Asian Journal of Probability and Statistics

20(3): 24-35, 2022; Article no.AJPAS.92094

ISSN: 2582-0230

Link Function for Quantile Estimation in Regression Settings

Md. Rezaul Karim [a*] and Sejuti Haque [a]

[a] *Department of Statistics, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh.*

Authors' contributions

This work was carried out in collaboration between both authors. Author MRK designed the study, performed the statistical analysis, wrote the protocol, and wrote the first draft of the manuscript. Author SH managed the analyses of the study and managed the literature searches. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2022/v20i3424

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/92094

Received: 15 July 2022

Accepted: 19 September 2022

Published: 18 October 2022

Original Research Article

Abstract

This paper studies the quantile estimation by using the link function under a broad family of asymmetric densities known as a generalized quantile-based asymmetric family. We proposed a link function and quantile estimation in regression settings. The estimator's asymptotic properties of the estimators are also discussed here. To demonstrate the proposed methods for estimating the quantile function, an actual data application including the proportion of daily SARS-Cov-2 infected persons tested for COVID-19 infection and meteorological factors such as temperature and humidity is included. We discovered that the amount of daily SARS-Cov-2 infected people tested for COVID-19 infection is significantly influenced by temperature and humidity.

Keywords: Generalized quantile-based asymmetric family; link function; quantile estimation; COVID-19.

**Corresponding author: E-mail: rezaul@juniv.edu;*

1 Introduction

Regression is one of the fundamental statistical tools that determines the strength and nature of the relationship between a set of response variables and a set of covariates. The mean regression focuses on an average relationship between a set of response variables and a set of covariates. It provides a single characteristic of a conditional distribution. It performs better results with nice mathematical properties for symmetric distribution (e.g., normal distribution). It is also not suitable when the data comes from the skewed distribution (see, for example, [1]). In regard to parameter estimation (or in general statistical inference and asymptotic properties) for a given distributions, a convenient class of families is the exponential family where the response variable Y given a covariate X is as follows

$$f_{Y|X}(y|X=x) = \exp\left(\frac{(y\theta(x)-b(\theta(x)}{a(\phi)} + c(y;\phi)\right), \tag{1.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are measurable functions (see, for example, [2]). The parameter function $\theta(\cdot)$ is called the canonical parameter and ϕ is a scale parameter.

$$m(x) \equiv E(Y|X=x) = b'(\theta(x))$$
$$\text{var}(Y|X=x) \equiv \alpha(\phi)b'(\theta(x))$$

The function $g(b')^{-1}$, which links the mean regression function to the canonical parameter function $(b')^{-1}(m) = \theta$ is called the canonical link.

However, the mean regression is highly influenced by extreme values. It is not usable when the quantile of the conditional distribution is the main interest (see, for example, [3]).

Therefore, [4] provided tick exponential family, whose role in the conditional quantile estimation is analog to the role of the linear exponential family (1.1) in the conditional mean estimation. The general form of the tick exponential family for $y \in \mathbb{R}$ is given by

$$f_\alpha(y;\eta) = \alpha(1-\alpha)g\prime(y)\begin{cases} \exp\left[-(1-\alpha)(g(\eta)-g(y))\right] & \text{if} \quad y \leq \eta \\ \exp\left[\alpha(g(\eta)-g(y))\right] & \text{if} \quad y > \eta. \end{cases} \tag{1.2}$$

It is noted that the tick-exponential family (1.2) is only used for the whole real line continuous variable. It is not useful for boundary response variables. Beside of this, the asymmetric Laplace is the only member of this family.

On the other hand, [5] proposed quantile regression which minimizes the tick function. It provides full characteristics of the distribution. This quantile regression is actually nonparametric because it does not need the underlying parametric assumption. It is more robust to outliers than mean regression. It is the only regression tool which is used for finding the effect of the covariate on different quantile level of the response variables. A nice discussion of quantile regression is presented by [3]. There are many problems that arise in nonparametric quantile regression due to the unknown underlying distribution. For example, crossing problem in quantile curves which leads to invalid inference, less efficiency etc.

Recently, [1] proposed the generalized quantile-based asymmetric family for any continuous variable Y which takes the form

$$f_\alpha^g(y;\eta,\phi) = \frac{2\alpha(1-\alpha)g'(y)}{\phi}\begin{cases} f\left((1-\alpha)\left(\frac{g(\eta)-g(y)}{\phi}\right)\right) & \text{if} \quad y \le \eta \\ f\left(\alpha\left(\frac{g(y)-g(\eta)}{\phi}\right)\right) & \text{if} \quad y > \eta, \end{cases} \tag{1.3}$$

where η is the location parameter and the αth quantile of Y and ϕ is the scale parameter. The speciality of family (1.3) is the location parameter(η) is a specific quantile of this family. There are many members of the family available in the literature such as asymmetric normal, asymmetric Laplace, asymmetric t and at least three big families that are subset of this family e.g, tick exponential (see, [6]), asymmetric power family (see, [6]), quantile-based asymmetric family (see, [6]). For any $\beta \in (0,1)$, the βth-quantile of Y equals

$$\{F_\alpha^g\}^{-1}(\beta;\eta,\phi) = \begin{cases} g^{-1}\left(g(\eta) + \frac{\phi}{1-\alpha}F^{-1}\left(\frac{\beta}{2\alpha}\right)\right) & \text{if} \quad \beta \le \alpha \\ g^{-1}\left(g(\eta) + \frac{\phi}{\alpha}F^{-1}\left(\frac{1+\beta-2\alpha}{2(1-\alpha)}\right)\right) & \text{if} \quad \beta > \alpha, \end{cases}$$

with in particular $\{F_\alpha^g\}^{-1}(\alpha;\eta,\phi) = \eta$. In the regression setting, the family (1.3) can be written as

$$\{f_{\alpha,Y|X}^g(y\,;\eta(x),\phi(x)|X=x) = \frac{2\alpha(1-\alpha)g'(y)}{\phi(x)}\begin{cases} f\left((1-\alpha)\left(\frac{g(\eta(x))-g(y)}{\phi(x)}\right)\right) & \text{if} \quad y \le \eta(x) \\ f\left(\alpha\left(\frac{g(y)-g(\eta(x))}{\phi(x)}\right)\right) & \text{if} \quad y > \eta(x), \end{cases} \tag{1.4}$$

where $\eta(x)$ and $\phi(x)$ are now the function of the covariate(s) x. In the setting of (1.4), the βth-conditional quantile function of Y given X = x (with $(0<\beta<1)$) is then

$$\{F_{Y|X,\alpha}^g\}^{-1}(\beta;\eta(x),\phi(x)|x) = g^{-1}(g(\eta(x)) + \phi(x).C_\alpha(\beta))$$

where

$$C_\alpha(\beta) = \frac{1}{1-\alpha}F^{-1}\left(\frac{1+\beta}{2\alpha}\right)\mathbb{I}(\beta<\alpha) + \frac{1}{\alpha}F^{-1}\left(\frac{1+\beta-2\alpha}{2(1-\alpha)}\right)\mathbb{I}(\beta\ge\alpha).$$

With F^{-1} the quantile function associated with the reference symmetric density f. The quantity $C_\alpha(\beta)$ is known as a constant and is a monotonic function of β. The family (1.3) depends on two vital elements:

- the reference symmetric density f and
- monotone strictly increasing link function g.

When the link function is identity (i.e., $g(y) = y$) then family tends to quantile-based asymmetric family given in [6]. In this study, the reference symmetric density f is assumed to be known. So the main focus is to estimate the link function g.

The link function is a crucial element in semiparametric quantile regression under the generalized quantile-based asymmetric family (see, [1]). It allows to explain any type of continuous response in terms of covariates. Besides, estimating the maximum likelihood estimator for unconditional setting and local likelihood estimator for conditional setting, the link function should be known (see, [1]). Usually we assume that, the link function in semiparametric quantile regression is known. For example, identity link, logit link, log link, canonical link, reciprocal link etc. But in real life data application, the link function is unknown. So it is very important to estimate link function.

Alternatively, logit-type link function could be a solution. So, in this research, we focus on the study of different link functions in semiparametric regression.

The main contribution of the study is to estimate the link function of the generalized quantile-based asymmetric family in regression settings. Section 2 presented a logit-type link function and derive its distribution. The estimation of the logit-type link function in conditional settings is described in Section 3. The real data application is added to demonstrate the proposed methodology. The concluding remarks are added in the final section.

2 Logit-type Link Function

Let the density of Y is a member of the generalized quantile-based family of distributions, and G is a distribution function of Y. Suppose g be a logit-type link function of Y depends on G such that

$$g(Y) = \text{logit}(G(Y)) = \ln\left(\frac{G(Y)}{1 - G(Y)}\right). \tag{2.1}$$

If we know the distribution function of G, we can easily derive the logit-type link function by using (2.1). The distribution of logit-type link function for quantile estimation in (2.1) has been studied in [7].

If η is the αth quantile of Y and g is the monotone strictly increasing link function, then $g(\eta)$ is also the αth quantile of $Z = g(Y)$ (see for example, [3]). By introducing the αth quantile parameter $\mu \in R$ and a scale parameter $\phi > 0$, we get

$$f_\alpha(z; \mu, \phi) = \frac{2\alpha(1-\alpha)}{\phi} \begin{cases} \dfrac{e^{-\alpha(\frac{z-\mu}{\phi})}}{\left(1+e^{-\alpha(\frac{z-\mu}{\phi})}\right)^2} & \text{if} \quad z > \mu \\ \dfrac{e^{-(1-\alpha)(\frac{\mu-z}{\phi})}}{\left(1+e^{-(1-\alpha)(\frac{\mu-z}{\phi})}\right)^2} & \text{if} \quad z \leq \mu, \end{cases} \tag{2.2}$$

where $F_\alpha^{-1}(\alpha) = \mu$. The density given in (2.2) is denoted by ALD(μ, ϕ, α) and called quantile-based asymmetric logistic density (ALD) proposed in [6]. We can easily find the quantile function of Z which is

$$F_\alpha^{-1}(\beta) = \begin{cases} \mu - \frac{\phi}{1-\alpha} \ln(\frac{2\alpha}{\beta} - 1) & ; \quad \text{if} \quad \beta < \alpha \\ \mu - \frac{\phi}{\alpha} \ln(\frac{1-\beta}{\beta-2\alpha+1}) & ; \quad \text{if} \quad \beta \geq \alpha. \end{cases}$$

3 Estimation of Logit-type Link Function for Conditional Setting

Let the density of Z_i is the link function of quantile-based asymmetric family of logistic distribution and F is distribution function of response y_i. The different link functions are presented in Fig.1 for the different densities.

$$Z_i = \ln\Big(\frac{F(y_i)}{1 - F(y_i)}\Big); \quad where, Z_i \sim \text{ALD}(\mu, \phi, \alpha)$$

In conditional setting, the link function Z_i is now function of covariate X_i takes the form,

$$Z_i|X_i \sim \text{ALD}(\mu(x), \phi(x), \alpha(x)); \quad where,\ Z_i|X_i = \ln\big(\frac{F(y|X)}{1 - F(y|X)}\big).$$

Where,

$$\mu(X) \in \mathbb{R},\ \phi(X) \in \mathbb{R} \text{ and } \alpha \in (0, 1).$$

The parameters are defined as,

$$\theta_1(\mathbf{X}) = \mathbf{X}\beta_1 = \mu(\mathbf{X}) \in \mathbb{R}$$
$$\theta_2(\mathbf{X}) = \mathbf{X}\beta_2 = \ln\phi(\mathbf{X}) \in \mathbb{R}$$
$$\theta_3(\mathbf{X}) = \mathbf{X}\beta_3 = \ln\frac{\alpha(\mathbf{X})}{1-\alpha(\mathbf{X})}$$

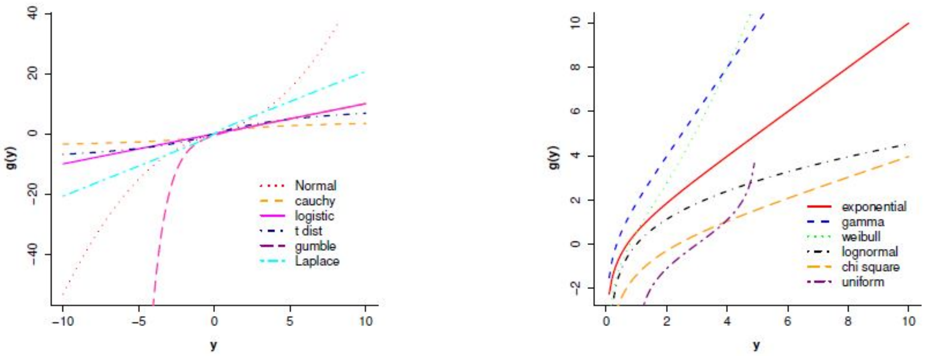

Fig. 1. Link function curve for (a). the real-valued random variable; (b). the semi-infinite supported random variable

Where,

$$\mathbf{z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}_{n\times 1} ; \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots X_{p1} \\ 1 & X_{12} & X_{22} & \dots X_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots X_{pn} \end{bmatrix}_{n\times(p+1)}$$

$$\text{and} \quad \beta_j = \begin{bmatrix} \beta_{j0} \\ \beta_{j1} \\ \beta_{j2} \\ \vdots \\ \beta_{jp} \end{bmatrix}_{(p+1)\times 1} ; j = 1, 2, 3.$$

We now turn to the regression setting involving one covariate. For conditional density of Z given $\mathbf{X} = \mathbf{x}$ we consider the density $f_{Z|X}(\cdot; \theta_1(x), \theta_2(x))$ in (2.2) and allow θ_1, θ_2 and index parameter α depend on $\mathbf{x}$ (See the density plot of ALD is given in Fig. 2). This leads to the conditional density

$$f_{Z|X,\theta_3(\mathbf{X})}(Z;\theta_1(\mathbf{X}),\theta_2(\mathbf{X})) = C \begin{cases} \frac{\exp\left[-\left(\frac{e^{\theta_3(\mathbf{X})}}{1+e^{\theta_3(\mathbf{X})}}\right)\left(\frac{Z-e^{\theta_1(\mathbf{X})}}{e^{\theta_2(\mathbf{X})}}\right)\right]}{\left[1+\exp\left(-(\frac{e_3^{\theta}(\mathbf{X})}{1+e^{\theta_3(\mathbf{X})}})(\frac{Z-e^{\theta_1(\mathbf{X})}}{e^{\theta_2(\mathbf{X})}})\right)\right]^2} & \text{if} \quad Z > e^{\theta_1(\mathbf{X})} \\ \frac{\exp\left[-\left(\frac{1}{1+e^{\theta_3(\mathbf{X})}}\right)\left(\frac{e^{\theta_1(\mathbf{X})}-Z}{e^{\theta_2(\mathbf{X})}}\right)\right]}{\left[1+\exp\left(-(\frac{1}{1+e^{\theta_3(\mathbf{X})}})(\frac{e^{\theta_1(\mathbf{X})}-Z}{e^{\theta_2(\mathbf{X})}})\right)\right]^2} & \text{if} \quad Z \leq e^{\theta_1(\mathbf{X})}, \end{cases} \tag{3.1}$$

where,

$$C = \frac{2e^{\theta_3(x)}}{(1+e^{\theta_3(x)})^2 e^{\theta_2(x)}}$$

The conditional likelihood function for $\theta(x) = (\mu(x), \phi(x), \alpha(x))^T$ can be written as

$$L(\theta) = \prod_{i=1}^{n} f_{Z|X,\theta_3(\mathbf{X})}(Z;\theta_1(\mathbf{X}),\theta_2(\mathbf{X}))$$

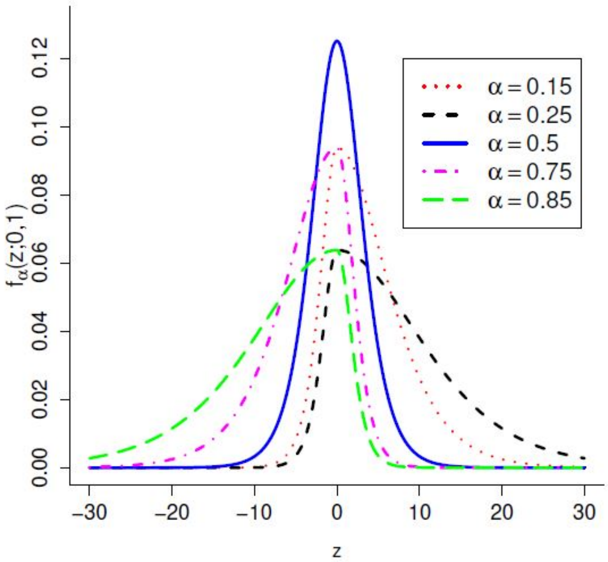

Fig. 2. The density plots of a quantile-based asymmetric logistic distribution with α = (0.15, 0.25, 0.50, 0.75, 0.85)th Quantile of $\mu = 0$ and $\phi = 1$

The conditional likelihood function for $\theta(x) = (\mu(x), \phi(x), \alpha(x))^T$ can be written as

$$\ln L = \sum_{i=1}^{n} \log f_{Z|X,\theta_3(\mathbf{X})}(Z;\theta_1(\mathbf{X}),\theta_2(\mathbf{X})).$$

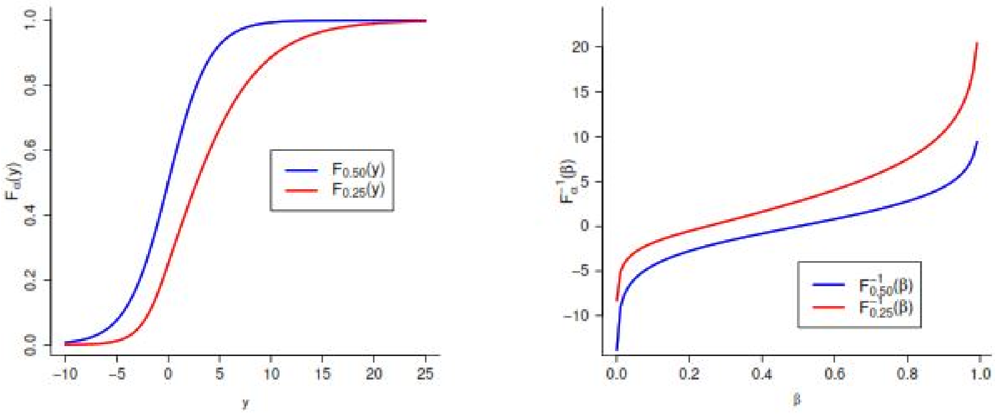

Fig. 3. Cumulative distribution function (left) and quantile function (right) for $\mu = 0$, $\phi = 1$ and $\alpha = (0.25, 0.50)$

For simplification we use, $l = \ln L$

$$
\begin{aligned}
l(\theta_1(\mathbf{X}), \theta_2(\mathbf{X}), \theta_3(\mathbf{X}); Z_i) = {} & n \ln\big(\frac{2e^{\theta_3}(\mathbf{X})}{1+e^{\theta_3}(\mathbf{X})}\big) - n\theta_2(\mathbf{X}) - \frac{1}{1+e^{\theta_3(\mathbf{X})}} \sum_{i=1}^{n} \frac{e^{\theta_1(\mathbf{X})} - Z_i}{e^{\theta_2(\mathbf{X})}} \\
& I(Z_i \le e^{\theta_1(\mathbf{X}))} \quad - 2\sum_{i=1}^{n} \ln\Big(1 + \exp\big(-\frac{e^{\theta_1(\mathbf{X})} - Z_i}{(1+e^{\theta_3(\mathbf{X})})e^{\theta_2(\mathbf{X})}}\big) \\
& I(Z_i \le e^{\theta_1(\mathbf{X}))} - \frac{e^{\theta_3(\mathbf{X})}}{1+e^{\theta_3(\mathbf{X})}} \sum_{i=1}^{n} \frac{Z_i - e^{\theta_1(\mathbf{X})}}{e^{\theta_2(\mathbf{X})}} I(Z_i > e^{\theta_1(\mathbf{X})})) \\
& - 2\sum_{i=1}^{n} \ln\Big(1 + \exp\big(-\frac{(Z_i - e^{\theta_1(\mathbf{X})})e^{\theta_3(\mathbf{X})}}{(1+e^{\theta_3(\mathbf{X})})e_2^{\theta}(\mathbf{X})}\big)\Big) I(Z_i > e^{\theta_1(\mathbf{X})}).
\end{aligned}
$$

$$
\begin{aligned}
l(\mathbf{X}\beta_1, \mathbf{X}\beta_2, \mathbf{X}\beta_3,) = {} & n \ln\big(\frac{2e^{\mathbf{X}\beta_3}}{1+e^{\mathbf{X}\beta_3}}\big) - n\mathbf{X}\beta_2 - \frac{1}{1+e^{\mathbf{X}\beta_3}} \sum_{i=1}^{n} \frac{e^{\mathbf{X}\beta_1} - Z_i}{e^{\mathbf{X}\beta_2}} I(Z_i \le e^{\mathbf{X}\beta_1}) \\
& - 2\sum_{i=1}^{n} \ln\Big(1 + \exp\big(-\frac{e^{\mathbf{X}\beta_1} - Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big) I(Z_i \le e^{\mathbf{X}\beta_1}) \\
& - \frac{e^{\mathbf{X}\beta_3}}{1+e^{\mathbf{X}\beta_3}} \sum_{i=1}^{n} \frac{Z_i - e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} I(Z_i > e^{\mathbf{X}\beta_1}) \\
& - 2\sum_{i=1}^{n} \ln\Big(1 + \exp\big(-\frac{(Z_i - e^{\mathbf{X}\beta_1})e^{\mathbf{X}\beta_3}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big) I(Z_i > e^{\mathbf{X}\beta_1}).
\end{aligned}
$$

3.1 Asymptotic properties of estimators

Differentiating the liklihood function with respect to β_1, β_2 and β_3 we get,

$$\frac{\delta l}{\delta \beta_1} = -K\sum_{i=1}^{n}\frac{e^{\mathbf{X}\beta_1}-Z_i}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_1}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{e^{\mathbf{X}\beta_1}-Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big) - Ke^{\mathbf{X}\beta_3}$$
$$\sum_{i=1}^{n}\frac{Z_i-e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_1}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{(Z_i-e^{\mathbf{X}\beta_1})^{e^{\mathbf{X}\beta_3}}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big),$$

$$\frac{\delta l}{\delta \beta_2} = -n\mathbf{X} + K\sum_{i=1}^{n}\frac{e^{\mathbf{X}\beta_1}-Z_i}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_2}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{e^{\mathbf{X}\beta_1}-Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)$$
$$+Ke^{\mathbf{X}\beta_3}\sum_{i=1}^{n}\frac{Z_i-e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_2}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{(Z_i-e^{\mathbf{X}\beta_1})^{e^{\mathbf{X}\beta_3}}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big),$$

$$\frac{\delta l}{\delta \beta_3} = 2Kn + \frac{Ke^{\mathbf{X}\beta_3}}{(1+e^{\mathbf{X}\beta_3})}\sum_{i=1}^{n}\frac{e^{\mathbf{X}\beta_1}-Z_i}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_3}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{e^{\mathbf{X}\beta_1}-Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)$$
$$+\frac{Ke^{\mathbf{X}\beta_3}}{(1+e^{\mathbf{X}\beta_3})}\sum_{i=1}^{n}\frac{Z_i-e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_3}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{(Z_i-e^{\mathbf{X}\beta_1})^{e^{\mathbf{X}\beta_3}}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big),$$

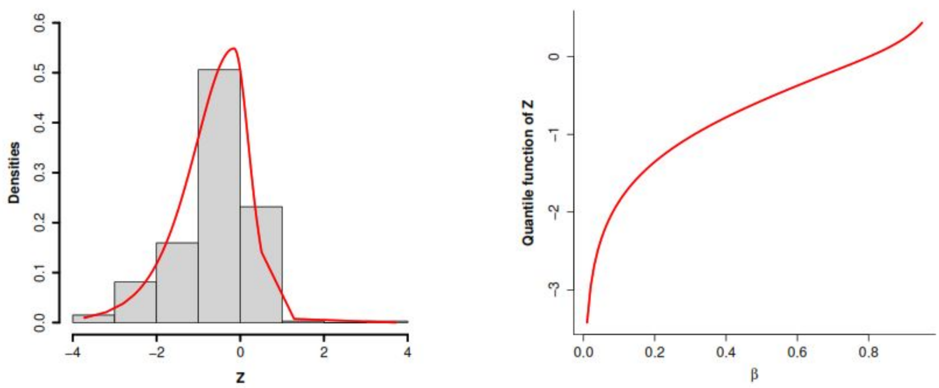

Fig. 4. (a). Histogram and fitted density estimate (solid red line) of the uniform logittype transformation of the proportion of daily SARS-CoV-2 infected people (left); (b). the estimated quantile function (right) the uniform logit-type transformation of the proportion of daily SARS-CoV-2 infected people

Now, the partial derivatives with respect to β_1, β_2 and β_3 is given by,

$$\frac{\delta^2 l}{\delta \beta_1\delta\beta_2} = K\mathbf{X}\sum_{i=1}^{n}\frac{e^{\mathbf{X}\beta_1}-Z_i}{e^{X\beta_2}} - 2\frac{\delta}{\delta\beta_1\beta_2}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{e^{\mathbf{X}\beta_1}-Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)$$
$$+K\mathbf{X}e^{\mathbf{X}\beta_3}\sum_{i=1}^{n}\frac{Z_i-e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_1\beta_2}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{(Z_i-e^{\mathbf{X}\beta_1})^{e^{\mathbf{X}\beta_3}}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big),$$

$$\frac{\delta^2 l}{\delta\beta_1\delta\beta_3} = \frac{-K\mathbf{X}e^{\mathbf{X}\beta_3}}{1+e^{\mathbf{X}\beta_3}}\sum_{i=1}^{n}\frac{e^{\mathbf{X}\beta_1}-Z_i}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_1\beta_3}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{e^{\mathbf{X}\beta_1}-Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)$$
$$+\frac{K\mathbf{X}e^{\mathbf{X}\beta_3}}{1+e^{\mathbf{X}\beta_3}}\sum_{i=1}^{n}\frac{Z_i-e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_1\beta_3}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{(Z_i-e^{\mathbf{X}\beta_1})^{e^{\mathbf{X}\beta_3}}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big),$$

$$\frac{\delta l}{\delta\beta_2\delta\beta_3} = \frac{-Ke^{\mathbf{X}\beta_3}}{1+e^{\mathbf{X}\beta_3}}\sum_{i=1}^{n}\frac{e^{\mathbf{X}\beta_1}-Z_i}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_2\beta_3}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{e^{\mathbf{X}\beta_1}-Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)$$
$$-\frac{K(e^{\mathbf{X}\beta_3})^2}{1+e^{\mathbf{X}\beta_3}}\sum_{i=1}^{n}\frac{Z_i-e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_2\beta_3}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{(Z_i-e^{\mathbf{X}\beta_1})^{e^{\mathbf{X}\beta_3}}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big),$$

$$\frac{\delta^2 l}{\delta{\beta_1}^2} = -K\mathbf{X}\sum_{i=1}^{n}\frac{e^{\mathbf{X}\beta_1}-Z_i}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_2\beta_3}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{e^{\mathbf{X}\beta_1}-Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)$$
$$-K\mathbf{X}e^{\mathbf{X}\beta_3}\sum_{i=1}^{n}\frac{Z_i-e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta\beta_2\beta_3}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{(Z_i-e^{\mathbf{X}\beta_1})^{e^{\mathbf{X}\beta_3}}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big),$$

$$\frac{\delta^2 l}{\delta{\beta_2}^2} = -K\mathbf{X}\sum_{i=1}^{n}\frac{e^{\mathbf{X}\beta_1}-Z_i}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta{\beta_2}^2}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{e^{\mathbf{X}\beta_1}-Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)$$
$$-K\mathbf{X}e^{\mathbf{X}\beta_3}\sum_{i=1}^{n}\frac{Z_i-e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta{\beta_2}^2}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{(Z_i-e^{\mathbf{X}\beta_1})^{e^{\mathbf{X}\beta_3}}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big),$$

$$\frac{\delta^2 l}{\delta{\beta_3}^2} = \frac{2Kn\mathbf{X}}{1+e^{\mathbf{X}\beta_3}} + \big[\frac{K\mathbf{X}e^{\mathbf{X}\beta_3}}{1+e^{\mathbf{X}\beta_3}} + \frac{K\mathbf{X}(e^{\mathbf{X}\beta_3})^2}{(1+e^{\mathbf{X}\beta_3})^2}\big]\sum_{i=1}^{n}\frac{e^{\mathbf{X}\beta_1}-Z_i}{e^{\mathbf{X}\beta_2}}$$
$$-2\frac{\delta}{\delta{\beta_2}^2}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{e^{\mathbf{X}\beta_1}-Z_i}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big) + \big[\frac{K\mathbf{X}e^{\mathbf{X}\beta_3}}{1+e^{\mathbf{X}\beta_3}} + \frac{K\mathbf{X}(e^{\mathbf{X}\beta_3})^2}{(1+e^{\mathbf{X}\beta_3})^2}\big]$$
$$\sum_{i=1}^{n}\frac{Z_i-e^{\mathbf{X}\beta_1}}{e^{\mathbf{X}\beta_2}} - 2\frac{\delta}{\delta{\beta_2}^2}\sum_{i=1}^{n}\ln\Big(1+\exp\big(-\frac{(Z_i-e^{\mathbf{X}\beta_1})^{e^{\mathbf{X}\beta_3}}}{(1+e^{\mathbf{X}\beta_3})e^{\mathbf{X}\beta_2}}\big)\Big),$$

where, $K = \frac{\mathbf{X}}{1+e^{\mathbf{X}\beta_3}}$.

It can be shown that, for large sample $\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \hat{\beta}_3\end{bmatrix} \sim N(\beta, \Sigma^{-1})$,

where,

$$\Sigma = \begin{bmatrix} \frac{\delta^2 l}{\delta{\beta_1}^2} & \frac{\delta^2 l}{\delta\beta_1\delta\beta_2} & \frac{\delta^2 l}{\delta\beta_1\delta\beta_3} \\ \frac{\delta^2 l}{\delta\beta_1\delta\beta_2} & \frac{\delta^2 l}{\delta{\beta_2}^2} & \frac{\delta^2 l}{\delta\beta_2\delta\beta_3} \\ \frac{\delta^2 l}{\delta\beta_1\delta\beta_3} & \frac{\delta^2 l}{\delta\beta_2\delta\beta_3} & \frac{\delta^2 l}{\delta{\beta_3}^2} \end{bmatrix}.$$

After differentiating the conditional log liklihood function, we have noticed that the derivatives are non linear and complex. That's why, for estimating parameter we have used a R package *QBAsyDist* which is introduced by [6]. In Section 3.2 we have applied a real data for illustrative purpose.

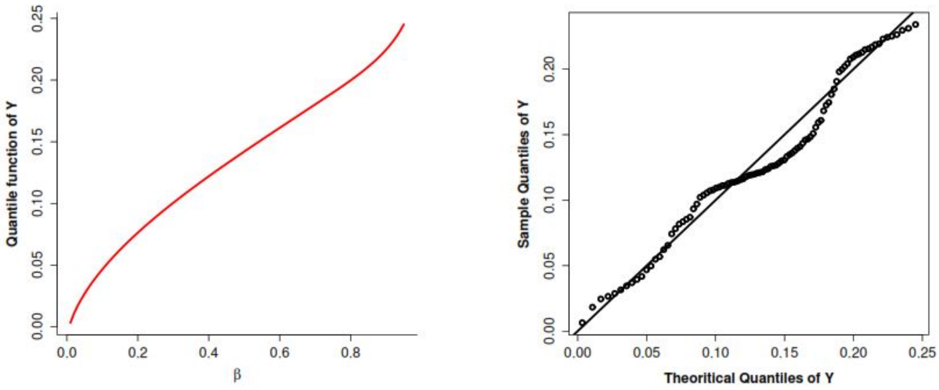

Fig. 5. (a). Estimated quantile function of Y ; (b). The QQ-plot (right)

3.2 Real data application

For illustrative purposes, we consider the daily proportion of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infected people who have tested for Coronavirus disease (COVID-19) infection from August 3, 2020 to February 12, 2021 in Bangladesh. The number of daily new SARS-CoV-2 infected cases and daily new tested peoples are reported by the Institute of Epidemiology Disease Control and Research (IEDCR), Dhaka, Bangladesh. The data are available on the website with web-link https://covid19.who.int/. It is observed that on average each day, 13.61% of peoples are infected who have tested for COVID-19 infection. Notice that the daily proportion of SARS-CoV-2 infected people (Y) is a bounded variable with support $[0, 1]$.

For the bounded random variable, we can not directly compute the quantile function of the distribution. Therefore, many authors including [8] and [9] used the uniform logit-type link function in quantile estimation (See Fig. 3 for the cdf and quantile function of Y.). That is the link function is $Z = \text{logit}(G(Y))$, where $G(y) = (y-a)/(b-a)$. We also consider this link function to estimate the quantile function of the proportion of daily SARS-CoV-2 infected cases among the people who have tested for COVID-19 infection on that day. In Section 2, we have shown that the distribution of Z is a quantile-based asymmetric logistic distribution given in (2.2). The data and quantile function of Z are presented in Fig. 4. The quantile function of Y and Q-Q plot are presented in Fig. 5.

For the uniform logit-type link function for this data set, we consider a as the minimum proportion of infected people minus k and b as the maximum infected people plus k, where k is very small number. In this case, we use $k = 0.01$. To add (subtract) a small value of k to b (a) to avoid the zero value of denomination (numerator) in the logit-type link function. The resulting link function is $z = g(y) = \ln\left(\frac{y-a}{b-y}\right)$ for $y \in (a, b)$. Using this link function $Z = g(Y)$, we estimate parameter $\theta = (\mu, \phi, \alpha)^T$ of the distribution of Z via the method of maximum likelihood estimation Figs. 4 and 5.

The daily proportion (Y) of SARS-CoV-2 infected people who have tested positive for COVID-19 infection is considered a response variable, and the daily temperature (X_1) and humidity (X_2) are considered covariates from August 3, 2020 to February 12, 2021 for illustrating the proposed method.

The data of daily temperature and humidity is available on the website https://www.timeanddate.com/weather/bangladesh/dhaka. In this case, the parametric functions can be written as

$$\theta_1(\mathbf{X}) = \mathbf{X}\beta_1 = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2,$$

$$\theta_2(\mathbf{X}) = \mathbf{X}\beta_2 = \exp(\beta_{20} + \beta_{21}X_1 + \beta_{22}X_2),$$
$$\theta_3(\mathbf{X}) = \mathbf{X}\beta_3 = \frac{\exp(\beta_{30} + \beta_{31}X_1 + \beta_{32}X_2)}{1 + \exp(\beta_{30} + \beta_{31}X_1 + \beta_{32}X_3)}.$$

Table 1. The summary statistics of the estimators for estimating $\mu(X;\beta_1)$, $\phi(X;\beta_2)$ and $\alpha(X;\beta_3)$ and the p-values obtained by using Bootstrapping

$\mu(X;\beta_1)$			$\phi(X;\beta_2)$			$\alpha(X;\beta_3)$		
β_1	$\widehat{\beta}_1\left(se(\widehat{\beta}_1)\right)$	P-value	β_2	$\widehat{\beta}_2\left(se(\widehat{\beta}_2)\right)$	P-value	β_3	$\widehat{\beta}_3\left(se(\widehat{\beta}_3)\right)$	P-value
β_{10}	2.3414(0.3994)	<0.005	β_{20}	1.0759(0.3160)	0.006	β_{30}	0.3947(0.5239)	<0.0001
β_{11}	0.7312(0.6943)	<0.0001	β_{21}	0.02031(0.2076)	<0.0001	β_{31}	0.0641(0.2431)	<0.0001
β_{12}	0.7514 (0.4129)	0.008	β_{22}	-0.0901(0.2836)	0.064	β_{32}	0.0166(0.2844)	0.424

The estimated parametric functions can be written as

$$\widehat{\mu}(X_i;\widehat{\beta}_1) = 2.3414 + 0.7312X_1 + 0.7514X_2,$$

$$\widehat{\phi}(X_i;\widehat{\beta}_2) = \exp(1.0759 + 0.02031X_1 - 0.0901X_2),$$

$$\widehat{\alpha}(X_i;\widehat{\beta}_3) = \frac{\exp(0.3947 + .0641X_1 + 0.0166X_2)}{1 + \exp(0.3947 + 0.0641X_1 + 0.0166X_3)}.$$

Table 1 shows the summary statistics of the estimated models. Regression coefficients for the temperature and humidity significantly impact the daily proportion of infected cases for the estimated function $\widehat{\mu}(X,\widehat{\beta}_1)$. For estimated $\widehat{\phi}(X,\widehat{\beta}_2)$, we see the regression coefficients for only temperature is statistically significant. Similarly, for $\widehat{\alpha}(X,\widehat{\beta}_3)$, we observe that temperature is highly statistically significant but humidity is not at 5% level of significance.

4 Concluding Remarks

In this research, we study the theory of quantile regression using a generalized quantile-based asymmetric family of densities. We provide the theory of logit-type link function for estimating quantile function in regression settings. In regression settings, we consider the response variable is the proportion of daily SARS-Cov-2 infected people tested for COVID-19 infection and two covariates: temperature and humidity. We noticed that the temperature and humidity have a significant impact on the proportion of daily SARS-Cov-2 infected persons tested for COVID-19 infection.

Declarations

- **Funding:** This research has been conducted with the financial support of a research grant from Jahangirnagar University.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Gijbels I, Karim R, Verhasselt A. Quantile estimation in a generalized asymmetric distributional setting. Steland, Ansgar and Rafajlowicz, Ewaryst and Okhrin, Ostap. Stochastic Models, Statistics and their Application. 2019b;Springer:13–40.

[2] Fan J, Gijbels I. Local Polynomial Modelling and Its Applications, Chapman & Hall/CRC; 1996.

[3] Koenker R. Quantile Regression. New York: Cambridge University Press; 2005.

[4] Komunjer I. Quasi-maximum likelihood estimation for conditional quantiles. J of Econometrics. 2005;128(1):137–164.

[5] Koenker R, Bassett Jr G. Regression quantiles. Econometrica: J of the Econometric Society. 1978;33–50.

[6] Gijbels I, Karim R, Verhasselt A. On Quantile-based Asymmetric Family of Distributions: Properties and Inference. International Statistical Review. 2019a;87:471-504.

[7] Karim R, and Haque S. Distribution of logit-type link function in a generalized quantile-based asymmetric distributional setting. Thailand Statistician (Accepted); 2022.

[8] Bottai M, Cai B, McKeown RE. Logistic quantile regression for bounded outcomes. Statistics in Medicine. 2010;29(2):309–317.

[9] Columbu S, Bottai M. Logistic Quantile Regression to Model Cognitive Impairment in Sardinian Cancer Patients in Di Battista T, Moreno E, Racugno W, Topics on Methodological and Applied Statistical Inference, Springer, 2016;65–73.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle5.com/review-history/92094