Journal of Econometrics 24 (1984) 397–403. North-Holland

LINEAR REGRESSION AND THE YULE DISTRIBUTION

Evdokia XEKALAKI

University of Missouri, Columbia, MO 65201, USA

Received March 1982, final version received August 1983

The Yule distribution is shown to have certain interesting properties in the area of regression analysis. In particular, it is shown that under certain conditions, a random variable Z will have linear regressions on another random variable X and on its observable part Y only when X has a Yule distribution. More generally, the regression on the observed part Y will be constant for a finite number of values of Y, say k, and linear otherwise, only when X has a Yule distribution with its first k frequencies truncated.

1. Introduction

Let X, Z be two non-negative random variables (r.v.'s) such that $\mathrm{E}(Z|X=x) = \alpha + \beta x$, where α and β are real constants. Regressions of this type can be of great importance to research workers in econometrics, especially when X is not observable. For example, in the area of demand–supply analysis for a commodity we can only observe the demand Y that can be met in the market instead of the actual demand X. Also, in income distribution analysis, reported income, say Y, is what is actually observed and not the true income X.

In such cases, one would naturally ask: How does the regression of Z on the observable part Y of an economic r.v. X ($\mathrm{E}(Z|Y=y)$) relate to the regression of Z on the original r.v. X? If it is reasonable to assume that Z has a linear regression on X, what are the conditions under which the regression on the observable part Y remains linear?

Krishnaji (1970b, p. 1), in the context of income underreporting, showed that if $Y = RX$ with the r.v. R distributed independently of X according to a Beta distribution, then $\mathrm{E}(Z|Y=y)$ is also linear if and only if the distribution of X is of the Pareto type. That is, among the continuous income distributions the Pareto is the only one that preserves the linearity of regression.

It is of interest to remark that the Pareto distribution can be considered to be an approximation of a more general distribution called the Yule distribu-

tion with probability function (p.f.)

$$p_x = \rho x!/(\rho+2)_{(x)}, \qquad \rho > 0, \quad x = 1,2,\ldots, \tag{1}$$

where $a_{(b)} = \Gamma(a+b)/\Gamma(a)$, and a, b are positive real numbers [Kendall (1961)]. Moreover, the Yule distribution can be considered as the discrete analogue of the Pareto distribution [Irwin (1975)]. The implication of this is, of course, that whenever continuous data on various economic quantities (thought of as described by the Pareto distribution) become discrete through grouping or due to the fact that we can only measure their quantities with a limited accuracy, the Yule distribution can be used instead of the Pareto distribution. Indeed, the Yule distribution has been shown to arise often in the context of economic problems, e.g. as the distribution of income [Simon (1955)] or as a demand distribution [Xekalaki (1983)]. It would, therefore, be interesting to examine whether a property analogous to the one studied by Krishnaji for the Pareto distribution in the continuous case holds for the Yule distribution in the discrete case. This is done in the sequel.

2. Linear regression and the Yule distribution

In this section we will be concerned with integer-valued r.v.'s X, Y, where Y is the observable part of X ($X > 0$). Specifically, we will assume that

$$Y = [RX], \tag{2}$$

where R is an r.v. independent of X and distributed in the interval $(0,1)$ with $[a]$ denoting the integral part of a. The problem to be studied, then, would be the effect of a Yule distribution on the regression $\mathrm{E}(Z|Y=y)$ of any r.v. Z (independent of R) on Y in cases where we can assume that the regression $\mathrm{E}(Z|X=x)$ is linear.

We first need to prove the following theorem.

Theorem 1. Let X be an r.v. on $\{k, k+1,\ldots\}$, $k > 0$, and let its p.f. be $p_{x'} - x \geq k$. Then,

$$P(X > r) = a(r+1)p_r, \qquad r = k, k+1,\ldots, \quad a > 0, \tag{3}$$

if and only if X has a $(k-1)$-truncated Yule distribution with parameter $1/a$, i.e., if and only if

$$p_r = (1/a)\big((k+1)_{(r-k)}/(1/a+k+1)_{(r-k+1)}\big), \qquad r \geq k, \tag{4}$$

Proof.

Necessity. Let X have a $(k-1)$-truncated Yule distribution with parameter ρ and p.f. given by (4). Then,

$$P(X>r)$$
$$=\rho \sum_{x=r+1}^{\infty} (k+1)_{(x-k)}/(\rho+k+1)_{(x-k+1)}$$
$$=\rho\big((k+1)_{(r-k+1)}/(\rho+k+1)_{(r-k+2)}\big)\sum_{x=0}^{\infty}(r+2)_{(x)}/(\rho+r+3)_{(x)}$$
$$=p_r((r+1)/(\rho+r+2))(\Gamma(\rho+r+3)\Gamma(\rho)/\Gamma(\rho+1)\Gamma(\rho+r+2))$$
$$=((r+1)/\rho)\,p_r.$$

Hence necessity has been established.

Sufficiency. Observe that (3) implies that

$$p_{r+1}=a\big[(r+1)p_r-(r+2)p_{r+1}\big], \qquad r\geq k.$$

This is equivalent to

$$p_{r+1}-((r+1)/(1/a+r+2))\,p_r=0, \qquad r\geq k,$$

whose the unique solution under the condition $\sum_{r=k}^{\infty}p_r=1$ is given by (4). This completes the proof of the theorem. Q.E.D.

We can now prove the main result.

Theorem 2. Let X be a positive integer-valued r.v. and consider another r.v. Z with distribution function $F(z)$, $z\in A\subseteq \boldsymbol{R}$ and such that

$$\mathrm{E}(Z|X=x)=\alpha+\beta x, \tag{5}$$

for some constants α and β, $\beta\neq 0$. Let Y be a non-negative and integer-valued r.v. related to X as in (2) with R independent of X and Z and uniformly distributed in $(0,1)$. Then, for $k>0$,

$$\mathrm{E}(Z|Y=y)=\gamma+(\gamma-\alpha)(k-1) \quad \textit{for} \quad y\leq k-1,$$
$$=\gamma+(\gamma-\alpha)y \quad \textit{for} \quad y\geq k, \tag{6}$$

if and only if X has a $(k-1)$-truncated Yule distribution with parameter $\beta/(\gamma-\alpha-\beta)$.

Proof.

Necessity. Let p_x, q_y denote the p.f.'s of X and Y, respectively. From the definition of Y it follows that

$$\mathrm{E}(Z|Y=y)$$
$$=\frac{1}{q_y}\int_A z \sum_{x=y+1}^{\infty} [\text{illegible}]\, P\left(\frac{y}{x}\le R<\frac{y+1}{x}\right) p_x \,\mathrm{d}F(z),$$

i.e.,

$$\mathrm{E}(Z|Y=y)=\frac{1}{q_y}\sum_{x=y+1}^{\infty}\mathrm{E}(Z|X=x)\frac{p_x}{x}. \tag{7}$$

But [see Krishnaji (1970a)]

$$q_y=\sum_{x=y+1}^{\infty}\frac{p_x}{x}. \tag{8}$$

Then, substituting for q_y and $\mathrm{E}(Z|X=x)$ from (8) and (5), respectively, we obtain

$$\mathrm{E}(Z|Y=y)=\alpha+\beta\left(\left(\sum_{x=y+1}^{\infty}p_x\right)\Big/\left(\sum_{x=y+1}^{\infty}\frac{p_x}{x}\right)\right). \tag{9}$$

Assume now that $X\sim(k-1)$-truncated Yule distribution with some parameter $\rho>0$. Xekalaki (1983) has shown that if X, Y are defined as in the statement of this theorem, then $X\sim(k-1)$-truncated Yule (ρ) distribution if and only if

$$\sum_{x=y+1}^{\infty}\frac{p_x}{x}=\frac{1}{1+\rho}p_y, \qquad y=k,k+1,\ldots. \tag{10}$$

Therefore, it follows from (10) and Theorem 1 that

$$\mathrm{E}(Z|Y=y)=\alpha+(\beta(\rho+1)/\rho)k, \qquad y\le k-1,$$
$$=\alpha+\beta(\rho+1)/\rho+(\beta(\rho+1)/\rho)y, \qquad y\ge k.$$

Then $\mathrm{E}(Z|Y=y)$ is of the form (6) and hence necessity has been established.

Sufficiency. Suppose that (5) and (6) are true. Then, using (9) we obtain

$$-\delta k\sum_{x=y+1}^{\infty}\frac{p_x}{x}+\beta\sum_{x=y+1}^{\infty}p_x=0 \quad \text{for} \quad y\le k-1, \tag{11}$$

and

$$\sum_{x=y+1}^{\infty}p_x-\frac{\delta}{\beta}(y+1)\sum_{x=y+1}^{\infty}\frac{p_x}{x}=0 \quad \text{for} \quad y\ge k, \tag{12}$$

where $\delta = \gamma - \alpha$. Specializing (11) for $y = r$ and $y = r - 1$ and subtracting the resulting equations, we obtain

$$p_r(-\delta k/r + \beta) = 0,$$

which cannot hold for $r = 1, 2, \ldots, k-1$ unless

$$p_r = 0, \qquad r = 1, 2, \ldots, k-1. \tag{13}$$

Hence, p_r is truncated at the point $k-1$. Applying the same technique to eq. (12) twice we obtain

$$(\beta - \delta)(p_r - p_{r+1}) + \delta((p_{r+1})/(r+1)) = 0,$$

or, equivalently,

$$p_{r+1} - ((r+1)/(\delta/(\delta - \beta) + r + 1))p_r = 0, \qquad r = k, k+1, \ldots. \tag{14}$$

The solution to this equation is of the form

$$p_r = p_k\big((k+1)_{(r-k)}/(\beta/(\delta - \beta) + k + 2)_{(r-k)}\big), \quad r = k, k+1, \ldots.$$

Because of (13), the constant p_k can be determined from the condition $\sum_{x=k}^{\infty} p_x = 1$. It can be checked that $p_k = \beta/(\delta + k(\delta - \beta))$. Recalling that $\delta = \gamma - \alpha$, it follows that the unique solution to (14) is the $(k-1)$-truncated Yule distribution with parameter $\beta/(\gamma - \alpha - \beta)$. [The positivity of $\beta/(\gamma - \alpha - \beta)$ is ensured by the fact that putting $y = 0$ in (9) yields $(\gamma - \alpha)/\beta = q_0^{-1} > 1$.]

In fact, what this theorem says is that, under (2) and provided that the regression $\mathrm{E}(Z|X = x)$ is linear, $\mathrm{E}(Z|Y = y)$ is also linear only when X has a $(k-1)$-truncated Yule distribution ($k > 0$). This is an interesting result that determines uniquely the distribution of X in the general case where only frequencies beyond some point $k-1$ are known (head truncation). Such cases arise very often in connection with applications (e.g. low incomes are not declared, hence their frequencies are unknown).

An interesting special case arises when $k = 1$ [X has a Yule distribution as given by (1)]. Then, by Theorem 2, the assumption of a Yule distribution for X implies that the linearity of regression of any r.v. Z on X is preserved under the transition from X to Y. Conversely, the form invariance of the regression is a sufficient condition for X to be a Yule r.v., i.e., if information on $\mathrm{E}(Z|Y = y)$ is accessible we can, under (2), infer about the distribution of X even in cases where X is not observable.

The implications of such a result in practice can be very important. The practical value of Theorem 2 is enhanced if it is combined with a theorem shown by Xekalaki (1983). This theorem (specialized for $k=1$) states that, under (2), X has a Yule distribution on $\{1,2,\dots\}$ with p.f. as given by (1) if and only if Y has a Yule distribution on $\{0,1,2,\dots\}$ with p.f.

$$q_r = \rho/(\rho+1), \qquad r=0,$$

$$= (1/(\rho+1))\big(\rho r!/(\rho+2)_{(r)}\big), \qquad r=1,2,\dots.$$

Then Theorem 2 in combination with the above mentioned result leads to the following corollary.

Corollary. Let X, Y, Z be defined as in Theorem 2. Then,

$$\mathrm{E}(Z|Y=y) = \gamma + (\gamma-\alpha)y, \qquad y=0,1,2,\dots,$$

if and only if Y has a Yule distribution on $\{0,1,2,\dots\}$.

The importance of the latter result in practice lies in that for a Yule population not much information is lost by the fact that we may observe only part of X (i.e., Y). Any r.v. Z that is presumed to regress linearly on X will, under (2), regress linearly on its observable part (Y), too.

As an example, consider a family expenditure survey problem where the relationship

$$\mathrm{E}(Z|X=x) = \alpha + \beta x, \qquad \beta \neq 0,$$

between the family income X and the family consumption expenditure Z is desired to be estimated. People tend to underreport their income. Hence data on the actual family income X may not be available. Let Y be the reported income (observable part of X). It is not unreasonable to assume that underreporting is effected through (2). Also, let us follow Simon (1955) in assuming a Yule distribution on $\{0,1,2,\dots\}$ with parameter ρ for the reported income Y. Then, by the corollary to Theorem 2 one can arrive at estimates of α and β. Specifically, suppose that on the available data it is found that the distribution of Y is adequately described by the hypothesized Yule distribution. Then, from the corollary to Theorem 2 we know that the regression of Z on Y will be of the form

$$\mathrm{E}(Z|Y=y) = \gamma + \delta y, \qquad y=0,1,2,\dots,$$

and from the available data we can estimate γ and δ. Then, from the fact that

$\gamma = \alpha + \beta(\rho + 1)/\rho$ and $\delta = \beta(\rho + 1)/\rho$ we can estimate the coefficients α and β of the relationship between consumption expenditure and true income.

Moreover, if the Yule distribution on $\{0, 1, 2, \ldots\}$ provides a satisfactory fit to the observed distribution of Y, then by Xekalaki's (1983) result we are in a position to know that the unobserved distribution of X is the Yule on $\{1, 2, \ldots\}$ with the same parameter as that of the observed distribution of Y. That is, although we have no observations on the r.v. X, we can recover full information about the structural form of its distribution.

References

Irwin, J.O., 1975, The generalized Waring distribution, Journal of the Royal Statistical Society A 138, 18–31 (Part I), 204–227 (Part II), 374–384 (Part III).

Kendall, M.G., 1961, Natural law in the social science, Journal of the Royal Statistical Society A 124, 1–16.

Krishnaji, N., 1970a, A characteristic property of the Yule distribution, Sankhyā A 32, 343–346.

Krishnaji, N., 1970b, Characterization of the Pareto distribution through a model of under-reported incomes, Econometrica 38, 251–255.

Simon, H.A., 1955, On a class of skew distribution functions, Biometrika 42, 425–440.

Xekalaki, E., 1983, A property of the Yule distribution and its applications, Communications in Statistics, Theory and Methods 12, 1181–1189.