Metrika (1998) 48: 215–222

Metrika
© Springer-Verlag 1999

Linearity of regression for non-adjacent order statistics

Anna Dembińska, Jacek Wesołowski*

Mathematical Institute, Warsaw University of Technology, Plac Politechniki 1, 00-661 Warsaw, Poland (e-mail: wesolo@alpha.im.pw.edu.pl)

Received February 1998

Abstract. Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution with the corresponding order statistics $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$. All the distributions for which $E(X_{k+r:n}|X_{k:n}) = aX_{k:n} + b$ are identified, which solves the problem stated in Ferguson (1967).

Key words: Order statistics, characterization, exponential distribution, Pareto distribution, power distribution, linearity of regression

1 Introduction

Extending some earlier works Ferguson (1967) proved the following theorem: Let $X_1, X_2, \ldots, X_n$ be a sample from a continuous distribution such that

$$E(X_{k+1:n}|X_{k:n}) = aX_{k:n} + b$$

for some $1 \leq k < n$,
then only the following three cases are possible:

1. $a = 1$ and X_1 has an exponential distribution,
2. $a > 1$ and X_1 has a Pareto distribution,
3. $a < 1$ and X_1 has a power distribution.

Let us point out that Ferguson states his result assuming that $E(X_{k:n}|X_{k+1:n}) = aX_{k+1} - b$ for some $1 \leq k < n$ instead of using the regression $X_{k+1:n}$ on $X_{k:n}$ and arrives at distributions dual to that given in 1–3. Since the duality is obvious (take $Y = -X$), without loosing generality, here we use the regression $X_{k+1:n}$ on $X_{k:n}$.

In Nagaraja (1988) an analogue of this result for discrete distributions was obtained.

Investigations of characterizations of probability distributions by properties of regression involving different functions of order statistics were lead by many researchers. The state of art up to the early nineties with suitable references can be found in the books of Arnold, Balakrishnan, Nagaraja (1992) or Johnson, Kotz, Balakrishnan (1994).

A slight refinement of the original Ferguson (1967) result, allowing discontinuity in one of the support ends has been given more recently in Pakes, Fakhry, Mahmoud and Ahmad (1996).

As pointed out in the monograph Arnold, Balakrishnan, Nagaraja (1992) the question raised by Ferguson (1967) ("it is unknown what new distributions arise if any") about analogous characterizations for non-adjacent order statistics has not been settled until the very recent paper by Wesołowski and Ahsanullah (1997) (it was pointed out by the referee that the result was shown earlier in the PhD thesis of Pudeg (1991)). They solved the problem considering linearity of regression of $X_{k+2:n}$ on $X_{k:n}$:

Let $X_1, X_2, \ldots, X_n$ be a sample from an absolutely continuous distribution such that

$$E(X_{k+2:n}|X_{k:n}) = aX_{k:n} + b$$

for some $1 \leq k < n-1$
then the same three cases 1.–3. are the only possible.

In a paper by the present authors, Dembińska and Wesołowski (1997), it was shown that in the absolutely continuous case, instead of a single regression condition, a pair of identities $E(X_{k_i+r:n_i}|X_{k_i:n_i}) = X_{k_i:n_i} + b_i, i = 1, 2$, with $r = 3$, $n_1 - k_1 \neq n_2 - k_2$ or with any r and $n_1 - k_1 = n_2 - k_2 + 1$, characterizes the exponential distribution.

In the present paper we identify the cases 1.–3. as all possible continuous distributions with the property of linearity of regression for any non-adjacent order statistics – which solves completely the problem raised by Ferguson (1967).

It should be pointed out that in López-Blázquez and Moreno-Rebollo (1997) this problem was considered under the additional assumption of r-differentiability of the distribution function of X's. The method used by these authors was based on solving r-th order differential equation and, as such, differs considerably from the approach, making use of integrated Cauchy functional equation, adopted in the present paper.

2 Linearity of regression

In this section we are interested in the conditional moment $E(X_{k+r:n}|X_{k:n})$, not only in the exponential case, but also for the power and Pareto distributions. Denote by $\mathscr{POW}(\theta; \mu, \nu)$ a power distribution defined by the density

$$f(x) = \frac{\theta(\nu - x)^{\theta-1}}{(\nu - \mu)^{\theta}} I_{(\mu,\nu)}(x),$$

where $\theta > 0$, $-\infty < \mu < \nu < \infty$ are some constants. By $\mathscr{PAR}(\theta; \mu, \delta)$ denote

the Pareto distribution with the pdf

$$f(x) = \frac{\theta(\mu+\delta)^{\theta}}{(x+\delta)^{\theta+1}} I_{(\mu,\infty)}(x),$$

where $\theta > 0$, and μ, δ are some real constants such that $\mu + \delta > 0$. Finally by $\mathscr{EXP}(\lambda, \gamma)$ denote the exponential distribution with the density

$$f(x) = \lambda \exp(-\lambda(x-\gamma)) I_{(\gamma,\infty)}(x),$$

where $\lambda > 0$ and γ are some real constants.

Observe that if X has the df F and the pdf f then for $[L]$ a.a. $x \in (l_F, r_F)$ (where $l_F = \inf\{x : F(x) > 0\}$, $r_F = \sup\{x : F(x) < 1\}$ and $[L]$ denotes the Lebesgue measure)

$$E(X_{k+r:n} | X_{k:n} = x)$$

$$= \frac{(n-k)!}{(r-1)!(n-k-r)!(\bar{F}(x))^{n-k}} \int_x^{r_X} y(\bar{F}(x) - \bar{F}(y))^{r-1} (\bar{F}(y))^{n-k-r} f(y)\, dy.$$

Consequently it can be easily verified that in all three cases of the exponential, power and Pareto distributions the regression relation, we are interested in, is linear, i.e.

$$E(X_{k+r:n} | X_{k:n}) = aX_{k:n} + b, \tag{1}$$

where the constants a and b have the following forms:

1. For the $\mathscr{POW}(\theta; \mu, v)$ distribution

$$a = \frac{\theta(n-k)!}{(n-k-r)!} \sum_{m=0}^{r-1} \frac{(-1)^m}{m!(r-1-m)![\theta(n-k-r+1+m)+1]}$$

$$b = v \frac{\theta(n-k)!}{(n-k-r)!}$$

$$\times \sum_{m=0}^{r-1} \frac{(-1)^m}{m!(r-1-m)!\theta(n-k-r+1+m)[\theta(n-k-r+1+m)+1]} \tag{2}$$

2. For the $\mathscr{PAR}(\theta; \mu, \delta)$ distribution with $\theta > \dfrac{1}{n-k-r+1}$

$$a = \frac{\theta(n-k)!}{(n-k-r)!} \sum_{m=0}^{r-1} \frac{(-1)^m}{m!(r-1-m)![\theta(n-k-r+1+m)-1]}$$

$$b = \delta \frac{\theta(n-k)!}{(n-k-r)!}$$

$$\times \sum_{m=0}^{r-1} \frac{(-1)^m}{m!(r-1-m)!\theta(n-k-r+1+m)[\theta(n-k-r+1+m)-1]} \tag{3}$$

3. For the $\mathscr{EXP}(\lambda, \gamma)$ distribution

$$a = 1, \quad b = \frac{(n-k)!}{\lambda(n-k-r)!} \sum_{m=0}^{r-1} \frac{(-1)^m}{m!(r-1-m)!(n-k-r+1+m)^2} \tag{4}$$

The question we address here is the following: are the given above examples the only for which linearity of regression (1) holds? The affirmative answer given beneath is the main result of the paper.

Theorem 1. *Assume that $X_1, \ldots, X_n$ are i.i.d. rv's with a common continuous df F. Let $E(|X_{k+r:n}|) < \infty$. If for some $k \leq n - r$ and some real a and b the linearity of regression (1) holds, then only the following three cases are possible:*
1. *$a = 1$ and F is a df of an exponential distribution;*
2. *$a > 1$ and F is a df of a Pareto distribution;*
3. *$a < 1$ and F is a df of a power distribution.*

Before we give the proof of the above result let us recall, following Rao and Shanbhag (1994), an important result concerning possible solutions of an extended version of the integrated Cauchy functional equation. This theorem will be used later on in the course of the proof of Theorem 1.

Theorem 2. *Consider the integral equation:*

$$\int_{\mathbf{R}_+} H(x+y)\mu(dy) = H(x) + c \quad \text{for } [L] \text{ a.a. } x \in \mathbf{R}_+,$$

where μ is a non-arithmetic σ-finite measure on $\mathbf{R}_+$ and $H : \mathbf{R}_+ \mapsto \mathbf{R}_+$ is a Borel measurable, either non-decreasing or non-increasing $[L]$ a.e. function that is locally $[L]$ integrable and is not identically equal zero $[L]$ a.e. Then $\exists \eta \in \mathbf{R}$ such that

$$\int_{\mathbf{R}_+} \exp(\eta x)\mu(dx) = 1,$$

and H has the form

$$H(x) = \begin{cases} \gamma + \alpha(1 - \exp(\eta x)) & \text{for } [L] \text{ a.a. } x \quad \text{if } \eta \neq 0 \\ \gamma + \beta x & \text{for } [L] \text{ a.a. } x \quad \text{if } \eta = 0 \end{cases}$$

where α, β, γ are some constants. If $c = 0$ then $\gamma = -\alpha$ and $\beta = 0$.

Now we are ready to prove our main result.

Proof of Theorem 1: Using the formulas for the joint distribution of $(X_{i:n}, X_{j:n})$ and the distribution of $X_{i:n}$ (see for instance the monograph of Arnold, Balakrishnan, Nagaraja (1992)) we can write:

$$E(X_{k+r:n}|X_{k:n}=x)$$

$$=\frac{(n-k)!}{(r-1)!(n-k-r)!}\int_x^\infty y\frac{[\bar{F}(x)-\bar{F}(y)]^{r-1}\bar{F}^{n-k-r}(y)}{\bar{F}^{n-k}(x)}d[-\bar{F}(y)]$$

F a.e., where $\bar{F}=1-F$. From (1) we get:

$$\frac{(n-k)!}{(r-1)!(n-k-r)!}\int_x^{r_F} y\left[\frac{\bar{F}(x)-\bar{F}(y)}{\bar{F}(x)}\right]^{r-1}\left[\frac{\bar{F}(y)}{\bar{F}(x)}\right]^{n-k-r}d\left[-\frac{\bar{F}(y)}{\bar{F}(x)}\right]$$

$$=ax+b \tag{5}$$

for F-almost all x's. Notice, following the reasoning of Ferguson (1967), that there does not exist an interval (c,d), $l_F<c<d<r_F$, over which F is constant since the right hand side of (5) is increasing in such an interval and the left hand side remains constant, while both sides are continuous, so that they could not possibly be equal at the next point of increase of F. (Observe that a has to be positive, which follows easily, for instance, from the next identity). Thus (l_F,r_F) is the support of the distribution defined by F and F is strictly increasing in this interval. Notice also that since both sides of (5) are continuous with respect to x we can assume that it holds for any $x\in(l_F,r_F)$.

Substituting $t=\bar{F}(y)/\bar{F}(x)$, i.e. $y=\bar{F}^{-1}(t\bar{F}(x))$ (observe that $\bar{F}^{-1}$ exists because $\bar{F}$ is strictly decreasing in (l_F,r_F)) into equation (5) we get:

$$\frac{(n-k)!}{(r-1)!(n-k-r)!}\int_0^1 \bar{F}^{-1}(t\bar{F}(x))(1-t)^{r-1}t^{n-k-r}\,dt=ax+b$$

Now substitute $\bar{F}(x)=w$, hence $x=\bar{F}^{-1}(w)$ and thus

$$\frac{(n-k)!}{(r-1)!(n-k-r)!}\int_0^1 \bar{F}^{-1}(tw)(1-t)^{r-1}t^{n-k-r}\,dt$$

$$=a\bar{F}^{-1}(w)+b,\quad w\in(0,1).$$

Divide both sides by a and substitute once again $t=e^{-u}$ and $w=e^{-v}$. Then

$$\frac{(n-k)!}{a(r-1)!(n-k-r)!}\int_0^\infty \bar{F}^{-1}(e^{-(u+v)})(1-e^{-u})^{r-1}e^{-(n-k-r)u}e^{-u}\,du$$

$$=\bar{F}^{-1}(e^{-v})+\frac{b}{a}$$

for any $v>0$.

Now let $G(v)=\bar{F}^{-1}(e^{-v})$. Consequently

$$\int_{\mathbf{R}_+} G(v+u)\mu(du)=G(v)+\frac{b}{a},\quad v>0$$

where μ is a finite measure on $\mathbf{R}_+$, which is absolutely continuous with respect

to the $[L]$ measure and is defined by

$$\mu(du) = \frac{(n-k)!}{a(r-1)!(n-k-r)!}(1-e^{-u})^{r-1}e^{-(n-k-r+1)u}\,du.$$

Observe that G is strictly increasing on $[0,\infty)$ since it is a composition of two strictly decreasing functions. Consequently the assumptions of Theorem 2 are fulfilled. Hence, since G is continuous, it follows that

$$G(v) = \begin{cases} \gamma + \alpha(1-\exp(\eta v)) & \text{if } \eta \neq 0 \\ \gamma + \beta v & \text{if } \eta = 0 \end{cases} \tag{6}$$

$v > 0$, where α, β, γ, η are some constants and

$$\int_{\mathbf{R}_+} \exp(\eta x)\mu(dx) = 1 \tag{7}$$

From (7) we get:

$$1 = \frac{(n-k)!}{a(r-1)!(n-k-r)!}\int_0^\infty e^{\eta x}(1-e^{-x})^{r-1}e^{-(n-k-r)x}e^{-x}\,dx$$

After substituting $t = e^{-x}$ we obtain (observe that $\eta < n-k-r+1$):

$$\begin{aligned} 1 &= \frac{(n-k)!}{a(r-1)!(n-k-r)!}\int_0^1 (1-t)^{r-1}t^{(n-k-r-\eta)}\,dt \\ &= \frac{1}{a}\,\frac{B(n-k-r-\eta+1,r)}{B(n-k-r+1,r)} \end{aligned}$$

where $B(.,.)$ is the complete beta function defined by

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,dt, \quad p,q>0.$$

Since $B(p,q) = \dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ then

$$\frac{\Gamma(n-k-r-\eta+1)\Gamma(r)}{\Gamma(n-k-\eta+1)}\,\frac{\Gamma(n-k+1)}{\Gamma(n-k-r+1)\Gamma(r)} = a \tag{8}$$

A slight rearrangement allows to rewrite (8) as

$$a = \frac{n-k}{n-k-\eta}\cdot\frac{n-k-1}{n-k-1-\eta}\cdot\ldots\cdot\frac{n-k-r+1}{n-k-r+1-\eta} = h(\eta), \tag{9}$$

say.

Observe that

1. $a < 1$ if $\eta < 0$,
2. $a > 1$ if $0 < \eta < n - k - r + 1$,
3. $a = 1$ if $\eta = 0$.

Moreover there is a unique η that fulfils (9), because the function h is strictly increasing.

Returning to (6), for a non-zero η, we can write

$$\bar{F}^{-1}(e^{-v}) = G(v) = \gamma + \alpha(1 - e^{\eta v})$$

which implies

$$e^{-v} = \bar{F}(\gamma + \alpha(1 - e^{\eta v})).$$

Let us substitute $z = \gamma + \alpha(1 - e^{\eta v})$. Then

$$e^{-v} = \left(1 - \frac{z - \gamma}{\alpha}\right)^{-(1/\eta)}$$

Hence $\bar{F}(z) = \dfrac{1}{\left(1 - \dfrac{z - \gamma}{\alpha}\right)^{1/\eta}}$ for $z > \gamma$.

Consider now three possible cases:

1. If $a < 1$ and $\eta < 0$ then

$$\bar{F}(z) = \left(\frac{\alpha + \gamma - z}{\alpha}\right)^{-(1/\eta)} = \left(\frac{\alpha + \gamma - z}{\alpha + \gamma - \gamma}\right)^{-(1/\eta)} = \left(\frac{v - z}{v - \mu}\right)^{\theta}$$

for $z \in (\mu, v)$, where $v = \alpha + \gamma$, $\mu = \gamma$, $\theta = -\dfrac{1}{\eta} > 0$. Observe that α has to be positive.

Thus $X_1 \sim \mathcal{POW}(\theta, \mu, v)$, where:

- $\theta = -\dfrac{1}{\eta}$ and η fulfils (9),
- v can be calculated from (2) with $\theta = -\dfrac{1}{\eta}$,
- $\mu < v$ is a real number.

2. If $a > 1$ and $\eta > 0$ then

$$\bar{F}(z) = \left(\frac{-\alpha}{z - \alpha - \gamma}\right)^{1/\eta} = \left(\frac{\gamma + (-\alpha - \gamma)}{z + (-\alpha - \gamma)}\right)^{1/\eta} = \left(\frac{\mu + \delta}{z + \delta}\right)^{\theta}$$

for $z > \mu$, where $\delta = -\alpha - \gamma$, $\mu = \gamma$, $\theta = \dfrac{1}{\eta} > 0$. Observe that α has to be negative.

Thus $X_1 \sim \mathscr{PAR}(\theta, \mu, \delta)$, where:

- $\theta = \dfrac{1}{\eta}$ and η fulfils (9),
- δ can be calculated from (3) with $\theta = \dfrac{1}{\eta}$,
- μ is a real number.

Observe that this is the only case in which $b = 0$ is allowed. Then $\delta = 0$ and $\mu > 0$.

3. If $a = 1$ and $\eta = 0$ then from (6) we get:

$$\bar{F}^{-1}(e^{-v}) = G(v) = \gamma + \beta v$$

$$e^{-v} = \bar{F}(\gamma + \beta v).$$

Let us substitute $z = \gamma + \beta v$. Then $\beta > 0$ and

$$\bar{F}(z) = e^{-(z-\gamma)/\beta} = e^{-\lambda(z-\gamma)}$$

for $z > \gamma$, where $\lambda = \dfrac{1}{\beta} > 0$.

Hence $X_1 \sim \mathscr{EXP}(\lambda, \gamma)$, where

- λ can be calculated from (4),
- γ is a real number.

Acknowledgement: The authors are greatly indebted to the referee for valuable remarks and first of all for turning their attention to some new references.

References

Arnold BC, Balakrishnan N, Nagaraja HN (1992) A first course in order statistics. Wiley, New York

Dembińska A, Wesołowski J (1997) On characterizing the exponential distribution by linearity of regression for non-adjacent order statistics. Demonstratio Mathematica 30:945–952

Ferguson TS (1967) On characterizing distributions by properties of order statistics. Sankhyā 29A:265–278

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, Vol. 1. Wiley, New York

López-Blázquez F, Moreno-Rebollo JL (1997) A characterization of distributions based on linear regression of order statistics and record values. Sankhyā 59A:311–323

Nagaraja HN (1988) Some characterizations of discrete distributions based on linear regressions of adjacent order statistics. Journal of Statistical Planning and Inference 20:65–75

Pakes AG, Fakhry ME, Mahmoud MR, Ahmad AA (1996) Characterizations by regressions of order statistics. Journal of Applied Statistical Science 3(1):11–23

Pudeg A (1991) Characterization of probability distributions via distributional properties of order statistics and record values. PhD Dissertation, Aachen University of Technology, Germany (in German)

Rao CR, Shanbhag DN (1994) Choquet-Deny type functional equations with applications to stochastic models. Wiley, New York

Wesołowski J, Ahsanullah M (1997) On characterizing distributions via linearity of regression for order statistics. Australian Journal of Statistics 39:69–78