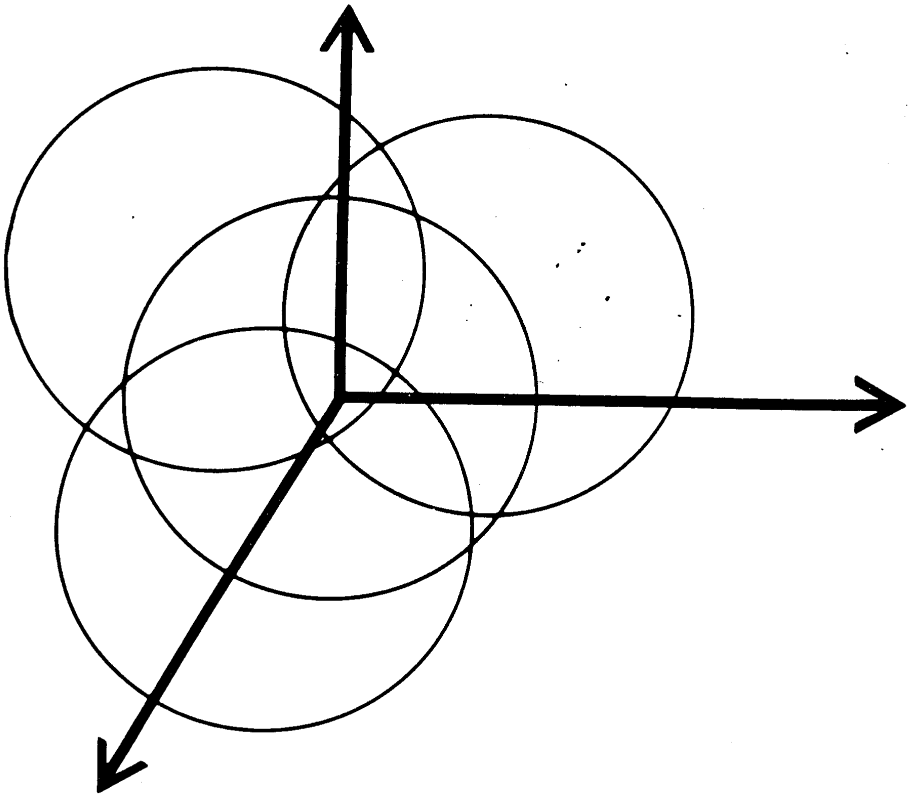

A reprint from

communications in statistics

COMMUN. STATIST.-THEORY METH. 19(8), 2757-2766 (1990)

IDENTIFIABILITY OF INCOME DISTRIBUTIONS IN THE CONTEXT OF DAMAGE AND GENERATING MODELS

Caterina Dimaki and Evdokia Xekalaki

Department of Statistics, The Athens University of Economics,
76 Patision St., 104 34 Athens, Greece.

Keywords: Regression function, Damage Model, Generating Model, Characterization, Income underreporting, Insurance Overreporting, Pareto Distribution, Power function distribution.

ABSTRACT

In the context of additive or multiplicative damage models, and under mild conditions, it is shown that the functional forms of suitably chosen regressions on a random variable X or/and its recorded part Y are characteristic of the distribution of X. The paper treats the cases where the recorded value is either an understatement or an overstatement of the true observation.

1. INTRODUCTION

Very often, the observed or recorded value Y of an observation differs from the actual value of the original observation X as a result of some **destructive** process so that the recorded observation Y can be regarded as the **undamaged** part of X. Such situations arise for example in area sampling (where the size of a herd of animals is recorded as equal to Y instead of X due to visibility bias), in labour force surveys (where registered unemployment Y is used instead of actual unemployment X), in income distribution analysis (where people tend to underreport their income for tax purposes) and in insurance claim distribution analysis (where people tend to overreport their true insurance claim). Most, if not all, of such practical situations can be viewed as falling within the framework of the **damage model** ,whereby an original observation X subjected to additive or multiplicative damage is recorded as Y≤X or within the framework of the **generating model** whereby X due to additive or multiplicative enhancement is recorded as Y≥X. The above mentioned two forms of the damage model have been considered in the literature as tools for obtaining characterization results based on conditions that relate the damaged part of X to its undamaged part. From among the results that have appeared in the literature in the context of damage models those for which damage is identified with underreporting of income, property or wealth will be in the focus of the present paper. Characterizations of this type often utilize appropriately chosen regression functions. These enhance the application potential of the characterization results as information on the regressors and the regressing random variable is more easily accessible. So, by considering multiplicative damage Krishnaji (1970), proved that the linearity of the regression of a random variable Z on X and on its observable part Y characterizes the distribution of the original random variable X as Pareto. Xekalaki (1984) proved that among the discrete income distributions the Yule is the only distribution that preserves the lineality of regression in Krishnaji's sense. Within the framework of the additive damage model Revankar et al (1974) proved that the linearity of the regression of the damaged part of X on the original part, characterizes the distribution of X as Pareto.

This paper investigates the question of whether results of more general nature hold in the context of the above models that will allow characterizations of more general families of distributions. The section that follows shows that indeed this is the case. So, section 2 provides two theorems that in the context of additive or multiplicative damage models show that the functional form of the regression on X or/and Y of an appropriately chosen random variable is characteristic of the distribution of X under assumptions that can be thought of as mild in connection with practical situations. These results consider X to be an absolutely continuous random variable. In order to cover cases such as an overreport of a true insurance claim, the previously described two forms of damage model may accordingly be converted so as to lead to further characterization results. So, in section 3 a generating model is considered whichby the recorded observation Y is an overstatement of the actual observation X i.e., $Y \geq X$. A model of this type and of an additive nature has been considered by Panaretos (1983) who refers to it as the generating model. Within the framework of this model, but allowing for both additive and multiplicative enhancement of the value of X it is shown that the functional form of the regression on X or Y of an appropriately chosen random variable is characteristic of the distribution of X which is again considered to be an absolutely continuous random variable.

2. CHARACTERIZATIONS IN THE CONTEXT OF UNDERREPORTING.

Let the actual observation be denoted by X, a random variable on $(0,+\infty)$. Let Y, $Y \leq X$ be the observable part of X and assume that $Y=RX$, where R is a random variable independent of X and distributed in the interval (0,1) according to the power function distribution. The problem to be studied then, would be the effect of the distribution of X on the regression $E(Z|Y=y)$ of any random variable Z (independent of R) on Y when the regression $E(Z|X=x)$ is of a given form. Conversely, the effect of the form of $E(Z|Y=y)$ on the distribution of X will be examined. The following theorem shows that the former uniquely determines the latter.

Theorem 2.1. Let X be an absolutely continuous random variable with a non-degenerate distribution and let

$$h(x) = E(Z|X=x), \quad x > 0 \tag{2.1}$$

be a non-constant function of x where Z is another random variable with an arbitrary distribution function. Further, let R be a random variable independent of Z and X with a density function given by:

$$f_R(r) = \rho\, r^{\rho-1}, \quad 0 < r < 1,\ \rho>0 . \tag{2.2}$$

Then the functions h(y), y>0 and

$$\lambda(y)= E(Z|Y=RX=y) \quad y>0 \tag{2.3}$$

uniquely determine the distribution of X.

Proof. Let $F_W(w)$ and $f_W(w)$ be the joint distribution function and joint probability density function of a random vector $W = (w_1,\dots,w_n)$, $n \geq 1$. Then, if Z takes on values in an interval denoted by R_Z,

$$E(Z|Y=y) = \int_{R_Z} z\, f_{Z|(Y=y)}(z)\, dz= \int_{R_Z} z\, \frac{f_{Z,Y}(z,y)}{f_Y(y)}\, dz. \tag{2.4}$$

But, from (2.2) and since R is independent of both X and Z, we have

$$F_{Z,Y}(z,y) = \int_0^1 F_{Z,X}\left(z, \frac{y}{r}\right) \rho r^{\rho-1}\, dr$$

which implies that,

$$f_{Z,Y}(z,y) = \rho y^{\rho-1} \int_y^\infty f_{Z,X}(z,x)\, x^{-\rho}\, dx. \tag{2.5}$$

Consequently, combining (2.4) and (2.5)

$$E(Z|Y=y) = \left[f_Y(y)\right]^{-1} \rho y^{\rho-1} \int_y^\infty \left[\int_{R_Z} z\, f_{Z|(X=x)}(z)\, dz\right] f_X(x)\, x^{-\rho} dx ,$$

or, using (2.1)

$$E(Z|Y=y) = \left[f_Y(y)\right]^{-1} \rho y^{\rho-1} \int_y^\infty h(x)\, f_X(x)\, x^{-\rho} dx. \tag{2.6}$$

But $F_Y(y) = \int_0^1 P(RX\leq y|R=r)\, f_R(r)dr = \int_0^1 F_X\left(\frac{y}{r}\right) f_R(r)\, dr.$

Therefore,

$$f_Y(y) = \rho y^{\rho-1}\int_y^{\infty} f_X(x)x^{-\rho}\,dx. \qquad (2.7)$$

Hence, combining (2.6) and (2.7)

$$E\,(Z|Y=y) = \int_y^{\infty} h(x)\;f_X(x)x^{-\rho}dx \Big/ \int_y^{\infty} h(x)\;f_X(x)x^{-\rho}dx.$$

Using (2.3) this equation becomes

$$\int_y^{\infty} h(x)\;f_X(x)x^{-\rho}dx = \lambda(y)\int_y^{\infty} f_X(x)x^{-\rho}\,dx.$$

Differentiating with respect to y it follows that

$$-h(y)f_X(y)y^{-\rho}= \lambda'(y)\int_y^{\infty} f_X(x)x^{-\rho}\,dx - \lambda(y)\;f_X(y)y^{-\rho}. \qquad (2.8)$$

Letting $K(y) = \int_y^{\infty} f_X(x)x^{-\rho}\,dx$ relationship (2.8) becomes

$$[h(y) - \lambda(y)]K'(y) = \lambda'(y)K(y).$$

Obviously, $h(y)-\lambda(y)\neq 0$ as otherwise $f_X(x)=0$ which would contradict the assumption that the distribution of X is non-degenerate. Hence

$$\frac{K'(y)}{K(y)} = \frac{\lambda'(y)}{h(y)-\lambda(y)}.$$

The solution of this differential equation is

$$K(y) = C\exp\left\{\int \frac{\lambda'(y)}{h(y)-\lambda(y)}\,dy\right\}.$$

This is equivalent to

$$\int_y^{\infty} f_X(x)x^{-\rho}\,dx = C\exp\left\{\int \frac{\lambda'(y)}{h(y)-\lambda(y)}\,dy\right\}$$

which by differentiation leads to

$$f_X(y) = Cy^{\rho}\,\frac{\lambda'(y)}{\lambda(y)-h(y)}\cdot\exp\left\{\int \frac{\lambda'(y)}{h(y)-\lambda(y)}dy\right\}. \qquad (2.9)$$

Hence, the result.

Corollary 2.1. (Characterization of the Pareto Distribution). Let X, Y and Z be defined as in theorem 2.1 and assume that $E(Z|X=x)=\delta+\beta x^{\alpha}$, $\alpha,\delta,\beta\in\mathbb{R}$. Then $E(Z|Y=y)=\delta+\gamma y^{\alpha}$, $\gamma\in\mathbb{R}$ if and only if X has a Pareto distribution provided that $\alpha\gamma/(\gamma-\beta)>\rho$.

Corollary 2.2. (Characterization of the F distribution). Let X, Y and Z be defined as in therem 2.1 and assume that

$E(Z|X=x)=\delta_1+\beta x$, $\delta_1,\beta \in \mathbb{R}$. Then $E(Z|Y=y)= \delta_0+\gamma y$, $\delta_0<\delta_1$, $\gamma<\beta$ if and only if X has Fisher's F distribution with $\beta-\gamma$ and $\delta_1-\delta_0$ degrees of freedom provided that $\beta-\gamma = 2(\rho+1)\in\mathbb{Z}$ and $\delta_1-\delta_0= 2\gamma/(\rho+1)-2(\rho+1)\in\mathbb{Z}$.

Remark 1. Krishnaji's (1970) result can be considered as a special case of corollary 2.1 for $\alpha=1$.

Remark 2. Let $f_X(x;\theta) = x^{-\theta} \exp \{-A(x)+B(\theta)\}$ be the probability density function of the log-exponential family .Substituting $f_X(x;\theta)$ into (2.9) and differentiating the resulting equation with respect to y it follows that

$(\theta+\rho)/y = D(y) - A'(y) - D'(y)/D(y)$ where $D(y)= \lambda'(y)/(h(y)-\lambda(y))$.

So, it is obvious that the form of $A(y)$ determines a unique relationship between $h(y)$ and $\lambda(y)$.

In the sequel, we treat the additive damage case.

Let X, a random variable with a non-degenerate probability distribution on $(m,+\infty)$ where $m>0$, denote the actual observation and let Y, $Y\le X$ be its observable part. Assume further that $Y=X-U$ and $0<U<\max\,(0,X-m)$. Then, the following theorem can be shown to hold.

Theorem 2.2. Let X,Y and U be defined as before and let (2.10) hold. Then the functions $h(x) = E\,(U|X=x)$, $x>m$ and $g(y)=E(U|X>y)$ uniquely determine the distribution of X.

Proof. Let $F_W(w)$ and $f_W(w)$ be the distribution function and probability density function respectively, of a random vector $W = (w_1,\ldots w_n)$, $n \ge 1$. Then, if U takes on values in an interval denoted by R_U,

$$g(y) = E(U|X>y) = \int_{R_U} u \; f_{U|(X>y)}(u)\, du.$$

Hence,

$$\int_y^\infty f_X(x) \; g(y)dx = \int_{R_U} u \left[\int_y^\infty f_{U,X}\,(u,x)\, dx\right] du$$

or, equivalently

$$\left[1-F_X(x)\right].g(y) = \int_y^\infty f_X(x) \left[\int_{R_U} u \, f_{U|(X=x)}(u)\, du\right] dx.$$

Substituting h(x) for E(U|X=x) we obtain

$$\left[1-F_X(y)\right].g(y) = E\left[h(x)\right] + \int_m^y h(x)\ d\left[1-F_X(x)\right].$$

Integrating by parts and since $F_X(m)=0$

$$\left[1-F_X(y)\right].g(y)=E\left[h(x)\right]+h(y)\left[1-F_X(y)\right]- h(m)-\int_m^y \left[1-F_X(x)\right] h'(x)\ dx.$$

Differentiating with respect to y and letting $\bar{F}_X(x)=1-F_X(x)$ it follows that

$$\bar{F}_X'(y)\left[g(y)-h(y)\right] = -\bar{F}_X(y).g'(y).$$

Obviously $g(y)-h(y)\neq 0$ since otherwise $\bar{F}_X(y)=0$ for every $y>0$ which would imply a degenerate distribution for X. Therefore

$$\bar{F}_X{}'(y)/\bar{F}_X(y) = g'(y)/(g(y)-h(y)).$$

The solution of this differential equation leads to

$$f_X(y) = C \frac{g'(y)}{g(y)-h(y)} .\exp\left\{\int \frac{g'(y)}{h(y)-g(y)}dy\right\}.$$

Therefore $f_X(y)$ is uniquely determined by the regression functions $g(y) = E(U|X=y)$ and $h(y) = E(U|X>y)$.

Corollary 2.2. (Characterization of the Pareto Distribution). Let X, Y and U be defined as before and assume that $E(U|X=x)=\delta+\beta x^{\alpha}$, $\alpha,\beta,\delta\in\mathbb{R}$. Then $E(U|X>y)=\delta+\gamma y^{\alpha}$, $\gamma\in\mathbb{R}$, $\alpha\gamma/(\gamma-\beta)>0$ if and only if X has a Pareto distribution.

Remark 3. Revankar et al.'s (1974) result can be regarded as a special case of corollary 2.2 for $\alpha=1$.

3. CHARACTERIZATIONS IN THE CONTEXT OF OVERREPORTING

Let X denote an actual observation and Y its observable (recorded) part. This section treats the case of problems in the context of the generating model where the recorded value Y is an overstatement of the true observation X. Before proceeding to the proof of the main results of this section we prove the following lemma.

Lemma 3.1. Let X be a random variable with an absolutely continous distribution function and let R be a random variable independent of X with probability density function given by (2.2). Assume, that $P(Y=X/R < x_0)>0$ for some $x_0>0$. Then the distribution of X/R truncated to the right at x_0 coincides with the distribution of X if and only if X follows a power distribution on $(0,x_0)$.

Proof. Adopting the notation of section 2 we have that

$$F_Y(y) = \int_0^1 F_X(yr)dF_R(r). \qquad (3.1)$$

Note that (3.1) holds for any arbitrary distribution of R provided that its range is contained in (0,1).

Necessity : Let X be distributed according to the power distribution on (0,x). Then

$$\bar{F}_X(x) = \begin{cases} (x/x_0)^{\alpha} \text{ for } x<x_0;\ x_0>0,\ \alpha>0 \\ 1 \text{ otherwise} \end{cases} \qquad (3.3)$$

Observing that

$F_Y(x_0) = \int_0^1 F_X\,(x_0r)d\,F_R(r)$ we have for any $y<x_0$,

$$F_Y(y) - F_Y(x_0)\,F_X(y) = \int_0^1 \left[F_X(yr)-F_X(x_0r)F_X(y)\right] dF_R(r).$$

But $\bar{F}(xyx_0) = \bar{F}(xx_0)\bar{F}(yx_0)$. Therefore $F_X(yr) = F_X(y)\,F_X(rx_0)$.i.e., $F_X(y)/F_Y(x_0) = F_X(y)$. Hence,

$$P\left[Y = \frac{X}{R} \le y\,|\,Y<x_0\right] = \frac{P(Y\le y,\ Y<x_0)}{P(Y<x_0)} = \frac{F_Y(y)}{F_Y(x_0)} = F_X(y)$$

Sufficiency: Combining (2.2) and (3.1) it follows that

$$F_Y(y) = \rho y^{-\rho}\int_0^y F_X(x)\ x^{\rho-1}\ dx$$

or, equivalently

$$\frac{1}{\rho}\ F_Y(y)y^{\rho} = \int_0^y F_X(x)\ x^{\rho-1}\ dx.$$

Differentiating with respect to y and letting $K(y)= F_X(y)\ y^{\rho}$yields

$$\frac{K'(y)}{K(y)} = \frac{\rho}{y\ F_X(x_0)}$$

whose solution leads to

$$F_X(y) = C\ y^{\rho\left[\frac{1}{F_Y(x_0)}-1\right]}$$

Hence the lemma has been established.

Consider now a multiplicative generating model whereby X denotes an actual observation and Y denotes its observable part which is at least equal to X. More specifically assume that the damage on X is effected through the relationship $Y=\frac{X}{R}$ where R is a random variable independent of X and distributed according to the power distribution in the interval (0,1). Then the following theorem can be shown.

Theorem 3.1. Let Z be a random variable with an arbitrary distribution and let X be a random variable with an absolutely continouous non-degenerate probability distribution. Let R be a random variable independent of Z and X with a density given by (2.2). Then, the functions $h(y)=E(Z|X=x)$ and $\lambda(y)=E\left[Z|Y=\frac{X}{R}=y\right]$ uniquely determine the distribution of X.

Corollary 3.1 (Characterization of the Pareto Distribution). Let X,Y and Z be defined as in theorem 3.1 and assume that $E(Z|X=x)=\alpha+\delta x$, $\alpha,\delta\in\mathbb{R}$. Then $E(Z|Y=y) = \alpha+\beta y$, $\beta\in\mathbb{R}$ if and only if X has a Pareto distribution of the first kind provided that $\beta/(\delta-\beta)<\rho$.

Considering now overreporting within the framework of an additive generating model the following theorem can be shown.

Theorem 3.2 Let Y,X,U be random variable defined as in theorem 2.2 and assume that Y=X+U.
Then, the functions $h(x)=E(U|X=x)$ and $g(y) = E\ (U|X<y)$ uniquely determine the distribution of X.

ACKNOWLEDGEMENT

The authors would like to thank the referees for constructive comments.

BIBLIOGRAPHY

1. **Krishnaji,** N. (1970). *Characterization of the Pareto Distribution Through a Model of Underreported Incomes.* **Econometrica**, 38, 251-255.

2. Panaretos, J. (1983). *"A Generating Model Involving Pascal and Logarithmic Series Distributions"* **Communications in Statistics - Theory and Methods**, 12, 841-848.

3. Revankar, N.S. Hartley, M.J. and Pagano, M. (1974). A Characterization of the Pareto Distribution. **The Annals of Statistics**, 2, 599-601.

4. Xekalaki, E. (1984). *Linear Regression and the Yule Distribution.* **Journal of Econometrics**, 24, 397-403.

Received by Editorial Board Member February 1990;
Revised June 1990.

Recommended by K. Kocherlakota, University of Manitoba,
CANADA.

Refereed by J. Keith Ord, Penn State University,
University Park, PA.and Anonymously.