Parametric and Non-Parametric Model Performance Differences in Estimation of Critical Stream Flow Discharge Levels

Kiarie F. K.

1Department of Management Science, Kenyatta University, P.O Box 43844-00100, Nairobi

Abstract: Nonparametric estimation based on quantile regression methodology of Koenker and Basset (1978) and conventional parametric regression approaches were applied to a river regime to estimate volatility in streamflow discharge levels. Consistency and asymptotic normality properties of estimators obtained from both approaches were given. From the study results non-parametric quantile regression approach yielded better results than other methods. Other than for boundary effects which require boundary modifications, the model validation results implied good performance of the nonparametric model in estimating critical streamflow discharge levels.

Keywords: Kernel estimate, quantile autoregression, consistency, asymptotic normality, critical discharge level

1. Introduction

Flood processes, by their nature, are inherently complex, nonlinear and many times life threatening. As a natural phenomenon, floods, in real sense, cannot be completely controlled. Engineering designs for flood control structures such as flood control reservoirs and dykes are based on previous flood events. The levels of these events can, however, are exceeded by future floods. One way of reducing losses due to floods is by use of flood early warning systems (FEWS). Such a system consists of streamflow monitoring and forecasting as well as public information system. Various methods are employed in hydrological stream flow monitoring and forecasting. Statistical models for hydrological applications are usually based on regression relationships derived from paired catchment and catchment treatment experiments, see Muthusi (2004). For a response variable of interest, for instance, streamflow discharge levels, Yt given a covariate Xt in Ft-1, both parametric and nonparametric regression methods can be applied to estimate critical streamflow discharge levels. This paper examines the difference in model performances in estimation of critical streamflow discharge levels using both parametric and nonparametric regression approaches. In this paper nonparametric estimation focuses on quantile regression methodology introduced in Koenker and Basset (1978) while parametric estimation is based on the conventional mean-variance regression.

2. The Study Model (Critical Streamflow discharge level) for Nonparametric Estimation

Assume that the underlying hydrological process of interest is of the form

\[ Y_t = m_\theta (X_t) + \epsilon_t \]  (2.1)

where \( Y_t \) is the stream flow discharge at time \( t \) measured in cubic meters per second, (cubecs). The variable \( X_t = (Y_{t-1}, \ldots, Y_{t-d}) \) is a d-dimensional vector consisting of the past observations of \( Y_t \). The conditional quantile function \( m_\theta(X_t) \) is the streamflow discharge level at \( \theta \in (0, 1) \). The errors, \( \epsilon_t \), are assumed to be zero quantile with some scale function \( \zeta_\theta \).

Here, model (2.1) can be viewed as a robust generalization of Autoregressive (AR) – Autoregression Conditional Heteroscedastic (ARCH) models introduced in Weiss(1984) and their nonparametric generalizations reviewed by Hardle (1989), see Franke and Mwita (2003) and Mwita (2005) for more details.

If we choose \( X_t = (Y_{t-1}, \ldots, Y_{t-d}, U_{t-1}) \) where the random vector \( U_t \) consists of observations from other time series such as soil moisture budget(SMB), precipitation, evapotranspiration, El-Nino Southern Oscillations(ENSO), Pacific Decadal Oscillations(PDO), then model (2.1) would become a quantile autoregressive model with exogenous components.

2.1 Nonparametric Estimation of critical streamflow discharge level

We consider the model (2.1), and define a true conditional distribution function \( F(x) \) of \( Y_t \) given \( X_t = x \) as

\[ F(x) = P(Y_t \leq x \mid X_t = x) = E[1_{[y \leq x]} \mid X_t = x] \]  (2.2)

where \( 1_{[y \leq x]} \) is an indicator function with \( P(Y_t \leq y \mid X_t = x) = 1 \) and 0 otherwise.

For any \( \theta \in (0, 1) \), we define the true critical streamflow discharge level as

\[ m_\theta(x) = \inf\{ y \in \mathbb{R} \mid F(x) \geq \theta \} \]  (2.3)

The distribution function in (2.2) can be estimated by the Nadaraya (1964) and Watson (1964) estimator as
where $K(u)$ is a d-dimensional kernel and $K_h(u) = h^{-d} K(u/h)$ is the rescaled kernel, see Franke and Mwita (2003) and Mwita (2005).

Therefore the kernel estimator for the critical streamflow discharge level is given by

$$\hat{F}_x(y) = \frac{\sum_{i=1}^{n} K_h (x - X_i) I_{1,y}}{\sum_{i=1}^{n} K_h (x - X_i)} \quad (2.4)$$

where $K(u)$ is a d-dimensional kernel and $K_h(u) = h^{-d} K(u/h)$ is the rescaled kernel, see Franke and Mwita (2003) and Mwita (2005).

Theorem 3.1

Assume that (A1)- (A3) hold. As $n \to \infty$, let the sequence of bandwidths $h > 0$ converge to 0 such that $nhd \to \infty$. Then the conditional quantile estimator is consistent, i.e.,

$$\hat{m}_q(x) \to m_q(x)$$

that is

$$E[\hat{m}_q(x)] - m_q(x) = h^d B_m(m(x)) + O(h^2)$$

(2.6)

Further, if the bandwidths are chosen such that $nhd+4$ is either 1 or converges to 0, then

$$\hat{m}_q(x) \text{ is asymptotically normal,}$$

$$\sqrt{nh^d} (\hat{m}_q(x) - m(x) - h^d B_m(m(x))) \Rightarrow N \left(0, \frac{V(\hat{m}_q(x))}{f_x^2(m_q(x))} \right)$$

(2.7)

where, $B(y)$ and $V_2(y)$ are the bias and variance expansions for the conditional distribution estimator in (2.4).

2.2 Asymptotic Normality

Assume that the time series (Yt, Xt) satisfies $\alpha$-mixing conditions. According to Marsy and Tjostheim (1995, 1997), both ARCH processes and nonlinear additive autoregressive models with exogenous variables are stationary and $\alpha$-mixing under some mild conditions. As Franke and Mwita (2003) demonstrated, if we choose $X_t = Y_t - d$ in (2.1) and assume the time series $Y_t$ is $\alpha$-mixing, we get an example of a quantile autoregressive process for which $(Y_t, X_t)$ and $(X_t, Y_t)$ are $\alpha$-mixing as well.

The following assumptions are necessary for proving asymptotic normality of $\hat{m}_q(x)$

**Henceforth,** $g(x)$ denotes the stationary probability density of $X_t$ at point $x$.

(A1) For all $u \in \mathbb{R}$

$K(u) \geq 0$

K is Lipschitz continuous i.e. $|K(u) - K(v)| \leq C K|u - v|$, for all $C K, u, v \in \mathbb{R}$ and $C K > 0$

$K(u) \leq K_{\infty}$, with $K_{\infty}$ being a constant

$K(u)du = 1, \int K(u)du = 0$ and $\int u K(u)du < \infty$

(A2) For all $y, x$ satisfying $0 < F(x) < 1$, $g(x) > 0$

1) $F(x)$ and $g(x)$ are twice continuously differentiable and bounded in $y, x$

$F(x)g(x) > 0$, for all $x$.

2) (A3) The process $(Y_t, X_t)$ is stationary and $\alpha$-mixing with mixing coefficients $\alpha(s) = O(s^{-2 + \delta})$ for some $\delta > 0$, $\alpha > 1$, and $\{s_n\}$ is an increasing sequence of positive integers.

The consistency and asymptotic normality properties of $\hat{m}_q(x)$, and their proofs can be found in Franke and Mwita (2003).

Here, we only state the theorems.

Theorem 3.2

Assume (A1), (A2), (B1), and (B2). If, as $n \to \infty$, the bandwidth $h \to 0$ such that

$$\hat{S}_n = nh^d (s_n \log n)^{-1} \to \infty$$

then (3.2.4) is uniformly consistent on $G$ in the strong sense.

That is, for $x \in G$

$$\sup_{x \in G} \frac{F(x) - F(x)}{s_n} \to 0$$

as $n \to \infty$.

In this section, we have shown that the estimate of our nonparametric quantile function is consistent and asymptotically normally distributed, and under suitable conditions, the estimator converges uniformly with an appropriate rate. The asymptotic normality property is used to construct the required confidence intervals for our estimator. These are strong properties that significantly imply sufficiency of the estimator is accurate in estimation of the critical streamflow discharge level.
3. Real Data Results

The application of the estimator was performed with data from the gauge at River Nyando, in Western Kenya, (River Station No. IGD03) in the wider Nyando Basin, located at 35.2°E longitude and -0.1°S latitude and covering an area of 3, 587 km². The drainage area downstream of the outlet of the catchment (IGD03) was found to accommodate all the discharge in the river channel. For this reason, monthly maximum streamflow data from gauging station IGD03 for the period 1970 – 1997 was used for calibrating the model. Also, the twenty-seven year period was considered long enough to capture diverse weather conditions, thus making the model to be a good representative of the basin.

Figure 3.1 gives the daily streamflow hydrograph of ground station gauging data for twenty-seven years, from 1970 – 1997. From the hydrograph, it is clear that the river regime experiences both peak and extremely high flows which are responsible for flood inundations experienced in flood plain areas of the Nyando basin.

Considering the critical streamflow discharge level to be our target variable, we first present hydrograph for monthly maximum streamflow for the period 1970 – 1997 in Figure 3.2.

The dotted curve represents the 0.95 conditional quantile while the dashed curve represents the 0.99 conditional quantile. Streamflow discharge levels above the 0.95-quantile curve represent critical streamflow discharge levels responsible for flood inundations at 95% confidence level. Such a level calls for some site-specific operational instructions to be issued by authorities monitoring the river catchment. The instructions may include shutting of floodgates and other engineering measures. Discharges above the 0.99 quantile curve represent extreme river flow levels. Such levels call for flood control teams to respond to imminent flood conditions and operate a warning system for the public as well as industries.

Figure 3.3 gives the volatility of the monthly maximum streamflow discharges. The hydrograph of these deviations depict the turbulence experienced by the Nyando River regime with an observable increase in trend.

Figures 3.4 gives the monthly maximum streamflow discharge levels together with 0.95 and 0.99 conditional quantiles respectively.
3.1 Parametric Estimation of critical stream flow discharge level

To compare model performances and the quality of our estimator quantitatively, a comparative assessment was carried out with two different assumptions on the river flow data.

3.1.1 Mean-Variance method

Under this method we assume normality of the streamflow discharge data. Our hydrologic model is then modified to take the following form of a quantile autoregressive-heteroscedastic process

\[ \text{Y}_t = \mu(X_t) + \zeta(X_t)\text{e}_t \]

where \( \mu(X_t) \) is the conditional mean of \( \text{Y}_t \) given \( X_t \) and \( \zeta(X_t) \) is the conditional standard deviation of \( \text{Y}_t \) given \( X_t \).

Using the Nadaraya-Watson estimator of \( \mu(X_t) \)

\[ \hat{\mu}(x) = \sum_{t=1}^{n} K_h(x-X_t)Y_t \]

and the estimator of the variance

\[ \hat{\sigma}^2(x) = \sum_{t=1}^{n} K_h(x-X_t)(Y_t-\hat{\mu}(x))^2 \]

Therefore, the estimator of the innovations \( \text{e}_t \) is given by

\[ \hat{\epsilon}_t = \frac{(Y_t-\hat{\mu}(x))}{\hat{\sigma}(x)} \]

The conditional quantile function estimator of \( m_\theta(X_t) \) using (3.1) becomes,

\[ \hat{m}_\theta(x) = \hat{\mu}(x) + \hat{\sigma}(x)\hat{\epsilon}_\theta \]

Where \( \theta \) are the true quantiles of \( \epsilon_t \sim N(0, 1) \) i.i.d random variables.

3.1.1.1 Asymptotic properties of the mean-variance estimator

The consistency of the conditional mean and conditional variance estimator is shown in the following proposition, see Hardle (1989).

**Proposition 3.1.1.2**

Assume the stochastic design model with a one-dimensional predictor variable \( X \) and define

\[ \lim_{n \to \infty} uK(u) = 0 \]

\[ \lim_{|u| \to \infty} uK(u) = 0 \]

\[ \text{E}Y^2 < \infty \]

\[ \text{hn} \to 0, \text{hn} \to \infty \text{ as } n \to \infty, \]

Then, at every point of continuity of \( \mu(x) \), \( f(x) \) and \( \sigma^2(x) \), with \( f(x) > 0 \)

\[ \hat{\mu}(x) \to \mu(x) \]

For convergence, we first define the mean squared error at a point \( x \) as follows

\[ d_M(x,h) = E[\hat{\mu}_h(x) - m(x)]^2 \]

The following theorem gives the speed of \( dM(x, h) \) as a function of \( h \) and \( n \). See Hardle (1989).

**Theorem 3.1.2**

Assume the random design model with a one-dimensional predictor variable \( X \) and define \( CK = \int K_2(u)du \), \( dK = \int u2K(u)du \).

Assume

\[ (F0) K \text{ has support } [-1, 1] \text{ with } K(-1) = K(1) = 0, \]

\[ (F1) \mu \in C_2, \]

\[ (F2) \max |X_i - X_j| = O(n^{-1}), \]

\[ (F3) \text{var}(X_i) = \sigma^2, i = 1, \ldots, n, \]

\[ (F4) n \to \infty, h \to 0, nh \to \infty. \]

Then

\[ dM(x, h) \approx (nh-1)\sigma^2ck + h^4dK_2[\mu''(x)]^2/4. \]

The mean squared error splits up into the two parts, variance and bias. The theorem shows that the variance and bias goes to zero as \( n \to \infty \), and \( h \to 0 \), respectively.

3.1.2 Historical simulation

Here, we assume independence and identical distribution (iid) of the streamflow discharge data.

Let \( Y_t, t = 1, 2, \ldots \ldots \) be a sequence of iid random variables, with a distribution function \( F(y) \), now to represent streamflow discharge level at time \( t \).

We denote \( Y_t = Y \) since \( Y_t \) are iid. For some probability \( \theta \in (0, 1) \), consider the \( \theta \)-quantile as

\[ y_{\theta} = \inf\{y \in R \mid F(y) \geq \theta \} \]

If \( F \) is continuous, then \( P(y = y_{\theta}) = 0 \) and \( F(y_{\theta}) = 0 \), while if \( F \) is discontinuous in \( y_{\theta} \), then \( P(y = y_{\theta}) > 0 \) and \( F(y_{\theta}) = 0 \).

Consider the order statistics \( y_1, n \leq y_2, n \leq \ldots \ldots \leq y_{k+1}, n \leq \ldots \ldots \leq y_{n} \), \( n \) is the sorted values of \( n \)-tuple \( (y_1, y_2, \ldots, y_n) \) and let \( k = [n(1-\theta)] \) be the integer part of \( n(1-\theta) \). The set of observations which constitute the 100(1-\( \theta \))% largest of the total values in the sample is represented by the largest \( k \) observations (outcomes)

\[ \{y_k, n \leq \ldots \leq y_1, n\}. \]

As usual, \( y_{k+1}, n \) denotes the empirical quantile which we may write as, \( q_{\theta}y_n \), where \( \theta \) stands for the proportion of observations below \( y_k, n \).

3.1.2.1 Asymptotic properties of \( y_{k+1}, n \) under the iid assumption

We assume that \( F \) has a density function, \( f \) and \( P( y_{k+1}, n \geq y_{k+1}, n) = 1 \).

With \( n-k = n\theta + o(n(1/2)) \), i.e \( k/n = 1 - \theta + o(n^{-1/2}) \), then by central limit theorem

\[ \theta(1-\theta) \]

\[ q_{\theta}y_n \sim N(q_{\theta}y, \ldots \ldots \ldots) \]

\[ n^{2/3}(q_{\theta}y) \]

where \( q_{\theta}y \) is the asymptotic quantile and \( \theta(1-\theta) \)
being the asymptotic variance, \( n \sigma^2(q \theta y) \)

See Mwita (2003) and references therein. This asymptotic normality property is used to construct the required confidence intervals for the estimator.

Figure 3.5 gives the 0.95 conditional quantiles obtained under the two assumptions (mean-variance and i.i.d) together with the 0.95-th quantile using model (2.1). The dotted straight line represents the 0.95-th quantile under iid assumption. The dotted curve represents the 0.95-th quantile under model (2.1) while the dashed curve (not visibly clear) with a similar pattern to the solid (actual streamflow discharge curve) represents the 0.95 conditional quantile under the assumption of normality. From the graph, both the iid and mean-variance methods appeared to underestimate the critical streamflow discharge level and therefore performing poorer than model (2.1) at the 95% level.

3.5: 0.95 quantile estimates Under mean variance and iid assumptions together with0.95 quantile under QAR.

Consequently, the nonparametric model is considered to be adequate and better in performance in the estimations of critical streamflow discharge levels.

4. Conclusion

In this paper, a nonparametric quantile regression approach was considered. By avoiding assumptions on the form of the conditional distribution of the streamflow discharge in a river regime, our method of estimating critical streamflow discharge level yields better results than other methods, which make assumptions on the underlying conditional distribution function of the streamflow discharges. Furthermore, the critical streamflow discharge level estimates obtained from the study model fits the quantile data well apart from the extreme ends, which is attributed to data scarcity at the extremes. However, the estimator obtained from our method suffers from boundary effects and therefore requires boundary modifications, see Mwita (2003). Other than the singular drawback of boundary effects, the study model’s validation results implies a good performance of the model in estimation of critical streamflow discharge levels.

References