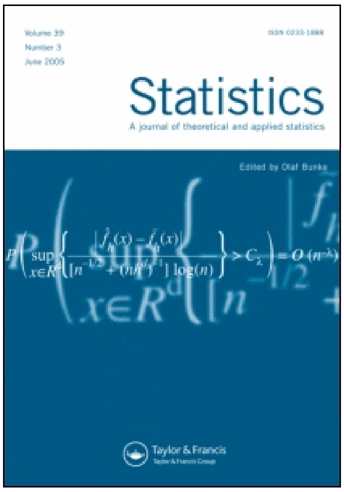

Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713682269

Linearity of regression for the past weak and ordinary records

Jacek Wesoowski [a]; Fernando López-Blázquez [b]
[a] Wydzia Matematyki i Nauk Informacyjnych, Politechnika Warszawska, Warszawa, Poland [b] Departamento de Estadistica e Investigación Operativa, Universidad de Sevilla, Sevilla, Spain

Online Publication Date: 01 December 2004

To link to this Article: DOI: 10.1080/02331880412331319288
URL: http://dx.doi.org/10.1080/02331880412331319288

Statistics, December 2004, Vol. 38(6), pp. 457–464

Taylor & Francis
Taylor & Francis Group

LINEARITY OF REGRESSION FOR THE PAST WEAK AND ORDINARY RECORDS

JACEK WESOŁOWSKI[a,*] and FERNANDO LÓPEZ-BLÁZQUEZ[b]

[a]*Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska, Warszawa, Poland;*
[b]*Departamento de Estadistica e Investigación Operativa, Universidad de Sevilla, Sevilla, Spain*

(Received 24 October 2000; Revised 16 May 2002; In final form 1 August 2004)

In this paper we study the property of linearity of backward regression for non-adjacent records. In the case of weak records, a characterization of the geometric distribution is obtained. It also appears that a related characterization for ordinary records does not hold, showing the difference in conditional behaviour between weak and ordinary records.

Keywords: Records; Weak records; Linearity of regression; Geometric distribution; Backward martingales

1 INTRODUCTION

Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed (iid) random variables (rvs) with a support of the form $\{0, 1, \ldots, N\}$, where $N \leq \infty$. Following Vervaat (1973), we introduce the concept of weak records. Firstly, we define weak record times by $V(1) = 1$ and $V(n) = \min\{j > V(n-1)\colon X_j \geq X_{V(n-1)}\}$, $n \geq 2$. Then, $W_n = X_{V(n)}$ is called the nth weak record, $n \geq 1$. The difference when comparing this notion with ordinary records, is in the inequality '$\geq$' instead of '$>$' in the definition of record times. This results in the observation that repeating a record is again a record, which seems to be more intuitive for iid sequences. What is more important for mathematical purposes is that the finite support is allowed, whereas for ordinary records the support is necessarily infinite. Let us note that in the continuous case the concepts of ordinary and weak records almost surely coincide.

The joint distribution of the first n weak records can be easily derived as

$$P(W_1 = k_1, \ldots, W_n = k_n) = p_{k_n} \prod_{r=1}^{n-1} \frac{p_{k_r}}{q_{k_r}}, \quad 0 \leq k_1 \leq \cdots \leq k_n \leq N, \tag{1}$$

where $p_k = P(X_1 = k)$ and $q_k = \sum_{j=k}^{N} p_j$, $k = 0, \ldots, N$. Hence, it immediately follows that the conditional distribution of two consecutive weak records in the forward direction, *i.e.*,

* Corresponding author. E-mail: wesolo@mini.pw.edu.pl

ISSN 0233-1888 print; ISSN 1029-4910 online
DOI: 10.1080/02331880412331319288

W_{n+1} given W_n, has a very simple form:

$$P(W_{n+1} = l \mid W_n = k) = P(W_2 = l \mid W_1 = k) = \frac{p_l}{q_k}, \quad 0 \le k \le l \le N.$$

This observation allowed us to characterize the family of distributions for which the regression of W_{n+1} on W_n is linear. The task was essentially accomplished in Stepanov (1994), with a slight complement in Wesołowski and Ahsanullah (2001). Non-linear regressions of W_{n+1} on W_n were considered in Aliev (1998). The case of linearity of regression of W_{n+2} on W_n was solved in Wesołowski and Ahsanullah (2001). The family of distributions of the parent sequence consists of geometric and two types of negative hypergeometric distributions.

Problems of that kind for the backward regression, *i.e.*, for the regression of W_m on W_n, for $m < n$, seem to be more difficult owing to the fact that the conditional distribution of W_m given W_n looks rather complicated. The first result in this direction was obtained by López-Blázquez and Wesołowski (2001) who characterized the family of distributions for which the regression of W_1 on W_2 is linear. One of the members of this class is the geometric distribution which covers the case $E(W_1 \mid W_2) = W_2/2$.

The main result of the present paper is a characterization of the geometric distribution by the condition

$$E(W_m \mid W_n) = \frac{m}{n} W_n,$$

for arbitrary fixed integers $1 \le m < n$. This is done in Section 3, whereas in Section 2 we derive the backward martingale property for the sequence of weak records when the parent distribution is geometric. It appears that in the case of ordinary records distributions other than geometric or geometric tail are allowed, as the discussion in Section 4 shows.

2 THE BACKWARD MARTINGALE PROPERTY

Assume that $X_1, X_2, \ldots,$ are iid geometric rvs, *i.e.*, $P(X_1 = k) = p(1-p)^k$, $k = 0, 1, \ldots,$ for some $p \in (0, 1)$. Then it follows easily that

$$P(W_n - W_{n-1} = i \mid W_{n-1} = j) = P(W_n = i + j \mid W_{n-1} = j) = \frac{p_{i+j}}{q_j} = p(1-p)^i,$$

for $i, j = 0, 1, \ldots$ Consequently, by the Markov property for the sequence $(W_n)_{n\ge1}$ (which follows immediately from the formula (1)), we conclude that

$$W_n = \sum_{i=1}^{n} Y_i, \tag{2}$$

where $(Y_i)_{i\ge1}$ is a sequence of iid rvs with the same geometric distribution as the original X_1. Hence,

$$E(W_m \mid W_n) = \sum_{i=1}^{m} E\left(Y_i \middle| \sum_{j=1}^{n} Y_j\right) = mE\left(Y_1 \middle| \sum_{j=1}^{n} Y_j\right) = \frac{m}{n} W_n,$$

for any $1 \le m < n$. Again, by the Markov property of the sequence $(W_n)_{n\ge1}$, it follows that $(W_n/n, \mathcal{F}_n)_{n\ge1}$, where $\mathcal{F}_n$ is the σ-field generated by $\{W_n, W_{n+1}, \ldots\}$, is a backward martingale. An immediate consequence of this fact is that $(W_n/n)_{n\ge1}$ converges a.s. Using the representation given in Eq. (2) and the strong law of large numbers, it can easily be checked that the a.s. limit is $(1-p)/p$.

Also, the representation (2) implies that W_n is a negative binomial rv with the probability mass function (pmf)

$$P(W_n = k) = \binom{n+k-1}{k} p^n (1-p)^k, \quad k = 0, 1, \ldots$$

3 CHARACTERIZATION OF THE GEOMETRIC DISTRIBUTION

In this section, we give the main result of the paper characterizing the geometric distribution by the linearity of backward regression for records that are not necessarily adjacent. This extends a result from our earlier paper, López-Blázquez and Wesołowski (2001), where the regression $E(W_1 \mid W_2)$ was considered. A slight drawback of the present approach is that we assume the slope in the regression of the exact form as derived in the previous section. This is the price we have to pay for considering non-adjacent records.

THEOREM 1 *Assume that*

$$E(W_m \mid W_n) = \frac{m}{n} W_n, \quad \text{a.s.}, \tag{3}$$

for fixed and arbitrary $1 \le m < n$. *Then, the parent distribution of* X_j*s is geometric.*

Proof Denote $c_j = p_j/q_j$, $j \in \{0, \ldots, N\}$. Observe that by Eq. (1) it follows that Eq. (3) is equivalent to

$$\sum_{0 \le i_1 \le \cdots \le i_{n-1} \le k} i_m c_{i_1} \cdots c_{i_{n-1}} = \frac{m}{n} k \sum_{0 \le i_1 \le \cdots \le i_{n-1} \le k} c_{i_1} \cdots c_{i_{n-1}}, \tag{4}$$

for any $k = 0, 1, \ldots, N$.

To show that the parent distribution is geometric it is equivalent to prove that $c_k = p$, for some $p \in (0, 1)$ and any $k = 0, 1, \ldots, N = \infty$. For that, apply induction with respect to k. Define $p = p_0$, then, just by the definition it follows that $c_0 = p$. Assume now that $p = c_0 = \cdots = c_{k-1}$. We will show that $c_k = p$.

Observe that by the induction assumption, it follows that Eq. (4) takes the form

$$\begin{aligned} p^{n-1} & \sum_{0 \le i_1 \le \cdots \le i_{n-1} \le k-1} i_m + c_k p^{n-2} \sum_{0 \le i_1 \le \cdots \le i_{n-2} \le k-1} i_m + \cdots + c_k^{n-m-1} p^m \sum_{0 \le i_1 \le \cdots \le i_m \le k-1} i_m \\ & + k c_k^{n-m} p^{m-1} \sum_{0 \le i_1 \le \cdots \le i_{m-1} \le k-1} 1 + \cdots + k c_k^{n-2} p \sum_{0 \le i_1 \le k-1} 1 + k c_k^{n-1} \\ = & \frac{m}{n} k \left(p^{n-1} \sum_{0 \le i_1 \le \cdots \le i_{n-1} \le k-1} 1 + \cdots + c_k^{n-2} p \sum_{0 \le i_1 \le k-1} 1 + c_k^{n-1} \right). \end{aligned} \tag{5}$$

Define now the quantities:

$$S_l(k) = \sum_{0 \le i_1 \le \cdots \le i_l \le k-1} i_1 = \binom{k-1+l}{l+1},$$

$$T_l(k) = \sum_{0 \le i_1 \le \cdots \le i_l \le k-1} 1 = \binom{k-1+l}{l},$$

for $l, k = 0, 1, 2, \ldots$

Observe also that for $l \geq m$,

$$\sum_{0 \leq i_1 \leq \cdots \leq i_l \leq k-1} i_m = -mS_l(k).$$

Consequently Eq. (5) can be rewritten as

$$\begin{aligned} k\frac{n-m}{n}&\{c_k^{n-1} + T_1(k)pc_k^{n-2} + \cdots + T_{m-1}(k)p^{m-1}c_k^{n-m}\} \\ &+ m\left\{\left(S_m(k) - \frac{k}{n}T_m(k)\right)p^m c_k^{n-m-1} + \cdots + \left(S_{n-2}(k) - \frac{k}{n}T_{n-2}(k)\right)\right. \\ &\left. \times\ p^{n-2}c_k + \left(S_{n-1}(k) - \frac{k}{n}T_{n-1}(k)\right)p^{n-1}\right\} = 0. \end{aligned}$$

Hence, it is enough to show that the only positive root of the polynomial

$$P(x) = a_0x^{n-1} + a_1px^{n-2} + \cdots + a_{n-1}p^{n-1}, \quad x \in \mathbb{R},$$

where

$$\begin{aligned} a_0 &= k\frac{n-m}{n}, \quad a_1 = k\frac{n-m}{n}T_1(k), \ldots, \quad a_{m-1} = k\frac{n-m}{n}T_{m-1}(k), \\ a_m &= m\left(S_m(k) - \frac{k}{n}T_m(k)\right), \ldots, \quad a_{n-1} = m\left(S_{n-1}(k) - \frac{k}{n}T_{n-1}(k)\right), \end{aligned}$$

is $x = p$.

Note that $P(x)$ can be rewritten in the form

$$P(x) = (x-p)(b_0x^{n-2} + b_1px^{n-3} + \cdots + b_{n-2}p^{n-2}) + b_{n-1}p^{n-1},$$

where $b_i = a_0 + \cdots + a_i$, $i = 0, 1, \ldots, n-1$. Now, to show that p is the only positive root, it suffices to check that $b_i > 0$, for all $i = 0, \ldots, n-2$ and $b_{n-1} = 0$. In fact, we will prove that

$$b_i = \begin{cases} k\dfrac{n-m}{n}T_i(k+1), & i = 0, \ldots, m-1 \\ mk\dbinom{k+i}{i}\left(\dfrac{1}{i+1} - \dfrac{1}{n}\right), & i = m, m+1, \ldots, n-1 \end{cases}$$

and such b_is obviously fulfils the previous requirements.

To this end, let us observe first that

$$T_0(k) + T_1(k) + \cdots + T_l(k) = T_l(k+1), \tag{6}$$

hence the formula for b_i, $i = 0, 1, \ldots, m-1$, follows immediately. Now let us consider any $i \in \{m, m+1, \ldots, n-1\}$. Then we get

$$\begin{aligned} b_i &= b_{m-1} + m\left\{\left(S_m(k) - \frac{k}{n}T_m(k)\right) + \cdots + \left(S_i(k) - \frac{k}{n}T_i(k)\right)\right\} \\ &= k\frac{n-m}{n}T_{m-1}(k+1) + m(S_m(k) + \cdots + S_i(k)) - \frac{mk}{n}(T_m(k) + \cdots + T_i(k)). \end{aligned}$$

Note that $S_l(k) = T_{l+1}(k-1)$. Then,

$$b_i = k\frac{n-m}{n}T_{m-1}(k+1) + m(T_{m+1}(k-1) + \cdots + T_{i+1}(k-1)) - \frac{mk}{n}(T_m(k) + \cdots + T_i(k)).$$

Applying Eq. (6), we obtain

$$\begin{aligned} b_i &= k\frac{n-m}{n}T_{m-1}(k+1) + m(T_{i+1}(k) - T_m(k)) - \frac{mk}{n}(T_i(k+1) - T_{m-1}(k+1)) \\ &= kT_{m-1}(k+1) - mT_m(k) + m\left(T_{i+1}(k) - \frac{k}{n}T_i(k+1)\right) \\ &= m\left(T_{i+1}(k) - \frac{k}{n}T_i(k+1)\right), \end{aligned}$$

and the final result follows by the definition of T_is.

So we proved that, if N is finite then

$$\frac{p_0}{q_0} = \frac{p_1}{q_1} = \cdots = \frac{p_N}{q_N} = p.$$

But for a finite N it follows that $p_N/q_N = 1$, which, due to the previous identity implies that the distribution of X_js is degenerated at zero. Otherwise the support of X_js is infinite and their common distribution is geometric. ■

4 REMARKS ON ORDINARY RECORDS

Given a sequence $X_1, X_2, \ldots,$ of iid rvs, let us denote by $(R_n)_{n\geq 1}$ the sequence of ordinary records.

For ordinary records linearity of regression of R_{n+1} on R_n was considered in Srivastava (1979) and Korwar (1984) – see also Arnold *et al.* (1998) for some comments. Related results involving constancy of regressions of $R_{n+2} - R_{n+1}$ or $(R_{n+1} - R_n)^2$ on R_n can be found in Nagaraja *et al.* (1989) and Balakrishnan and Balasubramanian (1995), respectively. Recently, it was shown in Dembińska and Wesołowski (2003) that the constancy of regression of $R_{n+2} - R_n$ on R_n characterizes the geometric-tail distribution. Here, we study linearity of regression for ordinary records in the opposite direction, *i.e.*, past given the present.

In the previous section, we have characterized the geometric distribution by the backward martingale property of weak records: $E(W_m/m \mid W_n) = W_n/n$, (a.s.), for fixed $1 \leq m < n$. It could be thought that if we use ordinary records rather than weak records we would obtain a similar characterization of the geometric-tail distributions. Such a duality was observed for some other results concerning regressional characterizations, see for instance remarks in Wesołowski and Ahsanullah (2001). In addition, in the case of ordinary records, it was shown in López-Blázquez and Wesołowski (2001) that linearity of regression of the first record on the second characterizes the parent distributions. A more general result for regressions of functions of the first record on the second has been obtained recently in Corrolary 9 of Franco and Ruiz (2001). However, it appears that the behaviour of the ordinary and weak records related to regressions with respect to the future is, in general, rather different. The aim of the present section is to show that a version of Eq. (3) for ordinary records does not characterize the geometric or geometric-tail distributions.

In particular, if $X_1 \sim Ge(p)$, $p \in (0, 1)$, it can be checked that the nth ordinary record admits a representation as

$$R_n = \sum_{i=1}^{n} Y_i + (n-1), \tag{7}$$

where $(Y_i)_{i\geq 1}$ is a sequence of iid rvs with the same geometric distribution as the original X_1. From Eq. (7), it is not difficult to conclude that $(Z_n, \mathcal{G}_n)_{n\geq 1}$, where $Z_n = [R_n - (n-1)]/n$, $n = 1, 2, \ldots$, is a backward martingale with a natural sequence of σ-fields $\mathcal{G}_n = \sigma(Z_n, Z_{n+1}, \ldots)$. Hence

$$E\left[\frac{R_m - (m-1)}{m} | R_n\right] = \frac{R_n - (n-1)}{n}, \quad \text{a.s.} \tag{8}$$

for $1 \leq m < n$. In particular, if $m = 2$ and $n = 3$, Eq. (8) gives

$$E(R_2 \mid R_3) = \frac{2}{3}R_3 - \frac{1}{3}, \quad \text{a.s.} \tag{9}$$

We will show that the condition (9) does not characterize the geometric (or geometric-tail) distribution.

Note that Eq. (9) is equivalent to

$$\sum_{0\leq i<j<k} j d_i d_j = \frac{2k-1}{3} \sum_{0\leq i<j<k} d_i d_j, \quad k = 2, 3, \ldots, \tag{10}$$

As

$$P(R_2 = j, R_3 = k) = p_k d_j \sum_{0\leq i<j} d_i, \quad 0 < j < k,$$

where $d_i = p_i/q_{i+1}$, $i = 0, 1, \ldots$

Observe that using the sequence $(d_i)_{i\geq 0}$ as defined earlier the pmf of X_1 can be written in the form

$$P(X_1 = 0) = \frac{d_0}{1+d_0}; \quad P(X_1 = k) = \frac{d_k}{1+d_k} \prod_{j=0}^{k-1} \frac{1}{1+d_j}, \quad k = 1, 2, \ldots \tag{11}$$

On the other hand it is known, see for instance López-Blázquez and Wesołowski (2001), that any infinite sequence of positive numbers $(d_i)_{i\geq 0}$ defines via Eq. (11) the pmf if only $\sum_{i=0}^{\infty} d_i = \infty$. Consequently to prove that there exist other distributions than geometric for which Eq. (9) holds it suffices to find appropriate solutions of Eq. (10). To this end, let us define $d_0 = (1 - \sqrt{1-4\alpha})/2$ for some $0 < \alpha \leq 1/4$ and $d_1 = 1 - d_0$. Then we will prove that

$$d_n = n\alpha(d_0 + \cdots + d_{n-1})^{-1}, \quad n \geq 1 \tag{12}$$

is a solution of Eq. (10) and generates the pmf according to Eq. (11).

Note that $d_1 = \alpha/d_0$, *i.e.*, Eq. (12) holds for $n = 1$. Now consider $k = 3$ in Eq. (10) to get

$$\alpha + 2d_2 = \frac{5}{3}(\alpha + d_2),$$

which implies immediately that $d_2 = 2\alpha/(d_0 + d_1) = 2\alpha$. Now assume that Eq. (12) holds for $n = 1, \ldots, l-1$. We will show that it holds also for $n = l$. Consider Eq. (10) for $k = l + 1$.

Then we get

$$\sum_{0\le i<j<l} j d_i d_j + l d_l \sum_{0\le i<l} d_i = \frac{2l+1}{3}\left(\sum_{0\le i<j<l} d_i d_j + d_l \sum_{0\le i<l} d_i\right).$$

Applying the earlier identity (10) for $k = l$ and then exploiting the induction assumption we get

$$\frac{l-1}{3} d_l \sum_{0\le i<l} d_i = \frac{2}{3}\sum_{j=1}^{l-1} d_j \sum_{i=0}^{j-1} d_i = \frac{2}{3}\alpha \sum_{j=1}^{l-1} j = \frac{2}{3}\alpha \frac{l(l-1)}{2},$$

which proves Eq. (12) for $n = l$.

Observe also that by Eq. (12) it follows that

$$d_n = \frac{n\alpha d_{n-1}}{(n-1)\alpha + d_{n-1}^2}, \quad n = 1, 2, \ldots \tag{13}$$

Hence $d_n < \sqrt{\alpha}$ for any $n > 2$ (just solve the respective quadratic inequality). Further, if $d_n < \sqrt{\alpha}$, $n > 2$, then Eq. (13) after some elementary algebra implies that the sequence $(d_n)_{n=3,4,\ldots}$ is increasing. Then its limit exists and equals $\sqrt{\alpha} > 0$ (just take the limits on both sides of Eq. (12)). Finally, we conclude that the series $\sum_{n=1}^{\infty} d_n$ diverges.

Note that for $\alpha < 1/4$ the pmfs obtained in such a way via Eq. (11) are neither geometric nor geometric tail, whereas for $\alpha = 1/4$ we obtain the geometric distribution with the parameter equal $1/2$.

Remark In Franco and Ruiz (2001), it was shown that in the case of discrete distributions the pair of regression functions

$$\phi_{n-1}(x) = E(h(R_{n-1})|R_n = x), \quad \phi_n(x) = E(h(R_n)|R_{n+1} = x),$$

where h is a real and strictly monotone function, characterizes the original distribution – see Theorem 8 of Franco and Ruiz (2001). In particular, linear ϕs lead to geometric-tail distributions – see Example 10 in that paper. However, as shown in the present section, for a single linearity of regression condition other distributions are also admissible.

Acknowledgement

The research of the second author was partially supported by MEC grant: BFM01-2378.

References

Aliev, F. A. (1998). Characterization of distributions through weak records. *Journal of Applied Statistical Science*, **8**, 13–16.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998). *Records*. Wiley, New York.

Balakrishnan, N. and Balasubramanian, N. (1995). A characterization of geometric distribution based on records. *Journal of Applied Statistical Science*, **2**, 277–282.

Dembińska, A. and Wesołowski, J. (2003). Constancy of regression for size-two record spacings. *Pakistan Journal of Statistics*, **19**, 143–149.

Franco, M. and Ruiz, J. M. (2001). Characterization of discrete distributions based on conditional expectations of record values. *Statistical Papers*, **42**, 101–110.

Korwar, R. M. (1984). On characterizing distributions for which the second record value has a linear regression on the first. *Sankhya, Series B*, **46**, 108–109.

López-Blázquez, F. and Wesołowski, J. (2001). Discrete distributions for which the regression of the first record on the second is linear. *Test*, **10**, 121–131.

Nagaraja, H. N., Sen, P. and Srivastava, R. C. (1989). Some characterizations of geometric tail distributions based on record values. *Statistical Papers*, **30**, 147–155.
Srivastava, R. C. (1979). Two characterizations of the geometric distribution by record values. *Sankhya, Series B*, **40**, 276–278.
Stepanov, A. V. (1994). A characterization theorem for weak records. *Theory of Probability and its Applications*, **38**, 762–764.
Vervaat, W. (1973). Limit theorems for records from discrete distributions. *Stochastic Processes and their Applications*, **1**, 317–334.
Wesołowski, J. and Ahsanullah, M. (2001). Linearity of regression for non-adjacent weak records. *Statistica Sinica*, **11**, 39–52.