Journal of Multivariate Analysis 102 (2011) 674–682

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Composing the cumulative quantile regression function and the Goldie concentration curve[☆]

SzeMan Tse

Department of Applied Mathematics, National Dong Hwa University, Hualien, Taiwan

ARTICLE INFO

Article history:
Received 10 December 2009
Available online 2 December 2010

AMS subject classifications:
60F15
60F17
62G05
62G30

Keywords:
Goldie concentration curve
Cumulative quantile regression function
Induced order statistics
Lorenz curve
Strong Gaussian approximations

ABSTRACT

The model we discuss in this paper deals with inequality in distribution in the presence of a covariate. To elucidate that dependence, we propose to consider the composition of the cumulative quantile regression (CQR) function and the Goldie concentration curve, the standardized counterpart of which gives a fraction to fraction plot of the response and the covariate. It has the merit of enhancing the visibility of inequality in distribution when the latter is present. We shall examine the asymptotic properties of the corresponding empirical estimator. The associated empirical process involves a randomly stopped partial sum process of induced order statistics. Strong Gaussian approximations of the processes are constructed. The result forms the basis for the asymptotic theory of functional statistics based on these processes.

1. Introduction

Let (X, Y) be a positive bivariate random vector with right continuous distribution function (df) $P(x, y)$, and let X have marginal df $F(x)$. Assume that Y is integrable, and denote by $m(x) = E[Y \mid X = x]$ the regression function of Y on X. Let $Q(u) := \inf\{x : F(x) > u\}$, $0 < u < 1$, denote the right continuous quantile function (qf) associated with $F(x)$. With $U = F(X)$ denoting the rank variable, Rao and Zhao [13] defined the quantile regression (QR) function of Y on X as

$$r(u) = \mathbf{E}[Y \mid U = u] = m \circ Q(u), \quad 0 \le u \le 1. \tag{1.1}$$

The cumulative QR (CQR) function and its standardized form are

$$\begin{aligned} M(u) &:= \int_0^u m \circ Q(t)\mathrm{d}t = \int_0^u r(t)\mathrm{d}t, \\ N(u) &:= \frac{1}{\mu} M(u) \end{aligned} \qquad 0 \le u \le 1, \tag{1.2}$$

where $\mu = M(1) = \mathbf{E}[Y]$. In econometrics, with (X, Y) representing income and tax respectively, $N(u)$ is the fraction of total tax contributed by the lowest uth fraction of income holders. In insurance, large claims are of more interest than small claims. With (X, Y) representing the risk index and the claim, respectively, the standardized form of the dual curve

☆ Work partially supported by NSC Grant 96-2118-M-259-001-MY2.
E-mail address: tse@server.am.ndhu.edu.tw.

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.11.010

$M(1) - M(u)$, called the tail conditional expectation, gives the fraction of the total claim due to the upper $(1-u)$th fraction of risk groups (see [8,9]).

The standardized CQR function can be regarded as the generalized version of the Lorenz curve in the presence of a covariate. For the special case $Y = X$, m becomes the identity function, the QR function $r(u)$ reduces to the qf $Q(u)$, and $N(u)$ reduces to the usual Lorenz curve (see [10]). The notion is particularly useful for comparing the concentration and inequality of the Y distribution in relation to that of X, and thus has a wide scope of application. Note, however, that $N(u)$ as defined in (1.2) may not lie below the line joining the points (0, 0) and (1, 1), which is a characteristic of a Lorenz curve.

Assuming that X has finite expectation, the Lorenz curve corresponding to X,

$$L_F(u) := \frac{1}{EX}\int_0^u Q(y)\mathrm{d}y, \quad 0 \le u \le 1, \tag{1.3}$$

is continuous and strictly increasing on [0, 1]. The inverse function L_F^{-1}, or simply L^{-1}, is well defined on [0, 1]. Goldie [11] called it the concentration curve associated with F. We shall follow Csörgő and Zitikis [6], and call it the Goldie curve. It is also continuous and strictly increasing. In the income and tax example above, $L^{-1}(u)$ gives the fraction of the lowest income group which accounted for the uth fraction of total income.

For the purpose of exposing the correlation between different variables regarding inequality in distribution, a more enlightening and direct alternative to $M(u)$ is the composite function $M \circ L^{-1}(u)$, with standardized counterpart

$$R(u) := \frac{M \circ L^{-1}(u)}{M \circ L^{-1}(1)}, \quad 0 \le u \le 1. \tag{1.4}$$

We shall refer at it as the R curve. The duals of L_F, L^{-1} and R can be defined similarly. Again in the context of the econometrics example above, $R(u)$ gives the fraction of the total tax from the lowest income group which accounted for the uth fraction of the total income. Thus, the graph of $R(u)$ is a fraction to fraction plot of total tax contributed versus total income earned, with the 45° line serving as the ideal reference line. The level of inequality in distribution of the response variable in relation to the explanatory variable is reflected by the degree of deviation in either direction from the reference line. Functional statistics based on this curve would be very helpful in making inference.

To appreciate the difference between $N(u)$ and $R(u)$, consider again the tax and income example. Suppose that 50% of the lowest income group contributed 70% of the total taxes, but accounted for only 30% of the total income. In terms of the function $N(u)$, we have $N(0.5) = 0.7$, while, in terms of the function $R(u)$, we have $R(0.3) = 0.7$. While both sets of numbers are informative, the latter representation, which amounts to a rescaling of the covariate quantiles by the Goldie curve, has the merit of putting the two variables on an equal footing, thus enhancing the visibility of distributional inequality when it is present.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. as (X, Y), with F_n and Q_n denoting the right continuous empirical df and qf of X, respectively, i.e., $Q_n(u) := \inf\{x : F_n(x) > u\}$ for $0 < u < 1$. Goldie [11] showed that uniform consistent estimators of $L_F(u)$ and $L^{-1}(u)$ are given by their empirical counterparts

$$\begin{aligned} L_n(u) &:= \frac{1}{\bar{X}}\int_0^u Q_n(y)\mathrm{d}y, \quad 0 \le u \le 1, \\ L_n^{-1}(u) &:= \inf\{y : L_n(y) > u\}, \quad 0 \le u \le 1, \end{aligned} \tag{1.5}$$

where $\bar{X}$ is the sample mean of the X_is. He also derived the weak convergence limits of the Lorenz process $l_n(u) := \sqrt{n}[L_n(u) - L_F(u)]$ and the concentration process $c_n(u) := \sqrt{n}[L_n^{-1}(u) - L^{-1}(u)]$ for $0 \le u \le 1$. Csörgő [2] and Csörgő et al. [3] gave a unified treatment of strong and weak approximations of the Lorenz and concentration processes. In particular, they established a strong invariance principle that Rao and Zhao [14] applied to derive a law of the iterated logarithm (LIL) for the Lorenz curve. Csörgő and Zitikis [4,5] proved a more general version of the LIL under weaker assumptions, and in [6] they constructed confidence bands for the Lorenz and Goldie curves. Tse [17] established the corresponding results for censored and truncated data.

To find an estimator for $M \circ L^{-1}(u)$, let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics of $X_1, \ldots, X_n$, and denote the Y associated with $X_{(i)}$ by $Y_{(i)}$. The latter are called induced order statistics or concomitants (see [12] for a review, especially pp. 228–229, as well as the references therein). The empirical CQR function and the associated CQR process are, for $u \in [0, 1]$,

$$\begin{aligned} M_n(u) &:= \frac{1}{n}\sum_{i=1}^{[nu]+1} Y_{(i)} \\ Z_n(u) &:= \sqrt{n}[M_n(u) - M(u)]. \end{aligned} \tag{1.6}$$

The empirical counterpart of $M \circ L^{-1}(u)$ is the randomly stopped sum process $M_n \circ L_n^{-1}(u)$ of the induced order statistics. It is closely related to the partial sum processes of induced order statistics studied in [7]. The standardized counterpart is

$$R_n(u) := \frac{M_n \circ L_n^{-1}(u)}{M_n \circ L_n^{-1}(1)}. \tag{1.7}$$