Survival Study of Extreme Record

Mohd Bakri Adam

Laboratory of Applied and Computational Statistics
Institute for Mathematical Research
Universiti Putra Malaysia

Mathematics Department
Faculty of Science
Universiti Putra Malaysia

bakri@putra.upm.edu.my

Abstract

This paper describes the survival study of extreme record of athletics performance. Two area of statistics are used to model and check the best model for the athletics data. We make use of the extreme value theory for minima and utilized the facility provided by the Kaplan-Meier to develop new goodness-of-fit test method via graphical approaches.

Introduction

Let $x_1, x_2, \ldots$ be a sequence of independent but not identically distributed random variables from a Generalized Extreme Value for Minima, GEVM. If we let w_i be the record value for the smallest data at time i of $\{X_i\}$, then the observations of w_i is a record of current minimum in the sequence within some specified interval of time which will change along the time. The initial observation of w_i should always be a record. The record is an extreme of extreme data, as data truncated to the left or right to form the smallest or the biggest record.

We do the analysis of the smallest record, estimating the parameters of the record involving the censored data using maximum likelihood estimate. We introduce a new method of goodness-of-fit to suit the record data by removing the existing trend from the model. The Kaplan-Meier estimation is incorporated together with the survival function to get the appropriate P-P and Q-Q plots.

Methodology

Let $z_1, \ldots, z_n$ be the sequence of annual minima from $\{x_i\}$ with

$$z_i \sim GEVM\ (\mu, \sigma, \xi)$$

where μ, σ and ξ are the location, scale and shape parameters respectively. The GEVM is defined as

$$1-\exp\left\{-\left[1-\xi\left(z-\mu\right)/\sigma\right]_+^{-1/\xi}\right\},$$

with $\mu \in (-\infty,\infty)$, $\sigma \in (0,\infty)$, $\xi \in (-\infty,\infty)$

$y_+ = \max(0, y)$.

Then, if we let $Y_i = \min(z_1, \ldots z_i)$ for $i = 1,2,\ldots, n$, and y_i is the record smallest value at time i amongst the annual minima data. It is easily seen that the distinct $\{y_i\}$ values are subset of the distinct w_i where $w_i = \min(x_1, .., x_n)$. Bunge & Goldie (1999) provide a detailed discussion about how the record value and record time relate to the extreme value theory.

We consider that an event has been censored at time t until other record breaking occurs. If Y is a continuous random variable then $S(y) = P(Y \geq y)$ is the survivor function of y. The survivor function of Y for records of annual minima is defined as

$$S(y)=\exp\left\{-\left[1-\xi(z-\mu)/\sigma\right]_+^{-1/\xi}\right\}.$$

The joint probability function of $Y_1,\ldots,Y_n$, is redefined as follows

$$f(y_1,\ldots,y_n)=\prod_{i=1}^{n}\left(\frac{1}{\sigma}\left[1-\xi\left(\frac{y_i-\mu_i}{\sigma}\right)\right]_+^{-1-\frac{1}{\xi}}\right)^{I_i}$$

$$\times\left(\exp\left\{-\left[1-\xi\left(\frac{y_i-\mu_i}{\sigma}\right)\right]_+^{-1/\xi}\right\}\right)^{I_i}$$

$$\times\prod_{i=1}^{n}\left(\exp\left\{-\left[1-\xi\left(\frac{y_i-\mu_i}{\sigma}\right)\right]_+^{-1/\xi}\right\}\right)^{1-I_i},$$

with $I_i = 1$ if at time I there is a record and $I_i = 0$ if there is not a record. The log-likelihood function for the record sequence $y=(y_1,..,y_n)$ is

$$L(y;\mu,\sigma,\xi)=A+B+C+D$$

where

$A=-\sum_{i=1}^{n} I_i\log\sigma$,
$B=-\sum_{i=1}^{n} I_i(1-1/\xi)\log E$,
$C=-\sum_{i=1}^{n} I_i E^{-1/\xi}$,
$D=-\sum_{i=1}^{n}(1-I_i)E^{-1/\xi}$ and

$E=1-\xi(y_i-\mu)/\sigma$ with $I_i=1$ if at time i there is a record and $I_i = 0$ if there is no record.

We clearly see that the sequence of $\{Y_i\}$ is neither independent nor identically distributed. As we know, $z_t \sim EVM(\mu,\sigma,\xi)$ then

$$Z_i=\mu_i+E_i,$$

where μ is the trend in Z_i and $E_i \sim GEVM(0,\sigma,\xi)$. Then the sequence of $\{E_i\}$ for $i=1,\ldots,n$ are independent and identically distributed by separating the trend from the data.

The E_i is used to construct P-P and Q-Q plots for assessing the goodness-of-fit of the record progression. See Adam (2007) pages 37-38 for more details.

Application

We used the women's athletics 1500 m and 3000m events from recognized international events over the period of 1972-1992. We focused on the record data only and assumed that the trend is

$$\mu=\alpha-\beta\left[1-\exp(-\gamma t)\right]$$

where $\beta>0,\gamma>0$ and t is the year, taking 1971 as the base year. This trend is based on the exponential decay of the annual minima over time. This model was used by Robinson and Tawn (1995).

The maximum likelihood estimates are as tabulated in Table 1 as follows,

Table 1: The parameter estimation for record for the women's athletics 1500 m and 3000 m track event from 1972 to 1992 and the numbers in parentheses are the standard errors.

	1500 m	3000 m
α	248(29)	548(9)
β	8.9(31.1)	37.7(13.9)
γ	0.261(0.609)	0.212(0.212)
σ	4.79(14.03)	4.12(7.23)
ξ	-0.567(2.546)	-0.020(1.110)

The 95% confidence interval for the shape parameters for 1500 m and 3000 m events are (-5.557,4.423) and (-2.196,2.156). These are unacceptably large confidence intervals. When maximizing the log-likelihood with fixed α,β and γ values at their estimated values (using all annual minima data), the standard errors for the shape are lowered, because separating the trend from the record process is hard if only the records are observed.

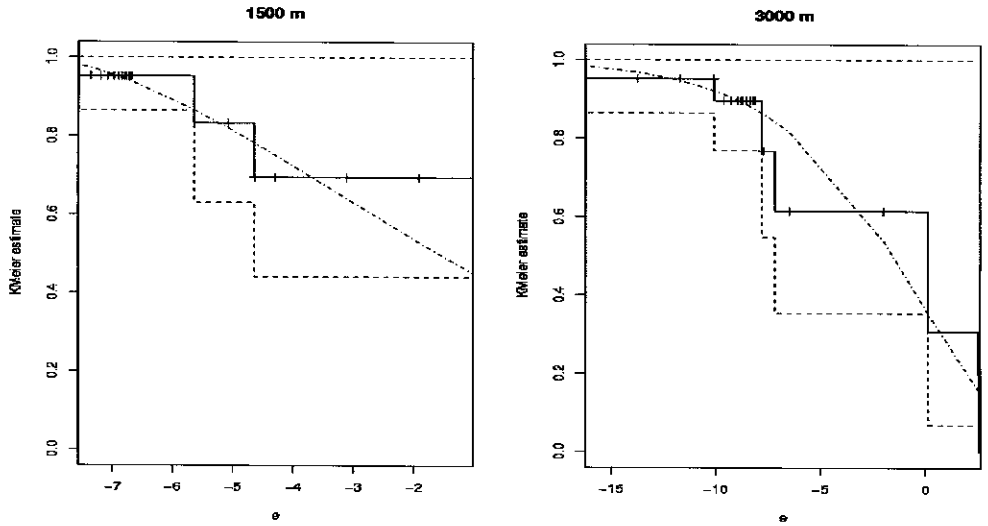

Figure 1: Kaplan-Meier estimates for women's athletics 1500 m and 3000 m of e_i with dash dotted line is survival function.

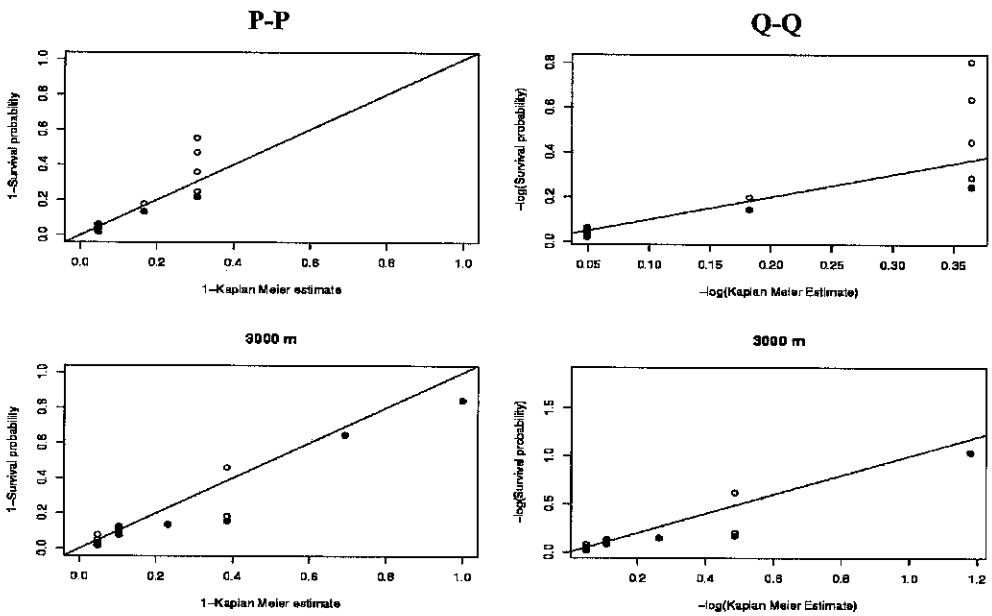

Figure 2: The P-P and Q-Q plots for women's athletics 1500 m and 3000 m, with record points marked with black circle.

In order to fit the model, as the record data involve censored data, we replace the empirical density function with a Kaplan-Meier estimate in P-P and Q-Q plots, see Figure 1.

The P-P and Q-Q plots in Figure 2 show that the model's fit is acceptable for women's athletics 1500 m and 3000 m events as there are signs of linear pattern in the plots for censored data and the record. We noted an important point, i.e. the number of a record depend wholly on the form of the trend of the location parameter.

References

[1] Adam. 2007. Extreme value modelling of sports data. PhD thesis, Lancaster University, Lancaster, UK.

[2] Robinson, M.E. and Tawn, J.A. 1995. Statistics for exceptional athletics. *J. Applied Statistics* **44**(4), 499-511.

[3] Bunge, J. and Goldie, C.M. 1999. Record sequences and their applications. (unpublish)