Studia Scientiarum Mathematicarum Hungarica 49 (4), 436–445 (2012)
DOI: 10.1556/SScMath.49.2012.4.1219

LINEARITY OF REGRESSIONS INSIDE TOP-k-LISTS AND RELATED CHARACTERIZATIONS

M. AHSANULLAH[1], G. G. HAMEDANI[2] and J. WESOŁOWSKI[3]

[1] Department of Management Sciences, Rider University, Lawrenceville, NJ 08648
e-mail: ahsan@rider.edu

[2] Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI 53201-1881
e-mail: g.hamedani@mu.edu

[3] Wydzial Matematyki i Nauk Informacyjnych, Politechnika Warszawska, Warszawa, Poland
e-mail: j.wesolowski@mini.pw.edu.pl

Communicated by E. Csáki

(Received July 24, 2010; accepted June 30, 2012)

Abstract

López-Blázquez and Wesołowski [6] introduced the top-k-lists sequence of random vectors and elaborated the usefulness of such data. They also developed the distribution of top-k-lists and their properties arising from various probability distributions, such as standard exponential distribution and uniform distribution on $(0, 1)$. In this paper, we study the linearity of regressions inside top-k-lists and then based on this study we present characterizations of certain distributions.

1. Introduction

First, we introduce notation and recall some basic results about order statistics (see e.g. David and Nagaraja [4]) and k^{th} records (see e.g. Ahsanullah [1] or Arnold et al. [2]). Given a list of k real numbers, $x_1, x_2, \ldots, x_k$, they can be arranged in an increasing order to obtain $x_{1:n} \leqq x_{2:n} \leqq \ldots \leqq x_{k:k}$. Then we define

$$\operatorname{ord}(x_1, x_2, \ldots, x_k) = (x_{1:k}, x_{2:k}, \ldots, x_{k:k}).$$

If F is a cumulative distribution function (*cdf*), the quantile function of F is

$$Q_F(u) = \inf\left\{ x : F(x) \geqq u \right\}, \quad \text{for} \quad u \in (0, 1].$$

2000 *Mathematics Subject Classification.* Primary 62E10, 62E15, 62E20, 60E05.

Key words and phrases. Top-k-lists, linearity of regression, characterizations.

Let $X_{i:k}$, $i = 1, 2, \ldots, k$ be order statistics of a sample $X_1, X_2, \ldots, X_k$ of independent and identically distributed (*iid*) random variables (*rv*'s) with an absolutely continuous *cdf* F and corresponding probability density function (*pdf*) f. Consider the k^{th} record times defined recurrently as

$$T_0^{(k)} = k,$$

$$T_{n+1}^{(k)} = \min\left\{j : j > T_n^{(k)} \text{ and } X_j > X_{T_n^{(k)}-k+1:\,T_n^{(k)}}\right\}, \quad n \geqq 0,$$

then the *rv* $R_n^{(k)} = X_{T_n^{(k)}-k+1:\,T_n^{(k)}}$, $n \geqq 0$ is the n^{th} k^{th} record from the sequence $(X_i)_{i\geqq 1}$. The *pdf* of $R_n^{(k)}$ is

$$f_{R_n^{(k)}}(x) = \frac{k^{n+1}}{n!}\left[-\ln\left(1-F(x)\right)\right]^n\left(1-F(x)\right)^{k-1} f(x), \tag{1.1}$$

for $x \in \left(Q_F(0^+), Q_F(1)\right)$.

A new concept of top-k-lists that was introduced by López-Blázquez and Wesołowski [6] is given in the following definition.

Definition 1.1. Let $(X_i)_{i\geqq 1}$ be a sequence of *iid rv*'s with a common *cdf* F. Then n^{th} top-k-list from $(X_i)_{i\geqq 1}$ is defined as follows:

$$L_n^{(k)} = (Y_{1,n}, Y_{2,n}, \ldots, Y_{k,n}), \quad n \geqq 0, \tag{1.2}$$

where $Y_{j,n} = X_{T_n^{(k)}-k+j:\,T_n^{(k)}}$, $j = 1, 2, \ldots, k$, $n = 0, 1, \ldots$.

Note that $Y_{1,n} = R_n^{(k)}$. The index i of the sequence $(X_i)_{i\geqq 1}$ in the Definition 1.1 can be viewed as a discrete time. The evolution of the list is as follows. At time k, the 0^{th} top-k-list $L_0^{(k)} = (X_{1:k}, X_{2:k}, \ldots, X_{k:k})$ is available. The list remains unaltered until time $T_1^{(k)}$. At this moment, the first element of the list $L_0^{(k)}$ is removed and the *rv* $X_{T_1^{(k)}}$ enters the list. Then $L_1^{(k)} = \operatorname{ord}\left(X_{T_1^{(k)}}, X_{2:k}, \ldots, X_{k:k}\right)$. From now on the process behaves in a similar way: an $(n-1)^{\text{th}}$ top-k-list $L_{n-1}^{(k)}$ remains unaltered until the n^{th} k^{th} record time $T_n^{(k)}$ occurs and then

$$L_n^{(k)} = \operatorname{ord}\left(X_{T_n^{(k)}}, Y_{2,n-1}, \ldots, Y_{k,n-1}\right), \quad n \geqq 1. \tag{1.3}$$

Note that the model of top-k-lists covers at least three important models for ordered statistical data:

- Order statistics as the 0^{th} top-k-list $L_0^{(k)}$.
- Records as the sequence $\big(L_n^{(1)}\big)$.
- k^{th} records as the sequence of the first components of $L_n^{(k)}$'s.

For an absolutely continuous *cdf* F with a *pdf* f, the *pdf* of the random vector $L_n^{(k)}$ was obtained by López-Blázquez and Wesołowski [6] in the context of the following theorem, which is given here for the sake of completeness.

Theorem 1.2 (Theorem 2 of López-Blázquez and Wesołowski [6]). *For any $w \in \big(Q_F(0^+), Q_F(1)\big)$, let $Z_1^{(w)}$, $Z_2^{(w)}, \dots, Z_k^{(w)}$ be iid rv's from the truncated pdf*

$$f_w(z) = \frac{f(z)}{1-F(w)}, \quad z \geqq w. \tag{1.4}$$

Then

(*a*) *For all $n \geqq 1$, the rv $X_{T_n^{(k)}}$ and the vector $\big(Y_{2,n-1}, Y_{3,n-1}, \dots, Y_{k,n-1}\big)$ are conditionally independent given $Y_{1,n-1}$. Moreover*

$$\big(X_{T_n^{(k)}} \mid Y_{1,n-1} = w\big) \stackrel{d}{=} Z_1^{(w)}. \tag{1.5}$$

(*b*) *For all $n \geqq 0$,*

$$\big(Y_{2,n}, Y_{3,n}, \dots, Y_{k,n} \mid Y_{1,n} = w\big) \stackrel{d}{=} \operatorname{ord}\big(Z_2^{(w)}, Z_3^{(w)}, \dots, Z_k^{(w)}\big). \tag{1.6}$$

(*c*) *The pdf of $L_n^{(k)}$ is*

$$f_{L_n^{(k)}}(y_1, y_2, \dots, y_k) = \frac{k^n k!}{n!}\big[-\ln\big(1-F(y_1)\big)\big]^n \prod_{j=1}^{k} f(y_j), \tag{1.7}$$

for $Q_F(0^+) < y_1 < y_2 < \dots < y_k < Q_F(1)$.

As we mentioned earlier, the marginal *pdf* of $Y_{1,n}$ which is the n^{th} k^{th} record is given by (1.1).

The following representations for conditional distribution, which are consequences of (1.6), will also be used in the next section:

$$(Y_{r,n} \mid Y_{1,n} = w) \stackrel{d}{=} Z_{r-1:k-1}^{(w)}. \tag{1.8}$$

Similarly, for any $1 < r < s \leqq k$ we have

$$\big((Y_{r,n}, Y_{s,n}) \mid Y_{1,n} = w\big) \stackrel{d}{=} \big(Z^{(w)}_{r-1:k-1}, Z^{(w)}_{s-1:k-1}\big). \tag{1.9}$$

Note that, due to the well-known conditional property of order statistics (see for instance Theorem 2.5 in David and Nagaraja [4]) the following further useful representations follow from (1.8) and (1.9) respectively

$$(Y_{r,n} \mid Y_{1,n}) \stackrel{d}{=} X_{r:k} \mid X_{1:k} \tag{1.10}$$

and

$$(Y_{r,n}, Y_{s,n}) \mid Y_{1,n} \stackrel{d}{=} (X_{r:k}, X_{s:k}) \mid X_{1:k}. \tag{1.11}$$

Throughout the paper when we talk about conditional moments, we tacitly assume that they do exist.

In Section 2 (below) we establish certain properties of the elements of top-k-lists. Then in Section 3 we study characterizations of certain distributions based on the results derived in Section 2.

2. Results

We start with the following proposition:

PROPOSITION 2.1. *(i) The marginal pdf of $Y_{r,n}$ for any $r \in \{2, 3, \ldots, k\}$ has the form*

$$f_{Y_{r,n}}(x) = \frac{k^n k! f(x)\big[1 - F(x)\big]^{k-r}}{n!(r-2)!(k-r)!} H_r(x), \tag{2.1}$$

where

$$H_r(x) = \int_{-\infty}^{u} \big[F(x) - F(t)\big]^{r-2}\big[-\ln\big(1 - F(t)\big)\big]^n f(t)\, dt.$$

(ii) The bivariate marginal pdf of $(Y_{r,n}, Y_{s,n})$ for any $r, s \in \{2, 3, \ldots, k\}$, $r < s$ has the form

$$f_{Y_{r,n},Y_{s,n}}(x, y) = \frac{k^n k! \big[F(y) - F(x)\big]^{s-r-1}\big[1 - F(y)\big]^{k-s}}{n!(r-2)!(s-r-1)!(k-s)!} \times f(x) f(y) H_r(x) I_{(-\infty,y)}(x), \tag{2.2}$$

and for any $r \in \{2, 3, \ldots, k\}$ *it is*

$$f_{Y_{1,n},Y_{r,n}}(x,y) = \frac{k^n k!\big[1-F(y)\big]^{k-r}\big[F(y)-F(x)\big]^{r-2}}{n!(r-2)!(k-r)!} \times \big[-\ln\big(1-F(x)\big)\big]^n f(x)f(y)I_{(-\infty,y)}(x), \tag{2.3}$$

where $I_{(-\infty,y)}(x)$ *is the indicator function.*

PROOF. (i) We rely on representation (1.10). Therefore, using the well-known formula for the marginal *pdf* of order statistics and (1.7) we obtain

$$\begin{aligned}
f_{Y_{r,n}}(x) &= \int_{-\infty}^{x} f_{X_{r:k}|X_{1:k}=w}(x|w) f_{Y_{1,n}}(w)\,dw \\
&= \int_{-\infty}^{x} \frac{(k-1)!}{(r-2)!(k-r)!} \frac{\big[F(x)-F(w)\big]^{r-2} f(x)\big[1-F(x)\big]^{k-r}}{\big[1-F(w)\big]^{k-1}} \\
&\quad \times \frac{k^{n+1}}{n!}\big[-\ln\big(1-F(w)\big)\big]^n\big[1-F(w)\big]^{k-1} f(w)\,dw \\
&= \frac{k^n k!}{n!(r-2)!(k-r)!} f(x)\big[1-F(x)\big]^{k-r} \int_{-\infty}^{x} \big[F(x)-F(w)\big]^{r-2} \\
&\quad \times \big[-\ln\big(1-F(w)\big)\big]^n f(w)\,dw,
\end{aligned}$$

which proves (2.1).

(ii) First we note that the integrand in the last formula is the *pdf* of $(Y_{1,n}, Y_{r,n})$, thus the formula (2.3) is proved.

Now we use the representation (1.9). Due to the formula for the *pdf* of a bivariate marginal of order statistics we obtain

$$f_{Y_{r,n},Y_{s,n}}(x,y) = \int_{-\infty}^{x} f_{Z^{(w)}_{r-1:k-1},Z^{(w)}_{s-1:k-1}}(x,y) f_{Y_{1,n}}(w)\,dw$$

$$= \int_{-\infty}^{x} \frac{(k-1)!}{(r-2)!(s-r-1)!(k-s)!}$$

$$\times \frac{\big[F(x)-F(w)\big]^{r-2} f(x)\big[F(y)-F(x)\big]^{s-r-1} f(y)\big[1-F(y)\big]^{k-s}}{\big[1-F(w)\big]^{k-1}}$$

$$\times \frac{k^{n+1}}{n!}\big[-\ln\big(1-F(w)\big)\big]^{n}\big[1-F(w)\big]^{k-1} f(w)\,dw.$$

Cancelling the term $\big[1-F(w)\big]^{k-1}$ we arrive at (2.2). □

Now we will establish the Markov property for elements of top-k-lists.

Proposition 2.2. *For any $n \geqq 1$ and for any $r, s \in \{1, 2, \dots, k\}$, such that $r < s$ the conditional distribution $Y_{s,n} \mid Y_{r,n}, Y_{r-1,n}, \dots, Y_{1,n}$ is the same as the conditional distribution $Y_{s,n} \mid Y_{r,n}$.*

Proof. It suffices to prove the result for $s = r+1$. Note that the conditional *pdf* of $Y_{r+1,n}$ given $Y_{r,n}$, $Y_{r-1,n}$, …, $Y_{1,n}$ has the form

$$f_{Y_{r+1,n}\mid Y_{r,n}=y_r,\dots,Y_{1,n}=y_1}(y_{r+1})$$

$$= \frac{f_{Y_{1,n},Y_{2,n},\dots,Y_{r+1,n}}(y_1, y_2, \dots, y_{r+1})}{f_{Y_{1,n},Y_{2,n},\dots,Y_{r,n}}(y_1, y_2, \dots, y_r)}.$$

Therefore through (1.7), upon cancellations we get that this is equal to

$$\frac{f(y_{r+1}) \int_{y_{r+1}<x_{r+2}<\dots<x_k} \prod_{j=r+2}^{k} f(x_j)\,dx_{r+2}\dots dx_k}{\int_{y_r<x_{r+1}<\dots<x_k} \prod_{j=r+1}^{k} f(x_j)\,dx_{r+1}\dots dx_k}.$$

Since the above expression is a function of y_r and y_{r+1} only, we conclude that it equals $f_{\,Y_{r+1,n}\mid Y_{r,n}=y_r}\,(y_{r+1})$ which completes the proof. □

Corollary 2.3. *The following representation is given for the conditional distribution $Y_{s,n} \mid Y_{r,n}$ for any $n \geqq 1$ and for any $r, s \in \{1, 2, \dots, k\}$, $r < s$*

$$Y_{s,n} \mid Y_{r,n} \overset{d}{=} X_{s:k} \mid X_{r:k}. \tag{2.4}$$

Proof. By the Markov property established in Proposition 2.2, we have

$$f_{Y_{s,n}\mid Y_{r,n}=y_r}(y_s) = f_{Y_{s,n}\mid Y_{r,n}=y_r,\,Y_{1,n}=y_1}(y_r)$$

$$= \frac{f_{Y_{1,n},Y_{r,n},Y_{s,n}}(y_1, y_r, y_s)}{f_{Y_{1,n},Y_{r,n}}(y_1, y_r)} = \frac{f_{Y_{r,n},Y_{s,n}\mid Y_{1,n}=y_1}(y_r, y_s)}{f_{Y_{r,n}\mid Y_{1,n}=y_1}(y_r)}.$$

Due to the representations (1.10) and (1.11), we have

$$\begin{aligned} f_{Y_{s,n}|Y_{r,n}=y_r}(y_s) &= \frac{f_{X_{r:k},X_{s:k}|X_{1:k}=y_1}(y_r, y_s)}{f_{X_{r:k}|X_{1:k}=y_1}(y_r)} \\ &= \frac{f_{X_{1:k},X_{r:k},X_{s:k}}(y_1, y_r, y_s)}{f_{X_{1:k},X_{r:k}}(y_1, y_r)} \\ &= f_{X_{s:k}|X_{r:k}=y_r,\ X_{1:k}=y_1}(y_s). \end{aligned}$$

Now the result follows through the Markov property for order statistics (see David and Nagaraja [4, page 17]). □

3. Characterizations

The first two characterizations presented below are based on the linearity of regression inside components of top-k-lists. The next two characterizations will be in terms of conditional distribution and conditional moments of spacings of the components of top-k-lists respectively. In the proofs below we will use known characterizations based on linearity of regressions for classical order statistics. We refer interested readers to Bieniek and Szynal [3] where these characterizations where extended to generalized order statistics.

PROPOSITION 3.1. *Assume that*

$$E(Y_{s,n} \mid Y_{r,n}) = aY_{r,n} + b \quad \textit{for} \ \ 1 \leqq r < s \leqq k. \tag{3.1}$$

Then only the following three cases are possible:

1. $a = 1$ *and* X_i *has an exponential distribution,*
2. $a > 1$ *and* X_i *has a Pareto distribution,*
3. $a < 1$ *and* X_i *has a power function distribution.*

PROOF. In view of Corollary 2.3, we see that for $r \geqq 1$ the condition of linearity (3.1) is equivalent to

$$E(X_{s:k} \mid X_{r:k}) = aX_{r:k} + b.$$

Now the result follows from Dembińska and Wesołowski [5]. □

REMARK 3.2. The analysis of regression condition in the opposite direction (i.e. $r > s$), in general case, seems much harder. Here we present the following special case.

PROPOSITION 3.3. *Assume that*

$$E(Y_{1,n} \mid Y_{2,n}) = aY_{2,n} + b. \tag{3.2}$$

Then the same three cases 1.–3. *are the only possibilities for the distribution of the* n^{th} *record* R_n *of the original sequence.*

PROOF. By Proposition 2.1 for $r = 1$ and $s = 2$ we obtain the following formula for the conditional *pdf* of $Y_{1,n}$ given $Y_{2,n}$

$$f_{Y_{1,n}|Y_{2,n}}(x|y) = \frac{\left[-\ln\left(1 - F(x)\right)\right]^n f(x)}{\int_{-\infty}^{y}\left[-\ln\left(1 - F(w)\right)\right]^n f(w)\,dw} I_{(-\infty,y]}(x). \tag{3.3}$$

Note that, the *pdf* g of the n^{th} record (putting $k = 1$ in (1.2)) has the form

$$g(x) = \frac{\left[-\ln\left(1 - F(x)\right)\right]^n f(x)}{n!}.$$

Therefore, using (3.3), the linearity of regression can be written as

$$\int_{-\infty}^{y} xg(x)\,dx = (ay + b)G(y),$$

where G is the *cdf* of R_n. Now, through the standard technique using differentiation we arrive at the desired result. □

REMARK 3.4. In the case of $r > 2$ the condition of linearity of regression

$$E(Y_{1,n} \mid Y_{r,n}) = aY_{r,n} + b$$

leads through (2.3) and (2.1) to the following integral equation

$$\int_{-\infty}^{y} x\left[F(y) - F(x)\right]^{r-2}\left[-\ln\left(1 - F(x)\right)\right]^n f(x)\,dx$$

$$= (ay + b)\int_{-\infty}^{y}\left[F(y) - F(x)\right]^{r-2} \times \left[-\ln\left(1 - F(x)\right)\right]^n f(x)\,dx.$$

The solution seems to be difficult even in the case of $r = 3$. In general, that is in the case of $1 < r < s < k$ the characterization of the parent law based on

$$E(Y_{r,n} \mid Y_{s,n}) = aY_{s,n} + b$$

seems to be very difficult.

PROPOSITION 3.5. *Let* $(X_i)_{i \geqq 1}$ *be a sequence of iid non-negative rv's with cdf* F *and* $F(0)=0$, $F(x)<1$ *for all* $x>0$. *If for a fixed* r, $1 \leqq r < k$, $Y_{r+1,n}-Y_{r,n}$ *and* $Y_{r,n}$ *are independent, then* $X_i \sim E(\lambda)$.

PROOF. The independence of $Y_{r+1,n}-Y_{r,n}$ and $Y_{r,n}$ implies

$$\left\{(k-r)\left[1-F(z+x)\right]^{k-r-1} f(z+x)\right\}\left[1-F(x)\right]^{-k+r} = C_z,$$

where C_z is independent of x.

Integrating the above expression with respect to z from z_0 to ∞, we obtain

$$\left[1-F(z_0+x)\right]^{k-r}\left[1-F(x)\right]^{-k+r} = b_{z_0},$$

where $b_{z_0} = \int_{z_0}^{\infty} C_z\, dz$. Letting $x \to 0$, we get $b_{z_0} = \left[1-F(z_0)\right]^{k-r}$. Thus

$$\left[1-F(z_0+x)\right]^{k-r} = \left[1-F(x)\right]^{k-r}\left[1-F(z_0)\right]^{k-r},$$

for all x, $z_0 \geqq 0$ and $k>r>0$. The solution of the last equation above is $1-F(x)=e^{-\lambda x}$, $x>0$ and some $\lambda>0$. $\square$

If F is *cdf* of a non-negative *rv*, we will call F "new better than used (NBU)" if

$$1-F(x+y) \leqq \left[1-F(x)\right]\left[1-F(y)\right], \quad x,y \geqq 0.$$

F is called "new worse than used (NWU)" if the above inequality is reversed. We say that $F \in C$ if F is either NBU or NWU.

PROPOSITION 3.6. *Let* $(X_i)_{i \geqq 1}$ *be a sequence of iid non-negative rv's with cdf* F *and* $F(0)=0$, $F(x)<1$ *for all* $x>0$. *If* $F \in C$ *and*

$$E\left[(Y_{r+1,n}-Y_{r,n})^m \mid Y_{r,n}=x\right] = b_m, \tag{3.4}$$

where b_m *is independent of* x, *then* X_i *has an exponential distribution.*

PROOF. From (3.4) we obtain

$$\int_0^{\infty} z^m (k-r)\left[\frac{\left(1-F(z+x)\right)}{1-F(x)}\right]^{k-r-1} \frac{f(z+x)}{1-F(x)}\, dz = b_m,$$

i.e.,

$$\int_0^{\infty} z^m (k-r)\left[1-F(z+x)\right]^{k-r-1} f(z+x)\, dz = b_m\left[1-F(x)\right]^{k-r},$$

and after simplification, we arrive at

$$\int_0^\infty mz^m \big[1 - F(z+x)\big]^{k-r} dz = b_m \big[1 - F(x)\big]^{k-r}.$$

Letting $x \to 0$ in the last equality we obtain

$$\int_0^\infty mz^{m-1} \big[1 - F(z)\big]^{k-r} dz = b_m.$$

Hence we can write

$$\int_0^\infty mz^{m-1} \Big[\big(1 - F(z+x)\big)^{k-r} - \big(1 - F(z)\big)^{k-r} \big(1 - F(x)\big)^{k-r}\Big] dz = 0. \tag{3.5}$$

Since $F \in C$, we must have

$$\big[1 - F(z+x)\big]^{k-r} - \big[1 - F(z)\big]^{k-r} \big[1 - F(x)\big]^{k-r} = 0, \tag{3.6}$$

for all $x > 0$ and almost all $z > 0$. The solution of (3.6) is $F(x) = 1 - e^{-\lambda x}$, $\lambda > 0$ and $x \geqq 0$.

REMARK 3.7. It can be shown that $b_m = \frac{1}{\Gamma(m-1)\left[(k-r)\lambda\right]^m}$.

Acknowledgements. The authors are grateful to a referee for the valuable suggestions which improved the presentation of the content of this work.

REFERENCES

[1] AHSANULLAH, M. *Record Values – Theory and Applications,* University Press of America, Lanham, Maryland, 2004.
[2] ARNOLD, B. C., BALAKRISHNAN, N. and NAGARAJA, H. N., *Records,* Wiley, New York, 1998.
[3] BIENIEK, M. and SZYNAL, D., Characterizations of distributions via linearity of regression of generalized order statistics. *Metrika,* **58** (2007), 259–271.
[4] DAVID, H. A. and NAGARAJA, H. N., *Order Statistics,* Wiley, New York, 2003.
[5] DEMBIŃSKA, A. and WESOŁOWSKI, J., Linearity of regression for non-adjacent order statistics, *Metrika,* **48** (1998), 215–222.
[6] LÓPEZ-BLÁZQUEZ, F. and WESOŁOWSKI, J., Top-k-lists, *Metrika,* **65** (2007), 69–82.