Prediction of record values by using quantile regression curves and distortion functions

Jorge Navarro¹

Received: 5 May 2021 / Accepted: 5 October 2021 / Published online: 9 November 2021
© The Author(s) 2021

Abstract

The purpose of the paper is to provide a general method based on conditional quantile curves to predict record values from preceding records. The predictions are based on conditional median (or median regression) curves. Moreover, conditional quantiles curves are used to provide confidence bands for these predictions. The method is based on the recently introduced concept of multivariate distorted distributions that are used instead of copulas to represent the dependence structure. This concept allows us to compute the conditional quantile curves in a simple way. The theoretical findings are illustrated with a non-parametric model (standard uniform), two parametric models (exponential and Pareto), and a non-parametric procedure for the general case. A real data set and a simulated case study in reliability are analysed.

Keywords Record values · Distorted distributions · Median regression · Confidence regions · Copula

Mathematics Subject Classification 62F99 · 62M20

1 Introduction

The prediction of future record values is a relevant topic nowadays. Several papers had studied and applied this topic to different problems (climate change, sports, etc.). In several practical situations, the usual point predictions obtained with linear regression do not provide good record predictions. The basic tools in this field can be seen in Dunsmore (1983), Raqab and Nagaraja (1995), Awad and Raqab (2000) and Nevzorov (2001) and the recent advances in Paul and Thomas (2016), Guo et al. (2020), Volovskiy and Kamps (2020a, b). Specific results for records from exponential distributions can be found in Awad and Raqab (2000), Basak and Balakrishnan (2003),

✉ Jorge Navarro
jorgenav@um.es

¹ Facultad de Matemáticas, Universidad de Murcia, 30100 Murcia, Spain

Raqab (2007), Volovskiy and Kamps (2020a, b) and the references therein. Analogously, the prediction of record values from a Pareto model were studied in Raqab et al. (2007), Volovskiy (2018) and Volovskiy and Kamps (2020b). Moreover, the prediction tools for generalized order statistics can also be applied to record values (see, Burkschat 2009).

The conditional median (or quantile) regression curves are a good alternative to the classical (mean) regression tool. The theoretical quantile regression curve to predict a response variable Y from an explanatory variable X can be obtained from the copula of (X, Y) (see, e.g., p. 217 in, Nelsen 2006). In practice, we can estimate the unknown parameters (in the copula or in the marginal distributions) or to use the empirical linear or non-linear estimators proposed in Koenker and Bassett (1978) and Koenker (2005). The package `quantreg` for the statistical program R can help us in this purpose. Moreover, the quantile regression curves can also be used to obtain confidence bands for the predictions and bivariate box plots, see Koenker and Bassett (1978) and Navarro (2020).

The aim of the paper is to use conditional median (or quantile) regression curves to predict future record values. Instead of use the general tools, we develop a specific procedure for upper record values. This procedure is based on the concept of multivariate distorted distributions introduced recently in Navarro et al. (2020). It is an alternative method to the classic copula approach to represent the dependence structure. A similar method can be used for lower record values (see Sect. 5). This is a novel approach that simplifies the expressions for the quantile curves.

The rest of the paper is scheduled as follows. The main results are given in Sect. 2. These general results are applied to three models in Sect. 3. A fixed model. We consider a uniform distribution but a similar method holds for other fixed distribution functions. A parametric model, the popular proportional hazard rate (PHR) model with applications to exponential and Pareto models. And finally, we also consider a non-parametric approach. A (classical) real data set and a simulated data set of lifetimes in a specific sampling procedure in reliability are studied in Sects. 4 and 5, respectively. The conclusions are placed in Sect. 6.

If f is a real valued function with n variables, then $\partial_i f$ denotes the partial derivative of f with respect to its i -th variable. Analogously, $\partial_{i,j} f = \partial_i \partial_j f$ and so on. Whenever we use these partial derivatives, we are tacitly assuming that they exist.

2 Main results

Let $(X_n)_n$ be an infinite sequence of independent and identically distributed (IID) random variables with an absolutely continuous cumulative distribution function F , survival (reliability) function $\bar{F} = 1 - F$ and probability density function (PDF) $f = F'$.

An observation X_j is called an *upper record value* in this sequence if it is greater than all previously observed values $X_1, \dots, X_{j-1}$. More specifically, the *upper record times* are defined as $T(1) = 1$ and

$$T(n+1) = \min\{j : X_j > X_{T(n)}\}$$

for $n = 1, 2, \dots$. Then the sequence $(R_n)_n$ of *upper record values* is defined by $R_n = X_{T(n)}$ for $n = 1, 2, \dots$. The *lower record values* $(R_n^*)_n$ are defined in a similar way. The purpose of the paper is to predict R_s from $R_1, \dots, R_n$ (or just from a single R_i with $i \leq n$) for $s > n$.

It is well known (see, e.g., Nevzorov 2001, p. 65) that the survival function $\bar{G}_n(t) = \Pr(R_n > t)$ of the n -th upper record value is given by

$$\bar{G}_n(t) = \bar{F}(t) \sum_{k=0}^{n-1} \frac{(-\log(\bar{F}(t)))^k}{k!} = \bar{q}_n(\bar{F}(t)), \quad (1)$$

where

$$\bar{q}_n(u) = u \sum_{k=0}^{n-1} \frac{(-\log(u))^k}{k!}$$

for $u \in [0, 1]$ and $n = 1, 2, \dots$. The function $\bar{q}_n : [0, 1] \rightarrow [0, 1]$ is continuous, increasing and satisfies $\bar{q}_n(0) = 0$ and $\bar{q}_n(1) = 1$. It is called *dual distortion function*, see Hürlimann (2004). A similar representation holds for the distribution function G_n of R_n that can be written as

$$G_n(t) = q_n(F(t)),$$

where $q_n(u) = 1 - \bar{q}_n(1 - u)$ for $u \in [0, 1]$. The function q_n is called the *distortion function* and G_n is a *distorted distribution* from F . The distorted distributions were introduced in Yaari (1987) in the context of the theory of choice under risk. Some applications and ordering results can be seen in Wang (1996) and Navarro et al. (2013).

Then, the PDF g_n of R_n can be obtained as

$$g_n(t) = f(t) \bar{q}'_n(\bar{F}(t)) = \frac{1}{(n-1)!} f(t) (-\log(\bar{F}(t)))^{n-1} \quad (2)$$

(see, e.g., Awad and Raqab 2000). Even more, the joint PDF $\mathbf{g}$ of $(R_1, \dots, R_n)$ is given by

$$\mathbf{g}(x_1, \dots, x_n) = h(x_1) \dots h(x_n) \bar{F}(x_n) \quad (3)$$

for $x_1 < \dots < x_n$ (see, Nevzorov 2001, p. 68), where $h = f/\bar{F}$ is the hazard (or failure) rate function associated to F . Hence, from (3), we obtain the following result.

Proposition 1 *The joint survival function $\bar{\mathbf{G}}$ of $(R_1, \dots, R_n)$ can be written as*

$$\bar{\mathbf{G}}(x_1, \dots, x_n) = \hat{D}(\bar{F}(x_1), \dots, \bar{F}(x_n)) \quad (4)$$

for a continuous increasing function $\hat{D} : [0, 1]^n \rightarrow [0, 1]$ which does not depend on F and satisfies $\hat{D}(0, \dots, 0) = 0$ and $\hat{D}(1, \dots, 1) = 1$.

Proof As F is continuous, $\bar{G}_1, \dots, \bar{G}_n$ are continuous and, from Sklar's theorem (see, e.g., Nelsen 2006, p. 46), the joint survival function $\bar{\mathbf{G}}$ of $(R_1, \dots, R_n)$ can be written as

$$\bar{\mathbf{G}}(x_1, \dots, x_n) = \hat{C}(\bar{G}_1(x_1), \dots, \bar{G}_n(x_n))$$

for all $x_1, \dots, x_n$ and a unique (survival) copula function $\hat{C}$. Hence, from (1), (4) holds for the increasing and continuous function

$$\hat{D}(u_1, \dots, u_n) = \hat{C}(\bar{q}_1(u_1), \dots, \bar{q}_n(u_n))$$

defined for $u_1, \dots, u_n \in [0, 1]$ that satisfies $\hat{D}(0, \dots, 0) = 0$ and $\hat{D}(1, \dots, 1) = 1$. $\square$

A similar representation can be obtained for the joint distribution function $\mathbf{G}$, that is,

$$\mathbf{G}(x_1, \dots, x_n) = D(F(x_1), \dots, F(x_n))$$

for a continuous increasing function $D : [0, 1]^n \rightarrow [0, 1]$ which does not depend on F . Note that D can be computed from $\hat{D}$ (and vice versa). We must say that the explicit expression of $\hat{D}$ (or D) is complicated. However, the expression for $\hat{C}$ is much more complicated (some examples are given below).

These representations are similar to the classic copula representations (see, e.g., Nelsen 2006) but here D and $\hat{D}$ are not necessarily copulas and F is not the marginal distribution of R_i for $i = 2, \dots, n$. Note that, as in (1), the joint survival $\bar{\mathbf{G}}$ is written in (4) as a distortion of the univariate survival function $\bar{F}$ at different points.

This kind of representations were called *multivariate distorted distributions* in Navarro et al. (2020). It can be proved that D and $\hat{D}$ can be extended to $\mathbb{R}^n$ and then they are multivariate distribution functions with supports included in $[0, 1]^n$. Analogously, the joint PDF of $(R_1, \dots, R_n)$ can be written as follows.

Proposition 2 *The joint PDF $\mathbf{g}$ of $(R_1, \dots, R_n)$ can be written as*

$$\mathbf{g}(x_1, \dots, x_n) = f(x_1) \dots f(x_n) \hat{d}(\bar{F}(x_1), \dots, \bar{F}(x_n)), \quad (5)$$

where $\hat{d} = \partial_{1, \dots, n} \hat{D}$ is the PDF of $\hat{D}$ given by

$$\hat{d}(u_1, \dots, u_n) = \frac{1}{u_1 \dots u_{n-1}} \quad (6)$$

for $1 > u_1 > \dots > u_n > 0$ (zero elsewhere).

Proof The expression of the PDF $\mathbf{g}$ in (5) can be obtained by differentiating in (4). Then the explicit expression for $\hat{d}$ can be obtained from (3) and (5). $\square$

Therefore, the explicit expression for $\hat{D}$ can be obtained from the expression for $\hat{d}$ given in (6). Moreover, the different marginal distributions of $(R_1, \dots, R_n)$ have also multivariate distorted distributions (see, Navarro et al. 2020). For example, if $1 \leq i < j \leq n$, then the joint survival function $\bar{G}_{i,j}$ of (R_i, R_j) can be written from (4) as

$$\bar{G}_{i,j}(x_i, x_j) = \hat{D}_{i,j}(\bar{F}(x_i), \bar{F}(x_j)), \quad (7)$$

where $\hat{D}_{i,j}(u, v) = \hat{D}(1, \dots, 1, u, 1, \dots, 1, v, 1, \dots, 1)$ and u and v are placed at the i -th and j -th variables, respectively.

Hence, to predict R_j from R_i for $i < j$, we can use here a similar technique to that developed in copula theory to compute the median regression curve. The result can be stated as follows.

Proposition 3 *The conditional survival function $\bar{G}_{j|i}$ of $(R_j | R_i = x_i)$ for $1 \leq i < j \leq n$ is given by*

$$\bar{G}_{j|i}(x_j | x_i) = \frac{(i-1)!}{(-\log \bar{F}(x_i))^{i-1}} \partial_1 \hat{D}_{i,j}(\bar{F}(x_i), \bar{F}(x_j)) \quad (8)$$

for $x_j \geq x_i$ whenever $f(x_i) > 0$, $0 < \bar{F}(x_i) < 1$ and $\lim_{v \rightarrow 0^+} \partial_1 \hat{D}_{i,j}(u, v) = 0$ for all $0 < u < 1$.

Proof From (7), the joint PDF of (R_i, R_j) is

$$\mathbf{g}_{i,j}(x_i, x_j) = \partial_{i,j} \bar{G}_{i,j}(x_i, x_j) = f(x_i) f(x_j) \partial_{1,2} \hat{D}_{i,j}(\bar{F}(x_i), \bar{F}(x_j)).$$

Therefore, from (2), the conditional PDF of $(R_j | R_i = x_i)$ is

$$\mathbf{g}_{j|i}(x_j | x_i) = \frac{\mathbf{g}_{i,j}(x_i, x_j)}{\mathbf{g}_i(x_i)} = \frac{(i-1)!}{(-\log \bar{F}(x_i))^{i-1}} f(x_j) \partial_{1,2} \hat{D}_{i,j}(\bar{F}(x_i), \bar{F}(x_j)),$$

for $x_j > x_i$ and x_i such that $f(x_i) > 0$. Hence, its survival function is

$$\begin{aligned} \bar{G}_{j|i}(x_j | x_i) &= \frac{(i-1)!}{(-\log \bar{F}(x_i))^{i-1}} \int_{x_j}^{\infty} f(z) \partial_{1,2} \hat{D}_{i,j}(\bar{F}(x_i), \bar{F}(z)) dz \\ &= \frac{(i-1)!}{(-\log \bar{F}(x_i))^{i-1}} \left[-\partial_1 \hat{D}_{i,j}(\bar{F}(x_i), \bar{F}(z)) \right]_{z=x_j}^{\infty} \\ &= \frac{(i-1)!}{(-\log \bar{F}(x_i))^{i-1}} \partial_1 \hat{D}_{i,j}(\bar{F}(x_i), \bar{F}(x_j)) \end{aligned}$$

where, in the last equality, we use that $\lim_{v \rightarrow 0^+} \partial_1 \hat{D}_{i,j}(u, v) = 0$ for all $0 < u < 1$ and that $0 < \bar{F}(x_i) < 1$. $\square$

As a consequence, the *median regression curve* to predict R_j from R_i is

$$m_{j|i}(x_i) = \bar{\mathbf{G}}_{j|i}^{-1}(0.5|x_i), \quad (9)$$

where $\bar{\mathbf{G}}_{j|i}^{-1}$ is the inverse function of $\bar{\mathbf{G}}_{j|i}$. This procedure can also be used to determine quantile confidence bands for this prediction. For example the centered 50% quantile confidence band is given by

$$\left[\bar{\mathbf{G}}_{j|i}^{-1}(0.75|x_i), \bar{\mathbf{G}}_{j|i}^{-1}(0.25|x_i) \right]$$

and the centered 90% quantile confidence band by

$$\left[\bar{\mathbf{G}}_{j|i}^{-1}(0.95|x_i), \bar{\mathbf{G}}_{j|i}^{-1}(0.05|x_i) \right].$$

To get these predictions from (8) we need to compute $\hat{D}$ (or $\hat{D}_{i,j}$) and to estimate $\bar{F}$. In the following subsections, we consider different relevant cases.

2.1 Case $i = 1$ and $j = 2$

A straightforward calculation from (3) proves that the joint survival function $\bar{\mathbf{G}}_{1,2}$ of (R_1, R_2) can be written as

$$\bar{\mathbf{G}}_{1,2}(x_1, x_2) = \bar{F}(x_2) + \bar{F}(x_2) \log \frac{\bar{F}(x_1)}{\bar{F}(x_1)} = \hat{D}_{1,2}(\bar{F}(x_1), \bar{F}(x_2)) \quad (10)$$

for $x_1 \leq x_2$, where

$$\hat{D}_{1,2}(u, v) = v + v \log \frac{u}{v}$$

for $1 > u \geq v > 0$. Note that $\hat{D}_{1,2}$ is not a copula since $\hat{D}_{1,2}(1, v) = v - v \log v \neq v$. Actually $\hat{D}_{1,2}(1, v) = \bar{q}_2(v)$, that is, it is the dual distortion function of the second upper record given in (1). Also note that $\bar{F}$ is equal to the first marginal survival distribution but it is not equal to the second one (so (10) is not a copula representation). To get the copula (or the survival copula) representation of (R_1, R_2) we need the explicit expression of the inverse function of G_2 .

The distortion function $\hat{D}_{1,2}$ can also be obtained from Proposition 2 as

$$\hat{D}_{1,2}(u, v) = \int_0^v \int_y^u \frac{1}{x} dx dy = \int_0^v (\log u - \log y) dy = v + v \log \frac{u}{v}$$

for $1 > u \geq v > 0$. Hence $\partial_1 \hat{D}_{1,2}(u, v) = v/u$ for $1 > u \geq v > 0$, and, from (8), we get

$$\bar{\mathbf{G}}_{2|1}(x_2|x_1) = \frac{\partial_1 \hat{D}_{1,2}(\bar{F}(x_1), \bar{F}(x_2))}{(-\log \bar{F}(x_1))^0} = \frac{\bar{F}(x_2)}{\bar{F}(x_1)}$$

for $x_2 \geq x_1$ and x_1 such that $0 < \bar{F}(x_1) < 1$ and $f(x_1) > 0$. This is a very well known result (see, e.g., Nevzorov 2001, p. 68). Then, the median regression curve to predict R_2 from R_1 is

$$m_{2|1}(x_1) = \bar{\mathbf{G}}_{2|1}^{-1}(0.5|x_1) = \bar{F}^{-1}(0.5\bar{F}(x_1)), \quad (11)$$

where $\bar{F}^{-1}$ is the inverse function of $\bar{F}$. Analogously, the centered 50% and 90% quantile confidence bands for this prediction are given by

$$\left[\bar{\mathbf{G}}_{2|1}^{-1}(0.75|x_1), \bar{\mathbf{G}}_{2|1}^{-1}(0.25|x_1) \right] = \left[\bar{F}^{-1}(0.75\bar{F}(x_1)), \bar{F}^{-1}(0.25\bar{F}(x_1)) \right]$$

and

$$\left[\bar{\mathbf{G}}_{2|1}^{-1}(0.95|x_1), \bar{\mathbf{G}}_{2|1}^{-1}(0.05|x_1) \right] = \left[\bar{F}^{-1}(0.95\bar{F}(x_1)), \bar{F}^{-1}(0.05\bar{F}(x_1)) \right].$$

They are plotted in Sect. 3 for different survival functions $\bar{F}$.

2.2 Case $i = n$ and $j = n + 1$

First, let us see how the joint survival function $\bar{\mathbf{G}}_{n,n+1}$ of (R_n, R_{n+1}) can be written as a distortion from the univariate survival function $\bar{F}$.

Proposition 4 *The joint survival function $\bar{\mathbf{G}}_{n,n+1}$ of (R_n, R_{n+1}) can be written as*

$$\bar{\mathbf{G}}_{n,n+1}(x_n, x_{n+1}) = \hat{D}_{n,n+1}(\bar{F}(x_n), \bar{F}(x_{n+1}))$$

for $x_n \leq x_{n+1}$, where

$$\hat{D}_{n,n+1}(u, v) = -\frac{1}{n!}v(-\log u)^n + \bar{\gamma}_{n+1}(-\log v) \quad (12)$$

for $1 > u \geq v > 0$ and

$$\bar{\gamma}_{n+1}(z) = \frac{1}{n!} \int_z^\infty x^n e^{-x} dx \quad (13)$$

is the survival function of a gamma distribution with scale parameter equal to one and shape parameter equal to $n + 1$.

Proof From Proposition 2 and (7), we know that $\hat{D}_{n,n+1}$ is a marginal distribution of the PDF given by

$$\hat{d}(u_1, \dots, u_{n+1}) = \frac{1}{u_1 \dots u_n}$$

for $1 > u_1 > \dots > u_{n+1} > 0$ and zero elsewhere. Therefore, for $1 > u > v > 0$, we have

$$\begin{aligned} \hat{D}_{n,n+1}(u, v) &= \int_0^v \int_{u_{n+1}}^u \int_{u_n}^1 \dots \int_{u_2}^1 \frac{1}{u_1 \dots u_n} du_1 \dots du_{n+1} \\ &= \int_0^v \int_{u_{n+1}}^u \int_{u_n}^1 \dots \int_{u_3}^1 (-\log u_2) \frac{1}{u_2 \dots u_n} du_2 \dots du_{n+1} \\ &= \int_0^v \int_{u_{n+1}}^u \int_{u_n}^1 \dots \int_{u_4}^1 \frac{1}{2} (-\log u_3)^2 \frac{1}{u_3 \dots u_n} du_3 \dots du_{n+1} \\ &= \dots \end{aligned}$$

and so on. Thus we finally arrive to

$$\begin{aligned} \hat{D}_{n,n+1}(u, v) &= \int_0^v \int_{u_{n+1}}^u \frac{1}{(n-1)!} (-\log u_n)^{n-1} \frac{1}{u_n} du_n du_{n+1} \\ &= \frac{1}{n!} \int_0^v [(-\log u_{n+1})^n - (-\log u)^n] du_{n+1} \\ &= -\frac{1}{n!} v (-\log u)^n + \frac{1}{n!} \int_0^v (-\log u_{n+1})^n du_{n+1} \end{aligned}$$

and, by doing the change $z = -\log u_{n+1}$, we obtain expression (12). $\square$

Therefore, from (12), we get

$$\partial_1 \hat{D}_{n,n+1}(u, v) = \frac{1}{(n-1)!} (-\log u)^{n-1} \frac{v}{u}$$

for $1 > u \geq v > 0$ and from (8) the conditional survival function is

$$\bar{\mathbf{G}}_{n+1|n}(x_{n+1}|x_n) = (n-1)! \frac{\partial_1 \hat{D}_{n,n+1}(\bar{F}(x_n), \bar{F}(x_{n+1}))}{(-\log \bar{F}(x_n))^{n-1}} = \frac{\bar{F}(x_{n+1})}{\bar{F}(x_n)}$$

for $x_{n+1} \geq x_n$ and x_n such that $\bar{F}(x_n) > 0$ and $f(x_n) > 0$. This expression is also a very well known result (see, e.g., Nevzorov 2001, p. 68). Even more, we can see that the record values form a Markov chain, that is,

$$\begin{aligned} \Pr(R_{n+1} > x_{n+1} | R_1 = x_1, \dots, R_n = x_n) &= \Pr(R_{n+1} > x_{n+1} | R_n = x_n) \\ &= \bar{\mathbf{G}}_{n+1|n}(x_{n+1}|x_n). \end{aligned}$$

Therefore the median regression curve to predict R_{n+1} from R_n (or from $R_1, \dots, R_n$) is the same as that obtained in the preceding subsection (see (11)) and so are the confidence regions. This is a good point since, if F is known, then the median regression curve and the confidence regions do not depend on n and the sequence of paired records (R_n, R_{n+1}) for $n = 1, 2, \dots$ can be plotted in the same graph (since they have the same conditional distribution). This is not the case if we need to estimate F or a parameter in F (see the different examples in Sect. 3).

2.3 Case $i = m$ and $j = n$

In the general case the joint survival function $\bar{\mathbf{G}}_{m,n}$ of (R_m, R_n) for $1 \leq m < n$ can also be written as a distortion from $\bar{F}$. The result is stated in the following proposition.

Proposition 5 *The joint survival function $\bar{\mathbf{G}}_{m,n}$ of (R_m, R_n) for $1 \leq m < n$ can be written as*

$$\bar{\mathbf{G}}_{m,n}(x_m, x_n) = \hat{D}_{m,n}(\bar{F}(x_m), \bar{F}(x_n))$$

for $x_n \leq x_{n+1}$, where

$$\hat{D}_{m,n}(u, v) = \bar{\gamma}_m(-\log v) + \frac{1}{(m-1)!} \int_{-\log u}^{-\log v} z^{m-1} e^{-z} \bar{\gamma}_{n-m}(-\log v - z) dz \quad (14)$$

for $1 > u \geq v > 0$ and $\bar{\gamma}_k$ is the survival function in (13).

Proof From (3) (or from (3.2) in, Awad and Raqab 2000) the joint PDF of (R_m, R_n) is

$$\mathbf{g}_{m,n}(x_m, x_n) = \frac{h(x_m)f(x_n)}{(m-1)!(n-m-1)!} H^{m-1}(x_m) [H(x_n) - H(x_m)]^{n-m-1}$$

for $x_m < x_n$ (zero elsewhere), where $H(x) = -\log \bar{F}(x)$ and $h(x) = H'(x)$ (the hazard rate function). Hence its survival function is

$$\begin{aligned} \bar{\mathbf{G}}_{m,n}(x_m, x_n) &= \int_{x_m}^{\infty} \int_{x_n}^{\infty} \mathbf{g}_{m,n}(y_m, y_n) dy_n dy_m \\ &= \int_{x_m}^{x_n} \int_{x_n}^{\infty} \mathbf{g}_{m,n}(y_m, y_n) dy_n dy_m + \int_{x_n}^{\infty} \int_{y_m}^{\infty} \mathbf{g}_{m,n}(y_m, y_n) dy_n dy_m \\ &= \int_{x_m}^{x_n} \int_{x_n}^{\infty} \frac{h(y_m)f(y_n)H^{m-1}(y_m)}{(m-1)!(n-m-1)!} [H(y_n) - H(y_m)]^{n-m-1} dy_n dy_m \\ &\quad + \int_{x_n}^{\infty} \int_{y_m}^{\infty} \frac{h(y_m)f(y_n)H^{m-1}(y_m)}{(m-1)!(n-m-1)!} [H(y_n) - H(y_m)]^{n-m-1} dy_n dy_m. \end{aligned}$$

Now we do the change $x = H(y_n) - H(y_m)$, with $dx = h(y_n)dy_n$ and $\bar{F}(y_m)e^{-x} = \bar{F}(y_n)$. Hence

$$\begin{aligned}\bar{\mathbf{G}}_{m,n}(x_m, x_n) &= \int_{x_m}^{x_n} \int_{H(x_n)-H(y_m)}^{\infty} \frac{h(y_m)\bar{F}(y_m)x^{n-m-1}e^{-x}}{(m-1)!(n-m-1)!} H^{m-1}(y_m) dx dy_m \\ &\quad + \int_{x_n}^{\infty} \int_0^{\infty} \frac{h(y_m)\bar{F}(y_m)x^{n-m-1}e^{-x}}{(m-1)!(n-m-1)!} H^{m-1}(y_m) dx dy_m \\ &= \int_{x_m}^{x_n} \frac{h(y_m)\bar{F}(y_m)H^{m-1}(y_m)}{(m-1)!} \bar{\gamma}_{n-m}(H(x_n) - H(y_m)) dy_m \\ &\quad + \int_{x_n}^{\infty} \frac{h(y_m)\bar{F}(y_m)H^{m-1}(y_m)}{(m-1)!} dy_m\end{aligned}$$

and by doing the change $z = H(y_m)$, with $dz = h(y_m)dy_m$ and $e^{-z} = \bar{F}(y_m)$, we get

$$\begin{aligned}\bar{\mathbf{G}}_{m,n}(x_m, x_n) &= \int_{H(x_m)}^{H(x_n)} \frac{z^{m-1}e^{-z}}{(m-1)!} \bar{\gamma}_{n-m}(H(x_n) - z) dz + \int_{H(x_n)}^{\infty} \frac{z^{m-1}e^{-z}}{(m-1)!} dz \\ &= \bar{\gamma}_m(H(x_n)) + \frac{1}{(m-1)!} \int_{H(x_m)}^{H(x_n)} z^{m-1}e^{-z} \bar{\gamma}_{n-m}(H(x_n) - z) dz \\ &= \hat{D}_{m,n}(\bar{F}(x_m), \bar{F}(x_n))\end{aligned}$$

for $x_m \leq x_n$ as stated above. $\square$

Therefore, from (14), we obtain

$$\begin{aligned}\partial_1 \hat{D}_{m,n}(u, v) &= \frac{1}{u} \frac{(-\log u)^{m-1}}{(m-1)!} e^{\log(u)} \bar{\gamma}_{n-m}(-\log v + \log u) \\ &= \frac{1}{(m-1)!} (-\log u)^{m-1} \bar{\gamma}_{n-m} \left(-\log \frac{v}{u} \right)\end{aligned}$$

for $1 > u > v > 0$ and, from (8), the conditional survival function of $(R_n | R_m = x_m)$ is

$$\bar{\mathbf{G}}_{n|m}(x_n | x_m) = (m-1)! \frac{\partial_1 \hat{D}_{m,n}(\bar{F}(x_m), \bar{F}(x_n))}{(-\log \bar{F}(x_m))^{m-1}} = \bar{\gamma}_{n-m} \left(-\log \frac{\bar{F}(x_n)}{\bar{F}(x_m)} \right) \quad (15)$$

for $x_n \geq x_m$ since $\lim_{v \rightarrow 0^+} \partial_1 \hat{D}_{m,n}(u, v) = 0$ for all $0 < u < 1$. Note that (15) proves that $\bar{\mathbf{G}}_{n|m}$ is also a distorted distribution since

$$\bar{\mathbf{G}}_{n|m}(x_n | x_m) = \bar{q}_{n|m}(\bar{F}(x_n) | \bar{F}(x_m))$$

with $\bar{q}_{n|m}(v | \bar{F}(x_m)) = \bar{\gamma}_{n-m}(-\log v + \log \bar{F}(x_m))$ for $0 \leq v \leq \bar{F}(x_m)$.

Therefore, the median regression curve to predict R_n from $R_m = x_m$ is given by

$$m_{n|m}(x_m) = \bar{\mathbf{G}}_{n|m}^{-1}(0.5|x_m) = \bar{F}^{-1}(c_{n-m}(0.5)\bar{F}(x_m)), \quad (16)$$

where $\bar{F}^{-1}$ is the inverse function of $\bar{F}$,

$$c_k(y) = \exp(-\gamma_k^{-1}(y)) \quad (17)$$

and γ_k^{-1} is the quantile function of a gamma distribution with shape parameter k and scale parameter equal to one.

Analogously, the 50% and 90% quantile confidence bands for this prediction are given by

$$\left[\bar{F}^{-1}(c_{n-m}(0.25)\bar{F}(x_m)), \bar{F}^{-1}(c_{n-m}(0.75)\bar{F}(x_m)) \right]$$

and

$$\left[\bar{F}^{-1}(c_{n-m}(0.05)\bar{F}(x_m)), \bar{F}^{-1}(c_{n-m}(0.95)\bar{F}(x_m)) \right].$$

Note that they only depend on $\bar{F}$ and on $k = n - m$. Hence, if $\bar{F}$ is known, then we have common confidence bands for the sequence of paired records $(R_1, R_{1+k}), (R_2, R_{2+k}), \dots$. They are plotted in Sect. 3 for different survival functions $\bar{F}$. This is not the case if we estimate $\bar{F}$ (or a parameter in $\bar{F}$). Of course, note that these general formulas can be used to obtain the particular formulas included in the preceding subsections.

3 Examples

We consider here some classical models to show how to proceed when F is known or when F_θ is known but it contains an unknown parameter θ . We also study the case in which F is completely unknown.

3.1 Uniform distribution

As we have mentioned before, if F is known, then the predictions and the confidence bands do not depend on n . Hence we can plot the paired records, the predictions for new records and the confidence bands in the same figure. Let us see some examples.

The simplest case is a standard uniform distribution with $F(x) = x$ for $0 \leq x \leq 1$. Hence, from (11), the median regression curve to predict R_{n+1} from $R_n = x_n$ (or from $R_1 = x_1, \dots, R_n = x_n$) is

$$m_{n+1|n}(x_n) = \bar{F}^{-1}(0.5\bar{F}(x_n)) = 1 - 0.5(1 - x_n) = 0.5 + 0.5x_n$$

Fig. 1 Plots of the paired records (R_n, R_{n+1}) (left) and (R_n, R_{n+2}) (right) from a standard uniform distribution jointly with the median regression curve (black) and the limit for the 50% (blue) and 90% (red) confidence bands

since $\bar{F}^{-1}(x) = 1 - x$ is the inverse function of $\bar{F}(x) = 1 - x$. The classical mean regression curve can be obtained as

$$\tilde{m}_{n+1|n}(x_n) = x_n + \int_{x_n}^1 \frac{1-x}{1-x_n} dx = 0.5 + 0.5x_n,$$

that is, in this case, both curves coincide.

Analogously, the 50% and 90% confidence bands for the predictions are

$$[0.25 + 0.75x_n, 0.75 + 0.25x_n] \text{ and } [0.05 + 0.95x_n, 0.95 + 0.05x_n].$$

They are plotted in Fig. 1 (left) jointly with $m_{n+1|n}$ and the sequence of paired records $(R_1, R_2), \dots, (R_8, R_9)$. The sequence of the first ninth records obtained by simulation is

0.31916, 0.78427, 0.87288, 0.90177, 0.95036, 0.98365, 0.98411, 0.99982, 0.99996.

The predicted values $m_{n+1|n}(R_n)$ and the confidence intervals $[l_n, u_n]$ (50%) and $[L_n, U_n]$ (90%) for these records are given in Table 1. Note that 5-out-of-8 points belong to the 50% confidence interval (a little bit above the expected coverage) and just 7-out-of-8 belong to the 90% confidence interval (a little bit below the expected coverage). We can compare the prediction for R_9 with that obtained from the (sample) regression line $m(x) = 0.6688 + 0.3097x$ based on the pairs $(R_1, R_2), \dots, (R_7, R_8)$. This prediction is $m(R_8) = 0.6688 + 0.3097 \cdot 0.999815 = 0.9784427$. The prediction in Table 1 is $m_{n+1|n}(R_8) = 0.9999075$ (median regression curve) and the exact value is $R_9 = 0.999959$. The respective absolute errors are 0.02151629 (estimated regression line) and 0.0000515 (exact median regression curve). As expected the second prediction is better (since it takes into account the information about the model).

Table 1 Predicted values and centered confidence intervals $[l_n, u_n]$ (50%) and $[L_n, U_n]$ (90%) for the first nine records from a standard uniform distribution

n	L_n	l_n	$m_{n+1 n}(R_n)$	R_{n+1}	u_n	U_n
1	0.3532067	0.4893737	0.6595825	0.784267	0.8297912	0.9659582
2	0.7950537	0.8382003	0.8921335	0.872883	0.9460667	0.9892133
3	0.8792389	0.9046623	0.9364415	0.901768	0.9682207	0.9936441
4	0.9066796	0.9263260	0.9508840	0.950356	0.9754420	0.9950884
5	0.9528382	0.9627670	0.9751780	0.983650	0.9875890	0.9975178
6	0.9844675	0.9877375	0.9918250	0.984107	0.9959125	0.9991825
7	0.9849016	0.9880802	0.9920535	0.999815	0.9960267	0.9992053
8	0.9998242	0.9998613	0.9999075	0.999959	0.9999537	0.9999907

Analogously, from (16), the median regression curve to predict R_{n+k} from $R_n = x_n$ (or from $R_1 = x_1, \dots, R_n = x_n$) is

$$m_{n+k|n}(x_n) = \bar{F}^{-1}(c_k(0.5)\bar{F}(x_n)) = 1 - c_k(0.5)(1 - x_n) = 1 - c_k(0.5) + c_k(0.5)x_n,$$

where $c_k(0.5) = \exp(-\gamma_k^{-1}(0.5))$ (see (17)). The values of $c_k(0.5)$ for $k = 1, 2, 3, 4, 5$ are

$$0.5, 0.186682309, 0.068971610, 0.025424023, 0.009363755. \quad (18)$$

The code in the statistical program R to get $c_k(u)$ is: `exp(-qgamma(u, k))`.

In a similar way, the 50% and 90% confidence bands for the predictions are

$$[1 - c_k(0.25) + c_k(0.25)x_n, 1 - c_k(0.75) + c_k(0.75)x_n]$$

and

$$[1 - c_k(0.05) + c_k(0.05)x_n, 1 - c_k(0.95) + c_k(0.95)x_n].$$

They are plotted in Fig. 1 (right) for $k = 2$ jointly with $m_{n+2|n}$ and the sequence of the first seventh paired records $(R_1, R_3), \dots, (R_7, R_9)$. The prediction obtained for the ninth record from the seventh one is

$$\hat{R}_9 = m_{n+2|n}(R_7) = 1 - 0.186682309 + 0.186682309 \cdot 0.984107 = 0.9970331$$

with confidence intervals $[0.9939225, 0.998924]$ (50%) and $[0.9888603, 0.9998617]$ (90%). In this case the exact value $R_9 = 0.999959$ is out of these intervals. The absolute error is 0.002925942. If we use the estimated regression line $m(x) = 0.8019 + 0.1832x$ obtained from $(R_1, R_3), \dots, (R_6, R_8)$, we get the prediction $m(R_7) = 0.9821884$ and the absolute error 0.0177706.

The same method can be applied to other (known) distributions. Some examples are given in the next subsections. Note that the quantile functions of the most usual models are included in statistical programs as R.

3.2 PHR model

If the baseline distribution function F_θ has a known parametric form with an unknown parameter θ and $R_1 = x_1, \dots, R_n = x_n$ are known, then we can use (3) to get the likelihood function for θ as

$$\ell(\theta) = h_\theta(x_1) \dots h_\theta(x_n) \bar{F}_\theta(x_n).$$

This expression can be used to obtain the maximum likelihood estimator (MLE) $\hat{\theta}$ of θ . Numerical methods can be used to get this estimator.

Then, as in the preceding subsection, we can use $\bar{F}_{\hat{\theta}}$ to get the estimated median regression curve and the estimated confidence bands replacing $\bar{F}$ with $\bar{F}_{\hat{\theta}}$ in the theoretical results obtained in Sect. 2.

In this subsection we consider a very popular model in reliability and survival studies, the proportional hazard rate (PHR) Cox model. This model contains useful models such as exponential and Pareto distributions (studied in the following subsections). It is defined by a proportional hazard rate function $h_\theta(x) = \theta h(x)$, where h is a known hazard rate function and $\theta > 0$ is an unknown parameter. Hence, its survival function is $\bar{F}_\theta(x) = \bar{F}^\theta(x)$, where $\bar{F}$ is the survival function associated to h . Hence, the likelihood function is

$$\ell(\theta) = \theta^n h(x_1) \dots h(x_n) \bar{F}^\theta(x_n).$$

To maximize ℓ is equivalent to maximize

$$\tilde{\ell}(\theta) = \log \ell(\theta) = n \log \theta + \theta \log \bar{F}(x_n) + \sum_{i=1}^n \log h(x_i)$$

with

$$\tilde{\ell}'(\theta) = \frac{n}{\theta} + \log \bar{F}(x_n).$$

Then the MLE of θ is

$$\hat{\theta}_n = -\frac{n}{\log \bar{F}(x_n)}. \quad (19)$$

Therefore, from (11), the exact median regression curve for the PHR model is

$$m_{n+1|n}(x_n) = \bar{F}_\theta^{-1}(0.5 \bar{F}_\theta(x_n)) = \bar{F}^{-1}(0.5^{1/\theta} \bar{F}(x_n))$$

since $\bar{F}_\theta^{-1}(y) = \bar{F}^{-1}(y^{1/\theta})$ is the inverse function of $\bar{F}_\theta$. If we replace θ with the MLE $\hat{\theta}_n$ given in (19), we obtain the maximum likelihood *estimated median regression curve* (EMRC) as

$$\hat{m}_{n+1|n}(x_n) = \bar{F}^{-1}(0.5^{1/\hat{\theta}_n} \bar{F}(x_n)) = \bar{F}^{-1}\left(\bar{F}(x_n) 0.5^{-\frac{1}{n} \log \bar{F}(x_n)}\right).$$

Analogously, the 50% and 90% *estimated quantile confidence bands* (EQCB) are

$$\left[\bar{F}^{-1}\left(\bar{F}(x_n) 0.75^{-\frac{1}{n} \log \bar{F}(x_n)}\right), \bar{F}^{-1}\left(\bar{F}(x_n) 0.25^{-\frac{1}{n} \log \bar{F}(x_n)}\right) \right]$$

and

$$\left[\bar{F}^{-1}\left(\bar{F}(x_n) 0.95^{-\frac{1}{n} \log \bar{F}(x_n)}\right), \bar{F}^{-1}\left(\bar{F}(x_n) 0.05^{-\frac{1}{n} \log \bar{F}(x_n)}\right) \right],$$

respectively.

In the general case, if we want to predict R_{n+k} from R_n for $k > 0$, from (16), the EMRC is

$$m_{n+k|n}(x_n) = \bar{F}_{\hat{\theta}_n}^{-1}\left(c_k(0.5) \bar{F}_{\hat{\theta}_n}(x_n)\right) = \bar{F}^{-1}\left(\bar{F}(x_n) (c_k(0.5))^{-\frac{1}{n} \log \bar{F}(x_n)}\right), \quad (20)$$

where c_k is defined in (17). The EQCB are obtained in a similar way by replacing 0.5 with 0.05, 0.25, 0.75, 0.95.

3.3 Exponential distribution

The exponential model with survival function $\bar{F}_\theta(x) = \exp(-\theta x)$ for $x \geq 0$ satisfies the PHR model with $h(x) = 1$ and $\bar{F}(x) = \exp(-x)$. Then, from the results given in the preceding subsection, the MLE for θ is $\hat{\theta}_n = n/x_n$. This is a well known result (see, e.g., Awad and Raqab 2000).

If θ is known, the exact median regression curve to predict R_{n+1} from $R_n = x_n$ (or from $R_1 = x_1, \dots, R_n = x_n$) is

$$m_{n+1|n}(x_n) = \bar{F}^{-1}(0.5^{1/\theta} \bar{F}(x_n)) = x_n - \log(0.5) \frac{1}{\theta}$$

since $\bar{F}^{-1}(x) = -\log(x)$ is the inverse function of $\bar{F}(x) = \exp(-x)$.

If θ is unknown, we replace θ with the MLE $\hat{\theta}_n$, obtaining the maximum likelihood EMRC as

$$\hat{m}_{n+1|n}(x_n) = x_n - \log(0.5) \frac{1}{\hat{\theta}_n} = x_n - \frac{\log(0.5)}{n} x_n. \quad (21)$$

The maximum likelihood *estimated regression curve (ERC)* $\tilde{m}_{n+1|n}(x_n) = E_{\hat{\theta}}(R_{n+1}|R_n = x_n)$ can be obtained as

$$\tilde{m}_{n+1|n}(x_n) = x_n + \int_{x_n}^{\infty} \frac{\exp(-\hat{\theta}x)}{\exp(-\hat{\theta}x_n)} dx = x_n + \frac{1}{\hat{\theta}_n} = x_n + \frac{1}{n}x_n, \quad (22)$$

that is, in this case, both curves (lines) do not coincide.

Another option is to use the joint MLP for (θ, R_{n+1}) from $R_1 = x_1, \dots, R_n = x_n$. It is well known that this procedure leads to $\hat{\theta} = (n+1)/x_n$ and the trivial estimator $\hat{R}_{n+1} = x_n$.

As for the exponential distribution both curves $\hat{m}_{n+1|n}(x_n)$ and $\tilde{m}_{n+1|n}(x_n)$ are straight lines, we could also use a (sample) linear mean (or quantile) regression to predict R_{n+1} from a sequence of paired records (as in Sect. 3.1).

From the general results for the PHR model, the 50% and 90% *estimated quantile confidence bands (EQCB)* for the exponential distribution are

$$\left[x_n - \log(0.75) \frac{1}{\hat{\theta}_n}, x_n - \log(0.25) \frac{1}{\hat{\theta}_n} \right] = \left[x_n - \log(0.75) \frac{x_n}{n}, x_n - \log(0.25) \frac{x_n}{n} \right]$$

and

$$\left[x_n - \log(0.95) \frac{1}{\hat{\theta}_n}, x_n - \log(0.05) \frac{1}{\hat{\theta}_n} \right] = \left[x_n - \log(0.95) \frac{x_n}{n}, x_n - \log(0.05) \frac{x_n}{n} \right],$$

respectively. They were obtained previously in Awad and Raqab (2000). Note that they depend on n .

If θ is known, the exact confidence bands are obtained just by replacing $\hat{\theta}_n$ with θ . In this case, they do not depend on n and we can plot together the regression curves, the confidence bands and the sequence of paired records. We do so in Fig. 2, left, for $\theta = 1$ including the exact regression curves and the sequence of the first eight paired records $(R_1, R_2), \dots, (R_8, R_9)$. The simulated sequence of records $R_1, \dots, R_9$ obtained is

0.45701, 1.19403, 1.74177, 2.00398, 2.25833, 4.97619, 5.84512, 6.90868, 14.16568.

The purple line represents the sample regression line $m(x) = -0.2967 + 1.6335x$ obtained from $(R_1, R_2), \dots, (R_8, R_9)$ which is a very bad approximation of both theoretical regression lines (it is strongly determined by the data and it does not take into account the information about the model).

If θ is unknown, the estimated confidence bands $[l_n, u_n]$ (50%) and $[L_n, U_n]$ (90%) depend on n . They are plotted (blue, red) in Fig. 2, right, for $\theta = 1$ by connecting the points obtained for $n = 1, \dots, 8$. We also include the estimated regression curves $\hat{m}_{n+1|n}$ (black) and $\tilde{m}_{n+1|n}$ (green) and the simulated sequence of paired records. These values are given in Table 2. In this sample, the real coverage probabilities for $[l_n, u_n]$ and $[L_n, U_n]$ are 0.625 (5-out-of-8) and 0.75 (6-out-of-8), respectively. The

Fig. 2 Plot of the paired records (R_n, R_{n+1}) from a standard exponential distribution jointly with the median regression curve (black), the theoretical and sample regression curves (green, purple) and the limits for the 50% (blue) and 90% (red) exact confidence bands (left). The same is done in the right plot by assuming that θ is unknown

MLE predictions from $\hat{\theta}_n = n/R_n$ of $\theta = 1$ for $n = 1, \dots, 8$ are

2.18813, 1.67499, 1.72239, 1.99603, 2.21403, 1.20574, 1.19758, 1.15796.

They are used both in $\hat{m}_{n+1|n}(R_n)$ and $\tilde{m}_{n+1|n}(R_n)$. The mean squared errors (MSE) for these eight predictions are 12.39879 and 13.70644, respectively. Analogously, the mean absolute errors (MAE) are 2.684359 and 2.877943. However, in this case, the joint maximum likelihood predictor (MLP) $\hat{R}_{n+1} = R_n$, provides the best predictions with respective errors 9.747176 and 2.320351. Note that the MLP are always out of the centered confidence bands. If we use this joint MLP, they should be replaced with the bottom confidence bands

$$\left[x_n, x_n - \log(0.5) \frac{1}{\hat{\theta}} \right] = \left[x_n, x_n - \log(0.5) \frac{x_n}{n} \right]$$

(50%) and

$$\left[x_n, x_n - \log(0.1) \frac{1}{\hat{\theta}} \right] = \left[x_n, x_n - \log(0.1) \frac{x_n}{n} \right]$$

(90%), where $x_n = R_n$.

Analogously, from (16), the exact median regression curve to predict R_{n+k} from $R_n = x_n$ (or from $R_1 = x_1, \dots, R_n = x_n$) is

$$m_{n+k|n}(x_n) = \bar{F}^{-1}(c_k(0.5)\bar{F}(x_n)) = x_n - \frac{1}{\theta} \log(c_k(0.5)), \quad (23)$$

where c_k is given in (17) (some of these values were given in (18)). The confidence bands and their estimated versions can be obtained in a similar way. For example, the

Table 2 Predicted values $\hat{R}_{n+1} = \hat{m}_{n+1|n}(R_n)$ and $\tilde{R}_{n+1} = \tilde{m}_{n+1|n}(R_n)$ and centered confidence intervals $[l_n, u_n]$ (50%) and $[L_n, U_n]$ (90%) for the first nine records from a standard exponential distribution when θ is unknown. R_{n+1} represents the exact values for $n = 1, \dots, 8$

n	L_n	l_n	$\hat{R}_{n+1}$	$\tilde{R}_{n+1}$	R_{n+1}	u_n	U_n
1	0.4804537	0.5884862	0.7737886	0.914024	1.19403	1.090565	1.826098
2	1.2246539	1.3657817	1.6078506	1.791047	1.74177	2.021670	2.982530
3	1.7715483	1.9087931	2.1442019	2.322357	2.00398	2.546636	3.481058
4	2.0296757	2.1481051	2.3512409	2.504973	2.25833	2.698504	3.504823
5	2.2814954	2.3882641	2.5713987	2.709994	4.97619	2.884469	3.611397
6	5.0187319	5.2147845	5.5510631	5.805556	5.84512	6.125935	7.460747
7	5.8879488	6.0853374	6.4239076	6.680135	6.90868	7.002697	8.346605
8	6.9529741	7.1571159	7.5072693	7.772263	14.16568	8.105861	9.495747

EMRC is

$$\hat{m}_{n+k|n}(x_n) = x_n - \frac{1}{\hat{\theta}_n} \log(c_k(0.5)) = x_n - \frac{x_n}{n} \log(c_k(0.5)). \quad (24)$$

Remark 1 Let us compare the different predictors for R_s from R_n in this model. The best linear unbiased predictor (BLUP) is $\hat{R}_s^{(1)} = \frac{s}{n} R_n$, the best linear invariant predictor (BLIP) is $\hat{R}_s^{(2)} = \frac{s+1}{n+1} R_n$, and the maximum likelihood predictor (MLP) is $\hat{R}_s^{(3)} = \frac{s}{n+1} R_n$. They were compared in Awad and Raqab (2000) as well. The predictor obtained from Volovskiy and Kamps (2020a) for this model is $\hat{R}_s^{(4)} = \frac{s-1}{n} R_n$. Finally, the one obtained from (24) is $\hat{R}_s^{(5)} = (1 - \frac{1}{n} \log(c_k(0.5))) R_n$. We obtain $m = 1000$ simulated values for R_n and R_s with $\theta = 1$, $n = 1$ and $s = 2, 3, 4, 5$. To compare the predictions we use the mean squared error

$$MSE_s(i) = \frac{1}{m} \sum_{j=1}^m (R_s(j) - \hat{R}_s^{(i)}(j))^2$$

and the mean absolute error

$$MAE_s(i) = \frac{1}{m} \sum_{j=1}^m |R_s(j) - \hat{R}_s^{(i)}(j)|$$

for $i = 1, 2, 3, 4, 5$, where $R_s(j)$ and $\hat{R}_s^{(i)}(j)$ represent the exact and estimated values obtained in the $m = 1000$ simulations for $j = 1, \dots, m$. The MSE and MAE obtained are given in Table 3. In this case (and with these error measures), the best predictions are obtained with the BLIP ($i = 2$). For more comparisons see Raqab and Nagaraja (1995) and Volovskiy and Kamps (2020a, b).

Table 3 $MSE_S(i)$ and $MAE_S(i)$ for the predictors considered in Remark 1 for the record values R_S and R_1 in an exponential model with $\theta = 1$

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
$MSE_2(i)$	2.134341	1.632646	2.146251	2.146251	1.706099
$MAE_2(i)$	1.0521102	0.8660707	1.0223854	1.0223854	0.9171513
$MSE_3(i)$	5.224474	3.633845	4.209561	3.633845	4.313979
$MAE_3(i)$	1.697259	1.443774	1.581928	1.443774	1.563959
$MSE_4(i)$	10.320037	6.890963	7.514841	7.150536	8.898777
$MAE_4(i)$	2.372153	2.073787	2.191918	2.087500	2.251814
$MSE_5(i)$	21.17106	12.26291	12.71946	14.57012	18.53319
$MAE_5(i)$	3.336490	2.822688	2.916739	2.937301	3.183455

3.4 Pareto distribution

Another popular model studied in the literature of records is the Pareto (type II) model with survival function $\bar{F}_\theta(x) = (1+x)^{-\theta}$ and hazard rate function $h(x) = \theta/(1+x)$ for $x \geq 0$ and $\theta > 0$. It is also included in the PHR model.

If θ is known, from (11), the exact median regression curve to predict R_{n+1} from $R_n = x_n$ (or from $R_1 = x_1, \dots, R_n = x_n$) is

$$m_{n+1|n}(x_n) = \bar{F}_\theta^{-1}(0.5\bar{F}_\theta(x_n)) = -1 + (1+x_n)0.5^{-1/\theta} \quad (25)$$

since $\bar{F}_\theta^{-1}(x) = -1 + x^{-1/\theta}$ is the inverse function of $\bar{F}_\theta$. The exact 50% and 90% confidence bands are obtained in a similar way as

$$\left[-1 + (1+x_n)0.75^{-1/\theta}, -1 + (1+x_n)0.25^{-1/\theta} \right]$$

and

$$\left[-1 + (1+x_n)0.95^{-1/\theta}, -1 + (1+x_n)0.05^{-1/\theta} \right],$$

respectively. They are plotted in Fig. 3 (left) for $\theta = 2$ jointly with the sequence of the first nine records obtained by simulation:

$$0.40174, 0.42488, 0.49916, 1.48359, 4.0376, 4.90751, \\ 10.24807, 39.27022, 49.69001.$$

Note that the curve and the regions do not depend on n .

In this case, the classical mean regression curve is

$$\tilde{m}_{n+1|n}(x_n) = E(R_{n+1}|R_n = x_n) = x_n + \int_{x_n}^{\infty} \left(\frac{1+x}{1+x_n} \right)^{-\theta} dx = \frac{1+\theta x_n}{\theta-1}$$

for $\theta > 1$. If $0 < \theta \leq 1$ it does not exist (however the curve $m_{n+1|n}$ in (25) always exists). The curve $\tilde{m}_{n+1|n}$ is also included in Fig. 3, left (green). Note that $m_{n+1|n} \neq \tilde{m}_{n+1|n}$. Actually, in this case ($\theta = 2$), $\tilde{m}_{n+1|n}$ coincides with the upper limit of the 50% confidence band (i.e. the third conditional quartile curve) since $-1 + 0.25^{-1/\theta}(1 + x_n) = 2x_n - 1$.

If θ is unknown, from the general results for the PHR, the MLE for θ is $\hat{\theta}_n = n / \log(1 + x_n)$. This is a well known result (see, e.g., Paul and Thomas 2016). Note that again it only depends on x_n . Therefore, we just replace θ with $\hat{\theta}_n$ in (25) to get the EMRC for the Pareto model as

$$\hat{m}_{n+1|n}(x_n) = -1 + (1 + x_n)0.5^{-1/\hat{\theta}_n} = -1 + (1 + x_n)0.5^{-\frac{1}{n} \log(1+x_n)}.$$

It is plotted in Fig. 3, right, for the nine simulated records stated above. In that plot we also include the estimated regression curves $\tilde{m}_{n+1|n}$ (green) and the estimated centered confidence bands $[l_n, u_n]$ (50%) and $[L_n, U_n]$ (90%) (that depend on n) obtained by connecting the points in the intervals for $n = 1, \dots, 8$. These values are given in Table 4. In this sample, the real coverage probabilities for $[l_n, u_n]$ and $[L_n, U_n]$ are 0.25 (2-out-of-8) and 0.5 (4-out-of-8), respectively (below the expected values). As in the exponential case, the joint MLP for R_{n+1} and θ gives $\hat{R}_{n+1} = x_n$ and $\theta = (n + 1) / \log(1 + x_n)$. The MSE for these eight predictions by using the EMRC, ERC and MLP are 89.88488, 141.3754 and 123.4542, respectively. Analogously, the MAE are 4.821381, 6.890559 and 6.161034. In this example, the EMRC provides the best fits.

The MLE $\hat{\theta}_n = n / \log(1 + R_n)$ for $\theta = 2$ that we obtain with the nine records are

$$2.96111, 5.64834, 7.40914, 4.39704, 3.09228, 3.37795, 2.89233, 2.16473, 2.29257.$$

Therefore, the prediction for the next upper record from the EMRC is

$$\hat{m}_{10|9}(R_9) = -1 + (1 + 49.690004)0.5^{-1/2.292568} = 67.58501.$$

Analogously, the one obtained with the estimated regression curve is 88.90651 and the estimated centered confidence intervals are $[56.46714, 91.79747]$ (50%) and $[50.83691, 186.2498]$ (90%).

Analogously, from (20), the median regression curve to predict R_{n+k} from $R_n = x_n$ is

$$m_{n+k|n}(x_n) = \bar{F}_\theta^{-1}(c_k(0.5)\bar{F}_\theta(x_n)) = -1 + (c_k(0.5))^{-1/\theta}(1 + x_n),$$

where c_k is given in (17). The confidence bands and their estimated versions are obtained in a similar way by replacing θ with the MLE $\hat{\theta}_n = n / \log(1 + R_n)$.

3.5 Non-parametric predictions

If $\bar{F}$ is unknown and we know the sequence of original values $X_1, \dots, X_N$ which lead to the record values $R_1, \dots, R_n$ ($n \leq N$), then the best option is to use the empirical

Table 4 Predicted values $\hat{R}_{n+1} = \hat{m}_{n+1|n}(R_n)$ and $\tilde{R}_{n+1} = \tilde{m}_{n+1|n}(R_n)$ and centered confidence intervals $[l_n, u_n]$ (50%) and $[L_n, U_n]$ (90%) for the first nine records from a Pareto distribution when $\theta = 2$ and it is unknown

n	L_n	l_n	$\hat{R}_{n+1}$	$\tilde{R}_{n+1}$	R_{n+1}	u_n	U_n
1	0.4262289	0.5447545	0.7714443	1.1165035	0.42488	1.2386634	2.855130
2	0.4378764	0.4993300	0.6109163	0.7314127	0.49916	0.8212445	1.421676
3	0.5095757	0.5585153	0.6461820	0.7330711	1.48359	0.8076212	1.246190
4	1.5127277	1.6515114	1.9076439	2.2146896	4.03760	2.4041071	3.908707
5	4.1218582	4.5287520	5.3033664	6.4453081	4.90751	6.8871741	12.27268
6	4.9978944	5.4326615	6.2530434	7.3917883	10.2481	7.9050504	13.34028
7	10.449328	11.424379	13.294103	16.192115	39.2702	17.165010	30.68830
8	40.235822	44.993841	54.468359	73.844976	49.6900	75.402327	159.6930

Fig. 3 Plots of the paired records (R_n, R_{n+1}) from a Pareto distribution with $\theta = 2$, jointly with the median regression curve (black), the mean regression curve (green) and the 50% (blue–green) and 90% (red) exact confidence bands (left). The same is done in the right plot by assuming that θ is unknown

(or kernel) estimator for F in (11) and (16). With the empirical estimator, the survival function $\bar{F}$ is approximated with

$$S(x) = \frac{1}{N} \sum_{i=1}^N 1(X_i > x)$$

and its inverse with

$$S^{-1}(x) = X_{i:N}$$

if $N - i \leq Nx < N - i + 1$ for $i = 1, \dots, N$, where $X_{1:N}, \dots, X_{N:N}$ represent the ordered data. Therefore, the EMRC from $X_1, \dots, X_N$ and (11) is

$$\hat{m}_{n+1|n}(x) = S^{-1}(0.5S(x)).$$

The estimated confidence regions and the estimated median regression curve $\hat{m}_{n+k|n}$ are obtained in a similar way.

For example, if we use the data which lead to the record values in Sect. 3.1 (uniform distribution), then we obtain EMRC and EQCB very similar to that plotted in Fig. 1 (since $N = 20000$). If we use the data in Sect. 3.3 (exponential distribution), then we obtain the EMRC and the EQCB plotted in Fig. 4, left (black and blue), where we also include the sequence of records (symbols '+'), the exact curves (red) and the EMRC obtained with the MLE of θ with $n = 9$ (green) when we know that the model is an exponential distribution. In Fig. 4, right, we plot the respective curves from the Pareto sample in Sect. 3.4.

If $\bar{F}$ is unknown and we just know the sequence of record values $R_1, \dots, R_n$, then $\bar{F}$ should be estimated from the likelihood function given in (3). By using the inversion

Fig. 4 Plots of the paired records (R_n, R_{n+1}) from a standard exponential distribution (left) and a Pareto distribution (right), jointly with the non-parametric predictions of the median regression curves and the confidence bands (black and blue), the exact ones (red) and the EMRC curve with the MLE of θ (green)

formula for the hazard rate function, it can be written as

$$\ell = h(x_1) \dots h(x_n) \exp \left(- \int_{-\infty}^{x_n} h(x) dx \right).$$

If

$$\tilde{\ell} = \log \ell = \sum_{i=1}^n \log h(x_i) - \int_{-\infty}^{x_n} h(x) dx, \quad (26)$$

we have to maximize $\tilde{\ell}$ for the functions h in the set of all the hazard rate functions of absolutely continuous distributions. This is a difficult task. Let us see some possible approximate solutions.

If the left-end point of the support $x_0 = \inf\{x : F(x) > 0\}$ is finite, then the integral in the expression of L can be approximated by

$$\int_{x_0}^{x_n} h(x) dx \approx \sum_{i=1}^n (x_i - x_{i-1}) h(x_i). \quad (27)$$

This is equivalent to assume that the hazard rate is constant but that this constant can change at the record values (a reasonable assumption in practice). With this approximation we have

$$\tilde{\ell} \approx \sum_{i=1}^n [\log h(x_i) - (x_i - x_{i-1}) h(x_i)] = \sum_{i=1}^n \phi_i(h(x_i)), \quad (28)$$

where $\phi_i(u) = \log u - (x_i - x_{i-1})u$, $\phi'_i(u) = 1/u - (x_i - x_{i-1})$ and $\phi''_i(u) = -1/u^2 < 0$. Therefore, the maximum value in (28) is obtained when

$$\hat{h}(x_i) = \frac{1}{x_i - x_{i-1}}$$

for $i = 1, \dots, n$. If we define $\hat{h}(x)$ as $\hat{h}(x_i)$ for $x_{i-1} < x \leq x_i$ and $i = 1, \dots, n$, then $\bar{F}$ can be estimated with

$$\begin{aligned} S(x) &= \exp\left(-\int_{x_0}^x \hat{h}(z)dz\right) \\ &= \exp\left(1 - j - (x - x_{j-1})\hat{h}(x_j)\right) \\ &= \exp\left(1 - j - \frac{x - x_{j-1}}{x_j - x_{j-1}}\right) \end{aligned} \quad (29)$$

if $x_{j-1} < x \leq x_j$ ($S(x) = 1$ for $x \leq x_0$ and $S(x) = 0$ for $x > x_n$). Therefore its inverse is estimated with the inverse of S given by

$$S^{-1}(y) = x_{j-1} + (x_j - x_{j-1})(1 - j - \log y)$$

for $e^{-j} \leq y < e^{1-j}$ and $j = 1, \dots, n$. If $0 < y < e^{1-n}$, it can be defined as $S^{-1}(y) = x_n$. Then we can use the predictions in (11) and (16) to get approximations of the median regression curve and the associated confidence regions.

We do that for the simulated record values obtained above from the exponential and Pareto models. Thus we get the estimated curves plotted in Fig. 5 (continuous lines). The dashed lines represent the exact curves (see Sects. 3.3 and 3.4). The green dashed lines represent the EMRC obtained with the MLE for θ . Note that the EMRC obtained from S fit better to the record values. This is due to the bias induced by the fact that S is computed from these record values (the same happen, e.g., in the linear regression models). We cannot expect the same accuracy for new records. Actually, the estimated curves based on the MLE fit better to the exact curves than the ones obtained from the non-parametric method (as expected).

We can use other options to maximize $\tilde{\ell}$ in (26). For example, we could use a better approximations for the integral in that expression than that used in (27).

Another popular option is to restrict the set of possible hazard functions. For example, we might assume only linear functions, that is, $h(x) = ax + b$ for $x \geq 0$ and $a, b \geq 0$. This is equivalent to consider the MLE method applied to the Linear Hazard Rate (LHR) model with $\bar{F}(x) = \exp(-bx - ax^2/2)$ for $x \geq 0$. This model contains the exponential model ($a = 0$), some Weibull models ($b = 0$) and it is equivalent to the normal (Gaussian) model when $a > 0$ and $x \rightarrow \infty$. Thus, from (26), we need to maximize

Fig. 5 Plots of the paired records (R_n, R_{n+1}) from a standard exponential distribution (left) and a Pareto distribution (right), jointly with the non-parametric predictions (continuous lines) obtained from (29) of the median regression curve (black) and the 50% and 90% confidence bands (blue, red). The dashed lines represent the respective exact curves and the green dashed lines the EMRC from the MLE of θ

$$\begin{aligned}\tilde{\ell}(a, b) &= \sum_{i=1}^n \log(ax_i + b) - \int_0^{x_n} (ax + b)dx \\ &= -\frac{1}{2}ax_n^2 - bx_n + \sum_{i=1}^n \log(ax_i + b).\end{aligned}\quad (30)$$

The partial derivatives are

$$\partial_1 \tilde{\ell}(a, b) = -\frac{1}{2}x_n^2 + \sum_{i=1}^n \frac{x_i}{ax_i + b} \quad (31)$$

and

$$\partial_2 \tilde{\ell}(a, b) = -x_n + \sum_{i=1}^n \frac{1}{ax_i + b}. \quad (32)$$

So we have to solve the equations

$$\sum_{i=1}^n \frac{x_i}{ax_i + b} = \frac{x_n^2}{2} \quad (33)$$

and

$$\sum_{i=1}^n \frac{1}{ax_i + b} = x_n. \quad (34)$$

We can use numerical methods to solve these equations. For example, we could apply the bivariate Newton-Raphson (NR) method (see, e.g., Gil et al. 2007) by using (31)

Fig. 6 Plots of the paired records (R_n, R_{n+1}) from a model with hazard rate $h(x) = 2x + 3$ jointly with the exact (left) and NR predictions (right) of the median regression curve (black) and the 50% and 90% confidence bands (blue, red)

and (32). Thus, we choose an initial point (a_0, b_0) (e.g. $a_0 = b_0 = 1$) and then we compute the next points as

$$(a_{m+1}, b_{m+1})' = (a_m, b_m)' - H^{-1}(a_m, b_m) V(a_m, b_m)$$

for $m = 0, 1, \dots$, where v' denotes the transpose of v ,

$$V(a_m, b_m) = \left(\partial_1 \tilde{\ell}(a_m, b_m), \partial_2 \tilde{\ell}(a_m, b_m) \right)'$$

is the vector with the first partial derivatives, $H(a_m, b_m) = (\partial_{i,j} \tilde{\ell}(a_m, b_m))$ is the Hessian matrix with the second partial derivatives and $H^{-1}(a_m, b_m)$ is its inverse matrix.

We apply this method to the first eleven records from a model with hazard rate $h(x) = 2x + 3$. After six iterations ($m = 6$) we obtain the MLE predictions of $a = 2$ and $b = 3$, $\hat{a} = 0.8406102$ and $\hat{b} = 4.7184054$ with a maximum value of $\tilde{\ell}(\hat{a}, \hat{b}) = 7.947086$. These estimations are not very good. However, the EMRC and the EQCB obtained from them, plotted in Fig. 6 (right), are very similar to the exact ones (given in the left plot).

4 A case with a real data set

We consider the data set studied in Awad and Raqab (2000), Sect. 5. It contains sizes of rocks that should be crushed. The data were published in Dunsmore (1983). The sizes are

9.3, 0.6, 24.4, 18.1, 6.6, 9.0, 14.3, 6.6, 13, 2.4, 5.6, 33.8.

Fig. 7 Plots of the paired records (R_n, R_{n+k}) for $k = 1$ (left) and $k = 2$ (right) from the real data set in Dunsmore (1983) jointly with the predictions of the median and mean regression curves (black and purple) and the 50% and 90% confidence bands (blue, red) under an exponential model

Hence, the upper record values are $R_1 = 9.3$, $R_2 = 24.4$ and $R_3 = 33.88$. The purpose is to predict $R_4, R_5, \dots$. In Awad and Raqab (2000), it is assumed that the data come from an exponential distribution with an unknown parameter θ . Then, as we have seen before, the MLE for θ is $\hat{\theta}_n = n/R_n$, obtaining the following predictions

$$\hat{\theta}_1 = 1/R_1 = 0.10752688, \quad \hat{\theta}_2 = 2/R_2 = 0.08196721, \quad \hat{\theta}_3 = 3/R_3 = 0.08854782.$$

The stability in that predictions may confirm that the exponential model is correct. The estimation with the complete sample is $\hat{\theta} = 0.08350731$. Then, to predict R_4 we can use the curves given in (21) and (22) with the estimation $\hat{\theta}_3 = 3/R_3 = 0.08854782$. They are plotted in Fig. 7 (left, black and purple lines) jointly with the corresponding 50% (blue) and 90% (red) confidence bands for these predictions. We also plot the paired record sequence (R_1, R_2) and (R_2, R_3) . As we can see they belong to the 50% confidence bands. Therefore, the exponential model cannot be rejected with these data. The respective predictions for R_4 are 41.70794 (black) and 45.17333 (purple) and the confidence intervals are $[37.12889, 49.53588]$ (50%) and $[34.45927, 67.7118]$ (90%).

If we want to predict R_5 from R_3 , we can use the curve obtained from (23)

$$\hat{m}_{n+2|n}(x_n) = x_n - \frac{1}{\hat{\theta}} \log(c_k(0.5)) = x_n + 1.678347 \frac{1}{\hat{\theta}}.$$

In our case, with $\hat{\theta}_3$, we get $m_{5|3}(x_3) = x_3 + 18.95413$ (see Fig. 7, right, black line) which gives the prediction $\hat{R}_5 = m_{5|3}(33.8) = 52.83413$. In this figure we also provide the estimated centered confidence bands $[44.73604, 64.28882]$ (50%) and $[37.89322, 87.45404]$ (90%) for R_5 . The only paired data available (R_1, R_3) is also plotted. Note that it belongs to the 50% confidence region.

Table 5 Available lifetimes X_I after 75 days in order of availability. The variable I represents the index in the original sample

Day k_I	11	13	23	25	30	41
I	8	9	2	23	27	25
X_I	3.848409	4.979474	21.59455	2.367513	3.260327	16.73338
Day k_I	42	46	51	51	52	55
I	19	21	26	31	7	38
X_I	23.66431	25.64172	25.14744	20.53256	45.24274	17.60179
Day k_I	59	60	60	61	65	68
I	34	53	13	49	42	33
X_I	25.10737	7.558565	47.89888	12.62512	23.51588	35.82169
Day k_I	68	70	71	73	73	75
I	52	43	66	36	4	44
X_I	16.84676	27.41309	5.142253	37.00595	69.04814	31.02384

The bold case indicates lower record values

5 A case study in reliability

In this section we show a simulated case study in reliability. We assume that in a production process, every day, a random chosen unit is put into operation (under a normal test or an accelerated test procedure). Let $X_1, X_2, \dots$ be the IID lifetimes of these units (think, e.g., on lifetimes of batteries in mobile phones or cars). After c days, only some of these values are available in a complete form, that is, are not censored. For example, after $c = 3$ days, X_1 is known if and only if $X_1 < 3$. In general, after c days, X_j is available if and only if $X_j < c - j + 1$ (or $X_j + j - 1 < c$). In other words, X_j is known the k_j -th day, with $k_j = [X_j + j]$, where $[x]$ represents the integer part of x . The procedure described below based on these assumptions can be adapted to other (similar) sampling schemes. The key idea is to use lower record values because they will be available before and with a small proportion of censored data.

We want to use this process to study early failures. Therefore, we can assume that the units do not have aging and that their lifetimes follow an exponential distribution F with an unknown mean $\mu > 0$ and constant hazard rate $h(t) = \theta = 1/\mu$ for $t \geq 0$. The purpose is to estimate θ (or μ) and the lower record values. Here it is important to discriminate small values from F and outliers (i.e. small values not obtained from F). To illustrate this process, we simulate 100 data. The 36 available data lifetimes after $c = 75$ days are given in Table 5 (in order of availability). The other data are censored. The available lower record values in this sequence are in bold case.

The incomplete (censored) sequence of lower records obtained from this table is $R_1^* = 75^+$, $R_2^* = 21.59455$, $R_3^* = 3.848409$ and $R_4^* = 2.367513$. The complete sequence of lower records from these 100 data points is 89.21383, 21.59455, 3.848409, 2.367513. To get the exact value of X_1 we have to wait until the 90-th

day. However, our censored sequence at day 75 is a good approximation of the exact sequence of lower records. To get the complete sample we have to wait until the 467-th day where $X_{15} = 452.5437$ will be available.

Instead, we want just to use the information available the 75-th day (given in Table 5). The likelihood function for the lower record values $(R_1^*, \dots, R_n^*)$ in the exponential model is

$$\ell(\theta) = \bar{h}(x_1) \dots \bar{h}(x_n) F(x_n) = \theta^n e^{-\theta x_n} \prod_{j=1}^{n-1} \frac{e^{-\theta x_j}}{1 - e^{-\theta x_j}},$$

where $\bar{h} = f/F$ is the reversed hazard rate. As a consequence, the joint distribution of $(R_1^*, \dots, R_n^*)$ can also be written as a distorted distribution

$$\mathbf{G}(x_1, \dots, x_n) = D(F(x_1), \dots, F(x_n))$$

for $x_1 > \dots > x_n$, where the PDF d of D is equal to the one in (6). Actually, D coincides with the distortion function $\hat{D}$ in (4).

To estimate θ from our censored sample with the MLE, we maximize ℓ obtaining $\hat{\theta} = 0.01645$ (or $\hat{\mu} = 60.77$). If we wait until the 90-th day to get the complete lower record sample, then the MLE prediction is $\hat{\theta} = 0.01437$ (or $\hat{\mu} = 69.55$).

The conditional distribution of $(R_{n+1}^* | R_1^* = x_1, \dots, R_n^* = x_n)$ is

$$F(x | x_1, \dots, x_n) = \frac{F(x)}{F(x_n)} \quad (35)$$

for $x \leq x_n$. Therefore, the median regression curve to predict R_{n+1}^* from R_n^* is

$$m_{n+1|n}^*(x_n) = F^{-1}(0.5F(x_n))$$

and in the case of an exponential distribution we obtain

$$m_{n+1|n}^*(x_n) = -\frac{1}{\theta} \ln(0.5 + 0.5 \exp(-\theta x_n)).$$

It is plotted in Fig. 8 by replacing θ with the predictions obtained above from the censored (left) and uncensored (right) samples. We also include in the plots the sequences of paired lower records and the centered 50% confidence band $[F^{-1}(0.25F(x_n)), F^{-1}(0.75F(x_n))]$ and 90% confidence band $[F^{-1}(0.05F(x_n)), F^{-1}(0.95F(x_n))]$. The respective predictions for the next lower record R_5^* are 1.1722 and 1.1736 (they are also included in the plots). We also plot the exact curves (dashed lines) and thus we can see that the predictions are good (in both cases).

To predict the next lower record we can also use the mean regression curve obtained from (35) as

$$\tilde{m}_{n+1|n}^*(x) = E(R_{n+1}^* | R_n^* = x_n) = \frac{1}{\theta} - \frac{x \exp(-\theta x)}{1 - \exp(-\theta x)}$$

Fig. 8 Plots of the paired lower records (R_n^*, R_{n+1}^*) for the censored (left) and uncensored (right) samples from the data in Table 5 jointly with the predictions of the median and mean regression curves (black and purple) and the 50% and 90% confidence bands (blue, red) under an exponential model. The dashed lines represent the exact curves (obtained with $\theta = 0.01$)

for $x \geq 0$. It is also plotted in Fig. 8 (purple lines) for the different estimations of θ . The predictions from this curve for R_5^* are 1.176064 (censored sample, day 75) and 1.177034 (complete record sample, day 90).

The available data in Table 5 can also be used to get other lower record sequences. For example, starting with $R_1^* = X_4 = 69.04814$, we get the sequence $R_2^* = X_7 = 45.24274$, $R_3^* = X_8 = 3.848409$ and $R_4^* = X_{23} = 2.367513$. With this sequence, the MLE for θ gives $\hat{\theta} = 0.0144$ (or $\hat{\mu} = 69.515$). These predictions can be used to improve the estimated curves plotted above.

6 Conclusions

We have proposed a very general method to predict record values from preceding records. We focused on upper records but the results for lower records are completely analogous (see Sect. 5). These predictions are based on conditional quantiles which can also be used to provide confidence bands. We first obtain the theoretical expressions and we then show how to apply them in practice in three different cases: a given probability (distribution) model, a given parametric model with an unknown parameter and a non-parametric method for arbitrary models. All these cases show that the proposed method is a good way to manage the uncertainty in the prediction of future record values. The confidence bands are more appropriate to this purpose than point estimators.

This paper is just a first step based on the concept of multivariate distorted distribution introduced in Navarro et al. (2020). The present study for records should be completed by considering MLE for other relevant models (we just study here uniform, PHR, exponential and Pareto models). Other estimation procedures can be considered as well (see e.g., Volovskiy and Kamps 2020a, and the references therein). The non-parametric estimations could be improved by using kernel estimators (that provide continuous curves). Moreover, similar techniques can be used to predict other ordered values such as order statistics (see, Navarro et al. 2020), sequential order statistics,

generalized order statistics, time series and sequences of non-identically distributed random variables. These topics are left for future research projects.

Acknowledgements I would like to thank the anonymous reviewers for several helpful suggestions. JN was supported in part by Ministerio de Ciencia e Innovación of Spain under Grant PID2019-103971GB-I00/AEI/10.13039/501100011033.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Awad AM, Raqab MZ (2000) Prediction intervals for the future record values from exponential distribution: comparative study. *J Stat Comput Simul* 65:325–340. <https://doi.org/10.1080/00949650008812005>
- Basak P, Balakrishnan N (2003) Maximum likelihood prediction of future record statistic. In: Lindqvist B, Doksum K (eds) *Mathematical and statistical methods in reliability*, chap 11. World Scientific Publishing, Singapore, pp 159–175
- Burkschat M (2009) Linear estimators and predictors based on generalized order statistics from generalized Pareto distributions. *Commun Stat Theory Methods* 39:311–326. <https://doi.org/10.1080/03610920902746630>
- Dunsmore IR (1983) The future occurrence of records. *Ann Inst Stat Math* 35:267–277. <https://doi.org/10.1007/BF02480982>
- Gil A, Segura J, Temme NM (2007) *Numerical methods for special functions*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia
- Guo B, Zhu N, Wang W, Wang H (2020) Constructing exact tolerance intervals for the exponential distribution based on record values. *Qual Reliab Eng Int* 36:2398–2410
- Hürlimann W (2004) Distortion risk measures and economic capital. *N Am Actuarial J* 8:86–95
- Koenker R (2005) *Quantile regression*. Cambridge University Press, Cambridge
- Koenker R, Bassett G Jr (1978) Regression quantiles. *Econometrica* 46:33–50
- Navarro J (2020) Bivariate box plots based on quantile regression curves. *Dep Mod* 8:132–156. <https://doi.org/10.1515/demo-2020-0008>
- Navarro J, del Águila Y, Sordo MA, Suárez-Llorens A (2013) Stochastic ordering properties for systems with dependent identically distributed components. *Appl Stoch Mod Bus* 29:264–278. <https://doi.org/10.1002/asmb.1917>
- Navarro J, Calí C, Longobardi M, Durante F (2020) Distortion representations of multivariate distributions. Submitted
- Nelsen RB (2006) *An introduction to copulas*, 2nd edn. Springer, New York
- Nevzorov VB (2001) *Records: mathematical theory*. American Mathematical Society, Providence
- Paul J, Thomas PY (2016) On generalized upper (k) record values from Pareto distribution. *Aligarh J Stat* 36:63–78
- Raqab MZ (2007) Exponential distribution records: different methods of prediction. In: Ahsanullah M, Raqab MZ (eds) *Recent developments in ordered random variables*, chap 16. Nova Science Publishers, Hauppauge, pp 239–251
- Raqab MZ, Nagaraja HN (1995) On Some predictors of future order statistics. *Metron* LIII-N.1-2:185–204
- Raqab MZ, Ahmadi J, Doostparast M (2007) Statistical inference based on record data from Pareto model. *Statistics* 41:105–118

- Volovskiy G (2018) Likelihood-based prediction in models of ordered data. Ph.D. thesis, RWTH Aachen University
- Volovskiy G, Kamps U (2020) Maximum observed likelihood prediction of future record values. *Test* 29:1072–1097. <https://doi.org/10.1007/s11749-020-00701-7>
- Volovskiy G, Kamps U (2020) Maximum product of spacings prediction of future record values. *Metrika* 83:853–868. <https://doi.org/10.1007/s00184-020-00767-1>
- Wang S (1996) Premium calculation by transforming the layer premium density. *Astin Bull* 26:71–92
- Yaari ME (1987) The dual theory of choice under risk. *Econometrica* 55:95–115

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.