences of Random Variables. Wolters-Noordhoff, Groningen (1969).

[17] Leadbetter, M.R., Lindgren, G., Rootzén, H., Extremes and Related Properties of Random Sequences and Processes. New York–Heidelberg–Berlin: Springer-Verlag, (1983).

[18] Mason, D.M.,Laws of large numbers for sums of extreme values. Ann. Probab. 10 (1982), 754-764.

[19] Mladenović, P., Piterbarg, V., On estimation of the exponent of regular variation using a sample with missing observations. Statist. Prob. Letters. 78 (2008), 327-335.

[20] Pickands, J., Statistical inference using extreme order statistics. Ann. Statist. 3 (1975), 119-131.

[21] Resnick, S.I., Extreme Values, Regular Variation and Point Processes. New York, Berlin, Heidelberg, London, Paris, Tokyo: Springer, (1987).

[22] Resnick, S. Starica, C., Consistency of Hill's estimator for dependent data. J. Appl. Probab. 32 (1995), 139-167.

[23] Resnick, S., Starica, C., Asymptotic behavior of Hill's estimator for autoregressive data. Stochastic Models 13 (1997), 703-723.

[24] Resnick, S. and Starica, C., *Tail index estimation for dependent data.* Ann. Appl. Probab. 8 (1998), 1156-1183.

[25] Seneta, E., Regularly varying functions. In: Lecture Notes in Mathematics, vol. 508., New York: Springer, 1976.

[26] Smith, R.L., Estimating tails of probability distributions. Ann. Statist. 15 (1987), 1174-1207.

Received by the editors October 1, 2008