Extremes
https://doi.org/10.1007/s10687-018-0322-z

Discussion on "Human life is unlimited but short" by Holger Rootzén and Dmitrii Zholud

Chen Zhou[1]

Received: 17 April 2018 / Accepted: 18 April 2018

Keywords Data truncation · Asymptotic bias

1 Introduction

The recent stimulating work of Rootzén and Zholud (2017) provides a detailed analysis on the tail of the distribution of human life length. One of the important contributions of this work is to address the impact of sampling scheme in the data of the age of long-lived people. The striking finding is that the length of human life follows a distribution without a finite endpoint. The data used in this study is the International Database on Longevity (IDL) data[1] which underwent a careful data validation procedure. Validating the data of extreme ages is in general plausible and necessary. Nevertheless this discussion will focus on an unintended impact of data validation: using the validated data record may lead to a biased conclusion regarding the tail of the original dataset.

Rootzén and Zholud (2017) draw the conclusion that human life length is "unlimited but short" by fitting a sample of high ages above 110 to the generalized Pareto distribution (GPD). Based on the estimated shape parameter, the null of $\gamma = 0$ is not

[1] See the website http://www.supercentenarians.org

✉ Chen Zhou
c.zhou@dnb.nl; zhou@ese.eur.nl

[1] Economics and Research Division, De Nederlandsche Bank and Erasmus University Rotterdam, 1000AB Amsterdam, The Netherlands

Published online: 28 April 2018

rejected. Notice that the GPD with $\gamma = 0$ corresponds to the exponential distribution which has no finite endpoint. Rootzén and Zholud (2017) denote the three cases $\gamma > 0$, $\gamma = 0$ and $\gamma < 0$ as "unlimited life length", "unlimited but short life length" and "limited life length" and therefore conclude from the statistical analysis that human life length is "unlimited but short".

It is debatable whether the case $\gamma = 0$ should be interpreted as "unlimited but short". The rationale behind fitting the high ages to the GPD follows from the domain of attraction condition in Extreme Value Theory (EVT). Suppose the distribution function of the human life length, F, belongs to the domain of attraction of a generalised extreme value distribution with extreme value index γ. Then the excesses above a high threshold follow approximately the GPD with a shape parameter γ. Although the GPD with $\gamma = 0$, i.e. the exponential distribution, has no finite endpoint, the original distribution function F with the extreme value index $\gamma = 0$ may still have a finite endpoint, see e.g. Gnedenko (1943, pp. 445). Therefore, the fact that human life length possesses an extreme value index $\gamma = 0$ may not be directly associated to the conclusion of "unlimited human life".

Apart from the interpretation, this discussion aims at providing further insight on the potential distortion to the (estimation of) extreme value index when the data record is not complete. The essential idea follows the lines of data validation in the IDL project: data that are classified as "invalid" will be removed from the dataset. Here I make the assumption that for the record of human life length, the probability of having an invalid observation depends on the age: high ages are more likely to be missing or classified as invalid in the record. Such an assumption is reasonable as the record of death and birth might not have been systematically organized one century ago.

I show that if the data observed is an incomplete subset of the full dataset, the extreme value index of the observed data might differ from that of the full data. In particular, if the full dataset possesses an extreme value index $\gamma < 0$, i.e. with a finite endpoint, the observed dataset may possess a higher extreme value index that is closer to zero. In some example, it is also possible to have the observed dataset with an extreme value index $\gamma = 0$. The situation depends on the probability of data being classified as invalid. Even if the observed data may share the same extreme value index as that of the full data, due to the distortion on the second order property of the observed data, the estimated extreme value index may be biased towards zero.

To summarize, if the data record is incomplete, it is more likely to obtain an (estimated) extreme value index that is closer to zero, and therefore less likely to reject the null of $\gamma = 0$. Overlooking this potential distortion may result in misleading conclusions.

The incomplete data setup in this discussion should be distinguished from random data censoring. Einmahl et al. (2008) considered random censoring in extremes. In their model, the original data may not be observed if it exceeds another random variable. In such a case, the random threshold is observed and the occurrence of "random censoring" is recorded. In the random censoring case, the observed data also possesses a different extreme value index compared to the original data. Different from random censoring, the model in this discussion can be regarded as "random truncation". Once truncated, no further information is observed: neither the value of the truncated threshold, nor the occurrence of truncation.

The discussion is organised as follows. Section 2 provides the evaluation for the impact of incomplete data record. Section 3 discusses the practical consequences.

2 The impact of incomplete data record

Assume that the *full* dataset of the human life lengths consists of i.i.d. observations $X_1, \cdots, X_n$ drawn from a distribution F. For simplicity, I shall consider an example $F(x) = 1 - (\theta - x)^{-1/\gamma}$ with $\theta > 0, \gamma < 0$. This distribution has a finite endpoint θ with an extreme value index $\gamma < 0$. For each given X_i, a Bernoulli random variable B_i indicates whether X_i will be observed. I assume that the probability of the age X_i being observed is negatively associated to the age itself, i.e. $\Pr(B_i = 1|X_i) = g(\theta - X_i)$, where $g : [0, +\infty) \to [0, 1]$ is an increasing function.

An alternative view for this setup is to consider a series of i.i.d. positive random variables $\{Z_i\}_{i=1}^n$ with a common distribution function F_Z. Each X_i is observed, i.e. $B_i = 1$, if and only if $X_i < Z_i$. Consequently, $P(B_i = 1|X_i) = 1 - F_Z(X_i)$. Comparing with the model setup, we can interpret the function g as $g(t) = 1 - F_Z(\theta - t)$. With this alternative view, the model can be interpreted as random truncation: X_i is truncated by the random threshold Z_i. If the truncation occurs, no information is recorded: neither the value of the truncated threshold, nor the occurrence of truncation itself.

Eventually, the *observed* dataset is $\{X_i : B_i = 1\}$. Denote $Y_i = X_i B_i$ for $i = 1, 2, \cdots, n$. Since extreme value analysis, such as estimating the extreme value index, only employs large observations in the observed data, the statistical inference can be regarded as being based on $\{Y_i\}$ while disregarding the number of observations n. Consequently, I will study the tail region of the distribution of $Y = XB$ while omitting the subscript as follows:

$$\Pr(Y > x) = \Pr(X > x, B = 1) = \int_x^\theta g(\theta - t) dF(t) \tag{1}$$

With three examples of the function g, I shall discuss there implication for the distribution of Y and the corresponding estimation of the extreme value index.

Example 1 The power function: $g(x) = x^\beta$ for $\beta > 0$.

With this function g, I assume that as the age is getting close to the endpoint θ, the probability of being observed decreases in a power speed towards zero. Moreover, in the view of data truncation, the random threshold Z follows a distribution $F_Z(x) = 1 - g(\theta - x)$, which has the same endpoint θ as that of X.

In this case, the distribution of Y can be derived as follows: for $x < \theta$,

$$\Pr(Y > x) = \frac{1}{-\gamma\beta + 1}(\theta - x)^{\beta - 1/\gamma}.$$

Consequently, the distribution of Y has the same finite endpoint θ as the distribution function F, but a different extreme value index $-1/(\beta - 1/\gamma) > \gamma$. We conclude that the extreme value index of the observed data is higher than that of the full dataset, and closer to zero. In other words, the random truncation distorts the *true* value of the extreme value index.

This example can be generalized to any F that belongs to the Weibull domain of attraction. The proof of the general result is omitted.

Example 2 Power function with a drift: $g(x) = \min(c + x^{\beta}, 1)$ for $\beta > 0$ and $0 < c < 1$.

Compared to the first example, the only difference in this example is that as the age is getting close to the endpoint θ, the probability of being observed decreases but is still bounded away from zero. Moreover, in view of data truncation, the random threshold Z satisfies $\Pr(Z > \theta) = c > 0$. This can be interpreted as that Z has a higher endpoint than θ, which allows all potential ages to be observed with a positive probability.

In this case, the distribution of Y can be derived as follows. For $x > \theta - (1-c)^{1/\beta}$, we have that $g(\theta - x) = c + (\theta - x)^{\beta}$, which implies that

$$\Pr(Y > x) = c(\theta - x)^{-1/\gamma} + \frac{1}{-\gamma\beta + 1}(\theta - x)^{\beta - 1/\gamma}.$$

Consequently, the distribution of Y has the same finite endpoint θ and the same extreme value index γ as the distribution function F. In this example, the random truncation does *not* distort the *true* value of the extreme value index.

Nevertheless, the tail part of the distribution of Y has an additional second order term. More specifically, it satisfies the so-called second order condition as in de Haan and Stadtmüller (1996), with a second order index $\rho_Y = \gamma\beta$ and second order auxiliary function $A_Y(t) = \beta\frac{c^{\gamma\beta-1}}{-\gamma\beta+1}t^{\gamma\beta} > 0$. (Details of the derivation are omitted here.) Notice that the original distribution function F has a second order index $-\infty$. Hence, the second order behavior of the tail distribution is distorted by the random truncation.

This example can be generalized to any F that belongs to the Weibull domain of attraction while satisfying the second order condition with an index $\rho_F < \gamma\beta$. Again, the proof of the general result is omitted. In other words, if $\rho_F < \gamma\beta$, the second order index of Y is higher than ρ_F and closer to zero.

The distortion to the second order index has a direct impact on the estimation of the extreme value index. Notice that Rootzén and Zholud (2017) use the maximum likelihood estimator (MLE) for estimating the extreme value index index γ. Not only for the MLE, but also for most of the existing estimators of γ in the literature, the optimal level of tail observations used, usually denoted as k, is at the level $O(n^{2\rho/(2\rho-1)})$, where ρ is the second order index of the underlying distribution. Since $\rho_Y > \rho_F$, it is a direct consequence that the optimal k that can be used for the observed dataset is at a lower level compared to that used for the full dataset. Consequently, there is a higher level of estimation uncertainty when using the observed dataset.

In practice, for a given n, one usually choose a fixed k. For example, in Rootzén and Zholud (2017), $k = 566$. In such a case the estimator of the extreme value index using the observed dataset can be severely biased. The asymptotic normality of the MLE (see Drees et al. (2004)) shows that the bias of the MLE has the same sign as the second order auxiliary function A. Since $A_Y(t) > 0$, the estimator based on the observed data is biased upwards, i.e. closer to zero.

To summarize, this example shows that if all ages have some positive chance to be observed, the observed dataset may still possess the same *true* extreme value index as the full dataset. However, when estimating the extreme value index based on the observed dataset, the estimator is either more inaccurate due to a lower choice of the number of tail observations, or more biased towards zero, which makes it more difficult to be detected as significantly different from zero.

Lastly, I provide an artificial example to show that with some specific choice of the g function, the observed dataset may exhibit a completely different endpoint, even with $\gamma = 0$.

Example 3 $g(x) = 0$ for $x \in [0, c]$ and $g(x) = \beta(x - c)^{-\beta-1} \exp\left\{-(x - c)^{-\beta}\right\}$ for $x > c$ and $\beta > 0$.

With this function g, I simply assume that no age in the region $[\theta - c, \theta]$ is observable. Consequently, the endpoint of Y is reduced to $\theta - c$. Moreover, in view of data truncation, the random threshold Z a lower endpoint $\theta - c$ which does not allow the ages in the region $[\theta - c, \theta]$ to be observed.

For this example, as $x \to \theta - c$,

$$\begin{aligned}\Pr(Y > x) &\sim \int_x^{\theta-c} \beta(\theta - c - t)^{-\beta-1} \exp\left\{-(\theta - c - t)^{-\beta}\right\} \frac{1}{-\gamma} c^{-1/\gamma-1} dt \\ &= \exp\left\{-(\theta - c - x)^{-\beta}\right\} \frac{1}{-\gamma} c^{-1/\gamma-1}\end{aligned}$$

The derivation shows that Y has an endpoint $\theta - c$ and an extreme value index zero.

3 Concluding remarks

In this discussion, I demonstrate that if the dataset is incomplete due to data validation, the observed dataset may have a different tail behavior compared to the original dataset. If the original dataset has a negative extreme value index with a finite endpoint, the *true* extreme value index may be distorted by data truncation: the observed dataset may have an extreme value index closer to zero, or even equal to zero. Even if the distortion is only for the second order tail behavior, the estimator of the extreme value index based on the observed dataset will be biased towards zero and/or suffer from a high estimation uncertainty.

I remark that such a distortion is particularly pronounced in practice if the number of large observations used for estimating the extreme value index is limited. Notice that the MLE has a rate of convergence $1/\sqrt{k}$ where k is the number of large observations used in the estimation. Further, the asymptotic variance is $(\gamma + 1)^2$. To test the null hypothesis that $\gamma = 0$ at a significance level α, the point estimate of the extreme value index must be below $-\Phi^{-1}(1 - \alpha/2)/\sqrt{k}$ in order to obtain a significant result. For the often used significance level $\alpha = 0.05$, with $k = 566$ as in

Rootzén and Zholud (2017), I get that the threshold is -0.082. Therefore even if the point estimate is only slightly distorted by the positive bias, it is quite likely that the result turned to be insignificant.

To summarize, using an incomplete data record may result in potential distortion to the true value and/or the point estimate of the extreme value index. Together with the low k used in estimation, one may not reject the null hypothesis that the extreme value index is zero, which can be false for the original dataset. On top of that, having a zero extreme value index does not necessarily imply having an infinite endpoint. Therefore, it is still less evident to conclude that human life is unlimited.

How to make inference on the tail behavior of the full dataset when the observed dataset is subject to random truncation is still open for research. From the three examples, it is clear that if the random threshold has a lower endpoint than that of the original data, such an inference is not feasible because the tail of the original dataset is not observed at all. If the random threshold has a higher endpoint than that of the original data, the distortion is limited to a high bias of the estimator. Bias correction might be a useful tool here for improving statistical inference. The most complicated case is when the endpoint of the random threshold coincides with that of the original dataset. This is left for future research.

References

de Haan, L., Stadtmüller, U.: Generalized regular variation of second order. J. Aust. Math. Soc. **61**(3), 381–395 (1996)

Drees, H., Ferreira, A., de Haan, L.: On maximum likelihood estimation of the extreme value index. Ann. Appl. Probab. **14**(3), 1179–1201 (2004)

Einmahl, J.H., Fils-Villetard, A., Guillou, A.: Statistics of extremes under random censoring. Bernoulli **14**(1), 207–227 (2008)

Gnedenko, B.: Sur la distribution limite du terme maximum d'une serie aleatoire. Ann. Math. **44**(3), 423–453 (1943)

Rootzén, H., Zholud, D.: Human life is unlimited–but short. Extremes **20**(4), 713–728 (2017)