A Simulation comparison of estimators of tail index

under right random censoring

مقارنة محاكاة لمقدري مؤشر الذيل تحت الرقابة العشوائية اليمنى

Hanane Arbia[1],* , **Badreddine Talbi**[2]

[1] Kasdi Merbah University, Ourgla, Algeria.

[2] Ecole Nationale Supérieure de Statistique et d'Economie Appliquée, Algiers, Algeria.

Received: 2020/11/19 Accepted: 2022/07/04

Abstract: Fabian et al. (2009) proposed a new estimator of extreme value index for heavy-tailed distributions, namely t-Hill estimator. We are interested in this paper on the extreme value theory under right random censoring. We adapted this estimator for this kind of data (censored data). A Simulation comparison of Hill estimator in the case of censoring and our estimator (adapted t-Hill estimator) show the robustness of the last one.

Keywords: t-Hill estimator; random censoring; simulation; robustness.

Résumé : Fabian et al. (2009) ont proposé un nouvel estimateur de l'indice des valeurs extrêmes pour les distributions à queue lourde, estimateur de t-Hill. Nous sommes intéressés dans cet article sur la théorie des valeurs extrêmes sous censure aléatoire droite. Nous avons adapté cet estimateur pour ce type de données (données censurées). Une comparaison par simulation de l'estimateur de Hill dans le cas de la censure et de notre estimateur (estimateur t-Hill adapté) montre la robustesse du dernier estimateur.

Mots-clés: estimateur t-Hill; censure aléatoire; simulation; robustesse.

ملخص : فابيان وأخرون (2009) اقترحوا مقدرا جديدا لمؤشر القيمة القصوى للتوزيعات ذات الذيل الثقيل, مقدرتهيل. نحن مهتمون في هذه الورقة بنظرية القيمة القصوى تحت الرقابة العشوائية اليمنى, قمنا بتكييف هذا المقدر لهذا النوع من البيانات (البيانات الخاضعة للرقابة). تظهر مقارنة المحاكاة لمقدار هيل في حالة الرقابة ومقدِّرنا تهيل المكيف مدى قوة هذا الأخير.

الكلمات المفتاح مقدرتهيل, الرقابة العشوائية, محاكاة, قوة.

* Hanane Arbia.

1- Introduction :

The extreme value theory is used to evaluate rare events. The applications of this theory is mainly based on the estimation of a parameter which gives the shape of the tail distribution; this parameter is called extreme value index. In the literature of the extreme values exists several estimators of this index, the most used are the estimator of Pikands Pickands et al. (1975), the estimator of moment Dekkers et al. (1989); this two estimators is for any values of the tail index, and the last one is the estimator of Hill Hill (1975), it is for positive tail index. We are interested in our works in the recent version of Hill estimator. The estimator of Hill Hill (1975) is the most common estimator for positive tail index.

It is defined by

$$\hat{\gamma}^{(\mathrm{Hill})} = \frac{1}{k}\sum_{i=1}^{k} log\frac{X_{n-i+1,n}}{X_{n-k,n}},$$

or $X_{1,n}, X_{2,n}, \dots, X_{n,n}$ are the statistics of order.

Consistency of this estimator was established by Mason (1982) and Deheuvels et al (1988). The asymptotic normality of Hill estimator was showed in de Haan et al. (2006).
A disadvantage of Hill estimator based on the maximum likelihood, it is not robust, and then it is sensitive to some observations.

Fabian et al. (2009) found a solution to this problem in the theory of extreme values, by construction of a robust estimator called t-Hill, it is defined by

$$\hat{\gamma}^{(\mathrm{tHill})} = \frac{1}{\frac{1}{k}\sum_{i=1}^{k}\frac{X_{n-k,n}}{X_{n-i+1,n}}} - 1,$$

t-Hill estimator is based on the harmonic mean, therefore it is robust (see Stehlik et al. (2012)), so it can give more realistic values for a large numbers of extreme values. Beran et al. (2013) proved the asymptotic normality of this estimator.

In the case of censored data, the estimation of the extreme value index approached recently by Beirlant et al. (2007) and Einmahl et al. (2008). Our aim in this paper is to estimate the tail index of right censured data. In this case the variable of interest

X is not completely observed, but it censured by another random variable Y independent of X.

So we are only observed $(Z_i; \delta_i)$ such as
$Z_i = min(X_i, Y_i)\ and\ \delta_i = \mathbb{I}\{X_i \leq Y_i\}$, 1 ≤ i ≤ n.

The random variables X, Y and Z have respectively the distributions F, G, and H, with extreme values indexes γ_1, γ_2, and γ respectively.

The goal is to estimate γ_1 index of F; several estimators of this index are introduced, they built by the same way, they based on usual estimators (Hill, moment, Pikands, ...):

$$\hat{\gamma}_1^{(c,n)} = \frac{\hat{\gamma}_{Z,k}}{\frac{1}{k}\sum_{j=1}^{k} \delta_{n-j+1}}.$$

So the estimator for censored data is given as a ratio of two quantities; the numerator which estimate the index extreme values for the variable Z not X, it divided by proportion of the uncensored data for estimated γ_1 index of F. Einmahl et al. (2008) proved the consistency and the asymptotic normality of $\hat{\gamma}_1$.

Our goal in this paper is organized as follows. In Section 2 the existing estimator is presented and we define the proposed estimator. In Section 3 we illustrate the robustness of our estimator by simulation comparison.

2–Estimators of tail index under right random censoring:

Let $X_1, X_2, \dots, X_n$ a realisation of random variable X (interest variable) with distribution function F and end-points x_F $(x_F = sup\{x, F(x) < 1\})$, censured by anathor random variable Y (independent of X) with distribution function G and end-points x_G $(x_G = sup\{x, G(x) < 1\})$. So X not be observed, the sample $(Z_i; \delta_i)$ is only observed,

$$\pm_i = \begin{cases} 1 & si\ X_i \leq Y_i \\ 0 & si\ X_i > Y_i \end{cases} \quad 1 \leq i \leq n, Z_i = min(X_i, Y_i)\ and$$

δ_i is an indicator variable, it determines if X has been censored or not. The estimation of the extreme value index under right random censoring is our aim Einmahl et al. (2008) suggested the folowing cases

$$\begin{cases} \text{case 1: } \gamma_1 > 0,\ \gamma_2 > 0 \\ \text{case 2: } \gamma_1 < 0,\ \gamma_2 < 0\ , x_F = x_G \\ \text{case 3: } \gamma_1 < 0,\ \gamma_2 < 0\ , x_F = x_G = +\infty \end{cases}$$

In our work, we consider the case 1, where the two distribution F and G are in the Pareto domain of attraction. In this case $\gamma = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2}$.

Beirlant et al. (2007) and Einmahl et al. (2008) proposed an estimator of tail index by adaptation of the classical estimators to censoring. Adapted Hill estimator is one of this proposed estimators when the tail index is positive, it defined by

$$\hat{\gamma}_1^{(c,Hill)} = \frac{\hat{\gamma}_{Z,k}^{(Hill)}}{\hat{p}}. \qquad (2.1)$$

Where $\hat{\gamma}_{Z,k}^{(Hill)}$ is the classical Hill estimator based in the sample Z:

$$\hat{\gamma}_{Z,k}^{(Hill)} = \frac{1}{k}\sum_{i=1}^{k} log \frac{Z_{n-k,n}}{Z_{n-i+1,n}},$$

and $\hat{p} = \frac{1}{k}\sum_{j=1}^{k} \delta_{n-j+1}$.

The asymptotic normality of adapted Hill estimator established recently by Brahimi et al. (2014), to approximate this estimator, they used the empirical process theory.

We used the approach defined in Beirlant et al. (2007) and Einmahl et al. (2008) to the so-called t-Hill estimator: it based in dividing the classical t-Hill estimator for Z by the proportion of the uncensored data. Our proposed estimator is:

$$\hat{\gamma}_1^{(c,tHill)} = \frac{\hat{\gamma}_{Z,k}^{(tHill)}}{\hat{p}}. \qquad (2.2)$$

is t-Hill estimator of tail index for the variable Z:$\hat{\gamma}_{Z,k}^{(tHill)}$

$$\hat{\gamma}_{Z,k}^{(tHill)} = \frac{1}{\frac{1}{k}\sum_{i=1}^{k} \frac{Z_{n-k,n}}{Z_{n-i+1,n}}} - 1.$$

Lastly, we assess the performance of these estimators by a simulation study.

3- Simulation study

In this section, we presented a simulation study for various sample sizes, to compare the performance of the adapted Hill and adapted t-Hill estimators for illustrate the robustness of our estimator. The following steps are proceed to derive the performance of estimators above.

S1: Generate 100 samples of size n (300, 600, 1000 and 2000) from the distributions F and G (of X and Y) whith tail index γ_1 and γ_2 respectively (Parto model).

S2: Choose a different values of proportion of the uncensored data, small and big value of p (0,25 and 0,85) (the percentage of censoring in the right tail is 75% and 15% respectively), such as

$\gamma_2 = \frac{p\gamma_1}{1-p}$.

S3: Consider the sample (Z_i, δ_i) where
$Z_i = min(X_i, Y_i)\ and\ \delta_i = \mathbb{1}\{X_i \leq Y_i\}, 1 \leq \mathrm{i} \leq \mathrm{n}.$

S4: Estimate γ_1 by the adapted Hill and the adapted t-Hill estimators $\hat{\gamma}_1^{(c,Hill)}$ and $\hat{\gamma}_1^{(c,tHill)}$ defined in (2.1) and (2.2) respectively. To determine the optimal numbers of upper extremes, we utilize

$$k_{opt} = argminMSE[\hat{\gamma}_1^{(c,.)}(k)].$$

S5: Calculate the empirical bias and Mean Square Error (MSE) to measure the performance of the estimators $\hat{\gamma}_1^{(c,Hill)}$ and $\hat{\gamma}_1^{(c,tHill)}$

$$bias = \frac{1}{R}\sum_{j=1}^{R}(\hat{\gamma}_{1,j}^{(c,.)} - \gamma_1)$$

and

$$MSE = \frac{1}{R}\sum_{j=1}^{R}[(\hat{\gamma}_{1,j}^{(c,.)} - \gamma_1)]^2.$$

The corresponding tables contains our results for two different values of p (0.25 and 0.85) and γ_1=1.5

Table (3.1): Empirical bias and MSE of the two estimators based on 100 samples of size n from Pareto model with γ_1=1.5 and p=0.25.

		$\hat{\gamma}_1^{(c,tHill)}$		$\hat{\gamma}_1^{(c,Hill)}$
N	**Bias abs**	**MSE**	**Bias abs**	**MSE**
300	0.1011508	0.1570170	0.1121830	0.1809540
600	0.0392026	0.0961756	0.0616084	0.1327581
1000	0.0213432	0.0392381	0.0281987	0.0658349
2000	0.0120958	0.0188821	0.0123296	0.0246296

Source: Results of simulation in R.

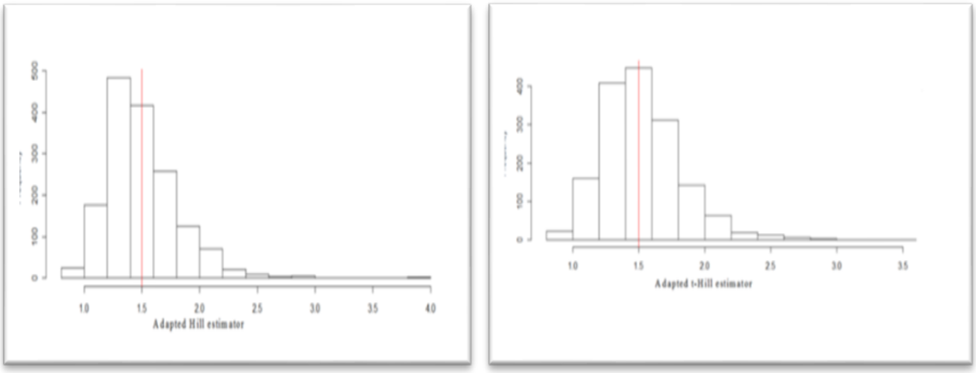
Table (3.2): Empirical bias and MSE of the two estimators based on 100 samples of size n from Pareto model with γ_1=1.5 and p=0.85.

	$\widehat{\gamma}_1^{(c,tHill)}$		$\widehat{\gamma}_1^{(c,Hill)}$	
N	**Bias abs**	**MSE**	**Bias abs**	**MSE**
300	0.0293469	0.0297518	0.0336163	0.0308642
600	0.0179344	0.0248831	0.0199505	0.0301349
1000	0.0088681	0.0125549	0.0119448	0.0160600
2000	0.0050819	0.0077457	0.0086366	0.0081563

Source: Results of simulation in R.

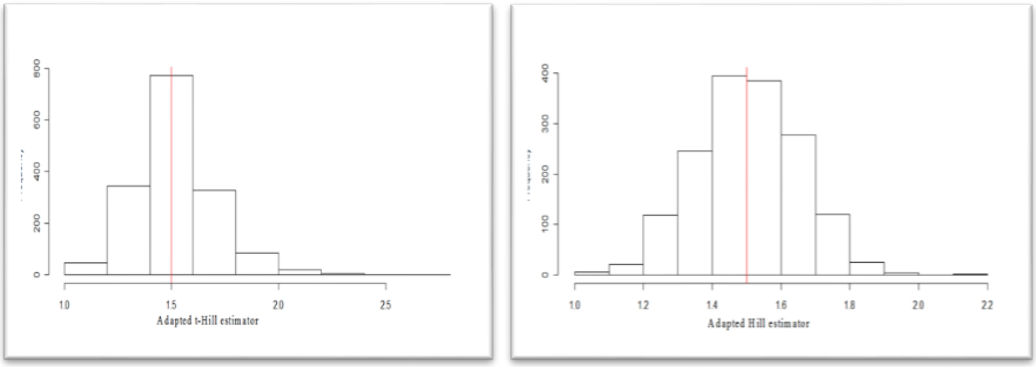
We can remark that when the censoring percentage decreases, the bias and MSE of the estimators $\widehat{\gamma}_1^{(c,Hill)}$ and $\widehat{\gamma}_1^{(c,tHill)}$ decreases. We also found that the adapted t-Hill estimator has less bias and MSE then the adapted Hill estimator for each p, it performs better for large simple size. This show the robustness of our estimator.

Figure (3.1): Histograms of the estimators $\widehat{\gamma}_1^{(c,tHill)}$ *and* $\widehat{\gamma}_1^{(c,Hill)}$ for γ_1=1.5 and p=0.25.



Source: Output of R.

Figure (3.2): Histograms of the estimators $\widehat{\gamma}_1^{(c,tHill)}$ *and* $\widehat{\gamma}_1^{(c,Hill)}$ for γ_1=1.5 and p=0.85.



Source: Output of R.

From this figures, histograms are well concentrated around the real tail index value γ_1=1.5 (the vertical line) in case where p=0.85 (high percentage of censoring in the right tail) and especially for our estimator $\widehat{\gamma}_1^{(c,tHill)}$.

4- Conclusion:

Estimation of the extreme value index under random censoring is a new research in theory of extreme value. Firstly it was reported in Reiss and Thomas et al. (1979), then it was approched recently by Beirlant et al. (2007) and Einmahl et al. (2008). In this paper we proposed a version of t-Hill estimator in the case of right censored data. Simulation study show that the proposed estimator is robust then adapted Hill estimator.

5- Bibliography

[1] Beirlant, J., Guillou, A., Dierckx, G., Fils-Villetard, A., 2007. Estimation of the extreme value index and extreme quantiles under random censoring. Extremes. 10, no. 3, 151-174.

[2] Beran, J., Schell, D. , Stehlík, M., 2013. The harmonic moment tail index estimator: asymptotic distribution and robustness. Ann. Inst. Statist. Math. 66, 193-220.

[3] Brahimi, B., Meraghni, D., Necir, A., 2014. Approximations of the tail index estimator of heavy-tailed distributions under random censoring and application. arXiv:1302.1666v5 [math.ST].

[4] Csörgő, S., 1996. Universal Gaussian approximations under random censorship. Ann. Statist. 24, no. 6, 2744-2778.

[5] Deheuvels, P., Häeusler, E., and Mason, D. M., 1988. Almost sure convergence of the Hill estimator. Math. Proc. Cambridge Philos. Soc., 104(02),371-381.

[6] Dekkers, A.L.M., de Haan, L., 1989. On the estimation of the extreme-value index and large quantile estimation. The annals of statistics 4, 1833-1855.

[7] Einmahl, J.H.J., Fils-Villetard, A. and Guillou, A., 2008. Statistics of extremes under random censoring. Bernoulli. 14, no.1, 207-227.

[8] Fabian, Z., 2001. Induced cores and their use in robust parametric estimation. Communication in Statistics.Theory and Methods 30:537.556.

[9] Fabian, Z., Stehlík, M., 2009. On robust and distribution sensitive Hill like method. IFAS Research Paper Series 2009-43.

[10] Gomes, M.I. and Neves, M.M., 2011. Estimation of the extreme value index for randomly censored data. Biometrical Letters. 48, no.1, 1-22.

[11] de Haan, L. and Stadtmüller, U., 1996. Generalized regular variation of second order. J. Australian Math. Soc. (Series A) 61, 381-395.

[12] de Haan L., Ferreira A., 2006. Extreme Value Theory. Springer, New York.

[13] Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. Ann. Statist. 3, no.5. 1163-1174.

[14] Kaplan, E. L., and Meier, P., 1958. Non parametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457-481.

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

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

[17] Reiss, R.-D., Thomas, M., 1997. Statistical analysis of extreme values. From insurance, finance, hydrology and other .elds, Birkh¨ auser Verlag, Basel.

[18] Shorack, G. A. and Wellner, J. A., 1986. Empirical Processes with Applications in Statistics. New York: Wiley.

[19] Stehlík, M., Fabian, Z., Stɥrelec, L., 2012. Small Sample Robust Testing for Normality against Pareto Tails, Communications in Statistics - Simulation and Computation, 41:7,1167-1194.

[20] Thomas, D.R., Pierce, D.A., 1979. Neyman.s smooth goodness-of-.t test when the hypothesis is composite. J. Amer. Statist. Assoc., 74, 441-445.