8-2014

QUANTILE REGRESSION FOR CLIMATE DATA

Dilhani Marasinghe
Clemson University, dmarasi@clemson.edu

QUANTILE REGRESSION FOR CLIMATE DATA

A Master Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
MASTER OF SCIENCE
Mathematical Sciences

by
DILHANI SHALIKA MARASINGHE
August 2014

Accepted by:
Dr. Collin Gallagher, Committee Chair
Dr. Christoper McMahan
Dr. Robert Lund

Abstract

Quantile regression is a developing statistical tool which is used to explain the relationship between response and predictor variables. This thesis describes two examples of climatology using quantile regression. Our main goal is to estimate derivatives of a conditional mean and/or conditional quantile function. We introduce a method to handle autocorrelation in the framework of quantile regression and used it with the temperature data. Also we explain some properties of the tornado data which is non-normally distributed. Even though quantile regression provides a more comprehensive view, when talking about residuals with the normality and the constant variance assumption, we would prefer least square regression for our temperature analysis. When dealing with the non-normality and non constant variance assumption, quantile regression is a better candidate for the estimation of the derivative.

Keywords: Quantile Regression, Conditional Quantile Function, Derivative, Autocorrelation

Acknowledgments

I would like to offer my heartfelt gratitude to my advisor Dr. Collin Gallagher, who gave me an excellent support throughout my thesis with great motivation. I am also grateful to my thesis committee Dr. Robert Lund and Dr. Chris McMahan for dedicating their time to read my thesis and for the valuable comments. I will be forever thankful to my friend Javier Ruiz-Ramírez for helping me immensely by sharing his knowledge. I warmly thank my parents for their encouragement and great blessings. Finally, a special thank goes to my husband Pubudu Lakmal, who is always nearby me with understanding and good caring.

Table of Contents

List of Figures

Chapter 1

Introduction

The weather is a very complex system in which a series of changes take place in a short period of time. Climate is merely the average weather over some larger time scale, usually in months or years [9]. It is measured by using averages of weather elements like temperature, precipitation, humidity and barometric pressure. Variations of these elements give rise to several phenomena that impact our daily lives. In this work, we will use data sets related to the study of temperature anomalies and tornadoes. Evidence of these anomalies have been recorded by research groups, namely, NASA Goddard Institute for Space Studies (GISS), NOAA National Climatic Data center and UK Met Office Hadley Centre [31]. These studies have shown that for example the average temperature across global land and ocean surfaces in 2013, compared with the base period that comprises the years 1901-2000 has raised 1.12 °F (0.62°C), a quantity which is significant in the field of global warming [33]. Moreover, according to NOAA's Storm Prediction Center, the tornado count of 856 for 1989 rose to 891 for 2013 [34].

In recent years, quantile regression has been widely used in the field of statistics, since it provides a more comprehensive view on the relationship of response and predictor variables [23, 26, 43]. However, climate studies mostly focus on average. The ordinary least square method (OLS) estimates the relationship between predictor and response variables by using the conditional mean function while quantile regression models explain that relationship using the conditional quantile function (see Section 1.2); quantile regression methods can detect more subtle relationships between independent and dependent variables and allow for potential heteroskedasticity.

In this thesis, we will be concerned in estimating derivatives of a conditional mean and/or conditional quantile function. Our work considers both, a parametric and non-parametric approach. In the parametric approach we use the bootstrap technique for deriving confidence intervals of the derivative of a quantile regression model. For the non-parametric method we use local polynomial quantile regression.

1.1 Quantiles

Let $F(x) = P(X \leq x)$ be the cumulative distribution function (CDF) and $f(x)$ the probability density function of a random variable X. The τ-th quantile is defined as

$$Q(\tau) = F^{-1}(\tau) = \inf \{x : F(x) \geq \tau\}, \tag{1.1}$$

where $\tau \in (0, 1)$ [27]. The median $Q(1/2)$, is a special case of quantiles. The asymmetric quantile loss function is defined as

$$\rho_\tau(u) = u(\tau - I(u < 0)), \tag{1.2}$$

which is illustrated in Fig. 1.1. Here $I(\cdot)$ is the indicator function.

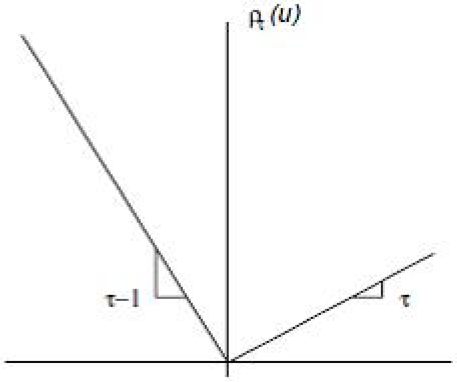


Figure 1.1: The Quantile Loss Function

In order to find the quantiles, we minimize $E(\rho_\tau(X - \xi))$ with respect to ξ.

$$
\begin{aligned}
E\left(\rho_\tau(X-\xi)\right) &= \int_{-\infty}^{+\infty} \rho_\tau(X-\xi)\ dF(x) \\
&= (\tau-1)\int_{-\infty}^{\xi} (x-\xi)\ dF(x) + \tau \int_{\xi}^{+\infty} (x-\xi)\ dF(x)
\end{aligned}
$$

Differentiating this expectation with respect to ξ,

$$
\begin{aligned}
&= \frac{d}{d\xi}\left[(\tau-1)\int_{-\infty}^{\xi}(x-\xi)\,dF(x) + \tau\int_{\xi}^{+\infty}(x-\xi)\,dF(x)\right] \\
&= \frac{d}{d\xi}\left[(\tau-1)\left(\int_{-\infty}^{\xi} x\,dF(x) - \xi\int_{-\infty}^{\xi} dF(x)\right) - \tau\left(\int_{+\infty}^{\xi} x\,dF(x) - \xi\int_{+\infty}^{\xi} dF(x)\right)\right] \\
&= (\tau-1)\left(\xi f(\xi) - \xi f(\xi) - 1.\int_{-\infty}^{\xi} dF(x)\right) - \tau\left(\xi f(\xi) - \xi f(\xi) - 1.\int_{+\infty}^{\xi} dF(x)\right) \\
&= (\tau-1)(-F(\xi)) - \tau(1-F(\xi)) \\
&= F(\xi) - \tau
\end{aligned}
$$

and finding the unique ξ that satisfies $F(\xi)-\tau=0$, gives us the minimum value. This claim is based on the fact that the second derivative of $E\left(\rho_\tau(X-\xi)\right)$ is the probability density function $f(\xi)$, which is a non-negative function. Hence, minimizing the quantile loss function applied to residuals leads us to an estimation of the quantiles of the response variable. In general, the distribution function $F(x)$ is unknown. Thus, we estimate the distribution function using the empirical CDF which is computed using sample observations.

$$
F_n(x) = \sum_{i=1}^{n} I(x_i \leq x)
$$

Then we minimize the expectation with the empirical distribution.

$$
\begin{aligned}
E\left(\rho_\tau(X-\xi)\right) &= \int_{-\infty}^{+\infty} \rho_\tau(X-\xi)\ dF_n(x) \\
&= \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(X-\xi)
\end{aligned}
$$

Since $1/n$ is a constant, minimizing the above expectation is the same as minimizing

$$\sum_{i=1}^{n} \rho_\tau(X - \xi).$$

Let $R(\xi) = \sum_{i=1}^{n} \rho_\tau(X - \xi)$. Suppose the optimal occurs at a point $\hat{\xi}$. This happens when the left and right derivatives of R are both non-negative at the point $\hat{\xi}$. In summary, the quantiles can be expressed as the solution to an optimization problem. This leads us to a more general method of estimating models of conditional quantile functions.

1.2 Quantile Regression

Quantile regression was introduced by Koenker and Bassett [23]. It provides more robust and efficient estimators compared to OLS and it does not make distributional assumptions on the error term in the model.

Consider the following simple linear mean regression model,

$$Y = X^T\beta + \epsilon, \tag{1.3}$$

with $E(\epsilon) = 0$. Therefore, $E(Y|X = x) = x^T\beta$. Here, β explains the change in the mean of the response variable Y due to a small change in x. In OLS regression β is estimated by solving,

$$\hat{\beta} = \underset{\{\beta\in\mathbb{R}^p\}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T\beta)^2. \tag{1.4}$$

A similar approach can be applied to estimate regression quantiles.

The τ-th conditional quantile function is defined as

$$Q(\tau|x) = x^T\beta(\tau), \tag{1.5}$$

for $\tau \in (0, 1)$. Here $\beta(\tau) = (\beta_1(\tau), \beta_2(\tau), \ldots, \beta_p(\tau))^T$ is the quantile coefficient vector. Thus, $Q(\tau|x) = \beta_1(\tau)x_1 + \beta_2(\tau)x_2 + \ldots, +\beta_p(\tau)x_p$, where $\beta_k(\tau)$, for $k = 1, 2, ..., p$ measures the change in the τ-th quantile of Y with respect to x_k.

Now, define the quantile regression model

$$y_i = x_i^T \beta(\tau) + \epsilon_i(\tau), \tag{1.6}$$

where $P(\epsilon_i(\tau) < 0) = \tau$. Analogous to (1.4), $\beta(\tau)$ can be estimated by solving the optimization problem,

$$\hat{\beta}(\tau) = \underset{\{\beta \in \mathbb{R}^p\}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta), \tag{1.7}$$

where $\rho_\tau(\cdot)$ is the quantile loss function defined in (1.2).

In order to find $\hat{\beta}(\tau)$, we rewrite the quantile regression model as,

$$\begin{aligned} y_i &= x_i^T \beta(\tau) + \epsilon_i(\tau) \\ &= x_i^T \beta(\tau) + (u_i - v_i), \end{aligned}$$

by introducing $2n$ artificial variables $u_i, v_i, i = 1, \ldots, n$ where $u_i = \epsilon_i I(\epsilon_i > 0)$ and $v_i = |\epsilon_i| I(\epsilon_i < 0)$ [27] , i.e. the residual vector splits into positive and negative parts. As a consequence, the problem introduced in (1.7) becomes,

$$\begin{aligned} \min_{\{\beta \in \mathbb{R}^p\}} \quad & \tau 1_n^T \mathrm{u} + (1-\tau) 1_n^T \mathrm{v} \\ \text{subject to} \quad & y - X^T \beta = \mathrm{u} - \mathrm{v} \\ & \mathrm{u} \geq 0, \mathrm{v} \geq 0 \end{aligned}$$

which has been solved using the Simplex algorithm [24], the Frisch-Newton interior point method and the Interior method with preprocessing. Among them, the simplex method is usually preferred.

1.3 Asymptotic Results

Recall the quantile regression function in (1.7). Quantile regression estimators are consistent, i.e. $\|\hat{\beta}_n(\tau) - \beta(\tau)\| \to 0$ in probability as $n \to \infty$, assuming the following regularity conditions:

1. The conditional distribution functions $F(Y|x_i)$ are absolutely continuous with continuous densities $f(Y|x_i)$ which are uniformly bounded away from 0 and ∞ at the τ-th quantile.

2. Q_0 and Q_1 are positive definite matrices such that,

(a) $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} x_i x_i^T = Q_0$

(b) $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} f_i^2(F^{-1}(\tau))x_i x_i^T = Q_1$

(c) $\max\limits_{i=1,\dots,n} \|x_i\|/\sqrt{n} \to 0$

Under the above conditions we have two scenarios for the asymptotic normal distribution of regression quantiles [27]. The first case is when the errors are independent and identically distributed,

$$\sqrt{n}\left(\hat{\beta}_n(\tau) - \beta(\tau)\right) \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f^2(F^{-1}(\tau))}Q_0^{-1}\right),$$

and the second when the errors are independent, but not identically distributed,

$$\sqrt{n}\left(\hat{\beta}_n(\tau) - \beta(\tau)\right) \xrightarrow{d} N\left(0, \tau(1-\tau)Q_1^{-1}Q_0Q_1^{-1}\right),$$

The asymptotic covariance between quantiles τ_i and τ_j is

$$Acov\left(\sqrt{n}\left(\hat{\beta}_n(\tau_i) - \beta(\tau_i)\right), \sqrt{n}\left(\hat{\beta}_n(\tau_j) - \beta(\tau_j)\right)\right) = \left(\tau_i \wedge \tau_j - \tau_i\tau_j\right) Q_1(\tau_i)^{-1}Q_0Q_1(\tau_j)^{-1},$$

In order to do statistical inference based on the asymptotic distribution of regression quantiles, first we have to estimate the covariance matrix. Under the iid assumption for the errors, the covariance matrix is

$$var\left(\sqrt{n}\hat{\beta}(\tau)\right) = \frac{\tau(1-\tau)}{\hat{f}^2(F^{-1}(\tau))}\hat{Q}_0^{-1},$$

where $\hat{Q}_0 = n^{-1}\sum_{i=1}^{n} x_i x_i^T$. The sparsity function $s(\tau) = \dfrac{1}{f(F^{-1}(\tau))}$ is estimated using the difference quotient of the empirical distribution function

$$\hat{s}_n(\tau) = \frac{\hat{F}_n^{-1}(\tau + h_n|\bar{x}) - \hat{F}_n^{-1}(\tau - h_n|\bar{x})}{2h_n},$$

where $\hat{F}_n^{-1}(\tau|\bar{x})$ provides the estimated conditional quantile of of Y given $\bar{x}$, $\bar{x}$ is the sample mean given by $\dfrac{\sum_{i=1}^{n} x_i}{n}$ and h_n is the bandwidth parameter where $h_n \to 0$ as $n \to \infty$.

Under non-iid error setting, the covariance matrix is

$$var\left(\sqrt{n}\hat{\beta}(\tau)\right) = \tau(1-\tau)\hat{Q}_1^{-1}\hat{Q}_0\hat{Q}_1^{-1},$$

where $\hat{Q}_1 = n^{-1} \sum_{i=1}^{n} \hat{f}_i^2(F^{-1}(\tau))x_i x_i^T$ and

$$\hat{f}_i(F^{-1}(\tau)) = \frac{2h_n}{x_i^T\hat{\beta}(\tau + h_n) - x_i^T\hat{\beta}(\tau - h_n)}$$

The bandwidth parameter h_n can be computed using the Bofinger method or the Hall and Sheather method [18]. For further reference read Chapter 4 of Koenker [27].

Once we estimate the covariance matrix the next step is to construct the confidence intervals. In order to do that, methods such as the Sparsity method, Rank score test and resampling techniques can be used for that purpose. One of the resampling methods, Bootstrap, is explained in the next section.

1.4 Bootstrapping

The bootstrap method is a general resampling procedure introduced by Efron [10] as a computer based method for estimating the distributions of statistics using the observations of the sample. It has many advantages. Although it provides inconsistent estimators in some occasions, it does not require distributional assumptions like normality. Bootstrap provides more accurate inferences even when the sample size is small. It can apply to statistics with sampling distributions that are difficult to derive asymptotically. The main characteristic of this method is that it generates a large number of repeated samples with replacement from the original sample in order to obtain a good estimate of the sampling distribution of interest. Once we know the sampling distribution of our statistic, we can find standard errors and confidence intervals for estimates such as mean, median [21] and regression coefficients [15]. Efron [11] has considered setting approximate confidence intervals for a single parameter θ in a parametric and non-parametric scenario. Later on, Efron and Tibshirani [12] worked together in deriving confidence intervals for time series data structures. Recent work shows that the bootstrap method can be implemented to construct confidence intervals for quantile regression estimates. Hahn [17] worked on bootstrapping quantile regression estimators and he showed that the constructed confidence intervals have asymptotically correct coverage probabilities. This work deals with a special method called the sieve bootstrap studied by Bühlmann [2] for time series data. In sieve bootstrap, the basic idea is to fit a parametric model first and then resample from the residuals. The algorithm is formally described as follows: Let $x_1, \ldots, x_n$ be a sample from a stationary process $\{x_t\}_{t\in\mathbb{Z}}$.

1. First we fit an autoregressive model of order p which is given by,

$$x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + z_t \quad t \in \mathbb{Z}, \tag{1.8}$$

2. Estimate $\hat{\phi}_1, \ldots, \hat{\phi}_n$ corresponding to the model (1.8). The residuals are then computed with

$$x_t = \sum_{j=1}^{p} \hat{\phi}_j x_{t-j} + \hat{z}_t. \tag{1.9}$$

3. Construct the resampling based on autoregressive residuals. For any $t \in \mathbb{Z}$, $z_t^* \overset{iid}{\sim} \hat{F}_z$, where F_z is the empirical CDF of $\hat{z}_t$. Define $\{x_t^*\}_{t\in\mathbb{Z}}$ by the recursion formula,

$$x_t^* = \sum_{j=1}^{p} \hat{\phi}_j x_{t-j}^* + \hat{z}_t^*. \tag{1.10}$$

4. Now consider any statistic $T_n = T_n(x_1, \ldots, x_n)$. Then we can define the bootstrapped statistic T_n^* by

$$T_n^* = T_n(x_1^*, \ldots, x_n^*). \tag{1.11}$$

For further reference of the bootstrap methods, we recommend the book of Davison and Hinkley [8].

1.5 Local Linear Quantile Regression

Local Linear Quantile Regression is an important non-parametric tool used for smoothing quantile regression curves. It is also useful in estimating derivatives of a particular estimate. The idea of smoothing by local regression was studied by Rosenblatt [37] and Parzen [36] with kernel density estimation methods. More general works on local regression have been written by Stone [39] and Cleveland [5]. Cleveland and Devlin [6] applied local linear and quadratic fitting to multivariate data. Local linear regression has been used as a basis for constructing projection pursuit estimates by Friedman and Stuetzle [15]. Hastie and Tibshirani [20] used local regression in additive models. A more detailed treatment on local regression can be found in Cleveland and Loader [7]. Local regression, together with quantile regression provide us information about smooth quantile curves. Some recent work on non-parametric estimation of conditional quantile functions can be found in Bhattacharya & Gangopadhyay [1], Koenker et al [25] and Chaudhuri [3]. This thesis focuses on estimating derivatives of a conditional quantile function. Chaudhuri [3] discussed the asymptotic behaviour of regression quantiles and Chaudhuri et al [4] applied those results in estimating average derivatives on local quantile regression. Suppose the sample $\{(x_i, y_i); i = 1, \ldots, n\}$ follows the model

$$y_i = m_\tau(x_i) + \epsilon_i(\tau), \tag{1.12}$$

where $m_\tau(\cdot)$ is an unknown function and x is uni-dimensional predictor . The quantile function $m_\tau(x)$ can be locally approximated with a polynomial by using a Taylor expansion in the neighborhood of x,

$$m_\tau(x_i) \approx \sum_{j=0}^{p} \frac{m_\tau^j(x)}{j!}(x_i - x)^j \equiv \widetilde{X}_i^T \beta_\tau,$$

where m_τ^j is the j-th derivative of m_τ and $\widetilde{X}_i = (1, (x_i - x), (x_i - x)^2, \ldots, (x_i - x)^p)^T$, $\beta_\tau = (\beta_{0\tau}, \beta_{1\tau}, \beta_{2\tau}, \ldots, \beta_{p\tau})$. Then the function is estimated by

$$\hat{m}_\tau(x) = \hat{\beta}_{0\tau}, \tag{1.13}$$

and the first derivative is given by

$$\hat{m}'_\tau(x) = \hat{\beta}_{1\tau}. \tag{1.14}$$

Then, the local polynomial quantile regression estimates, β_τ, are solved by using a weighted objective function

$$\underset{\{\beta \in \mathbb{R}^p\}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i(x) \rho_\tau(y_i - \widetilde{X}_i^T \beta), \tag{1.15}$$

where $w_i(x) = K((x_i - x)/h)$ with K as the bounded kernel function and h the bandwidth parameter.

Local composite quantile regression(CQR), proposed by Kai et al [22], is a new non-parametric regression method which provides more efficient estimators compared to the local linear estimators. In order to find the bandwidth, the authors initialize their method using a generalized version of the quantile loss function (1.2).

$$\rho_{\tau_k} u = u(\tau_k - I(u < 0)),$$

$k = 1, 2, \ldots, q$, with q loss functions and $\tau_k = k/(q+1)$.

Then the locally weighted CQR loss function is defined as follows,

$$\text{argmin} \sum_{k=1}^{q} \left[\sum_{i=1}^{n} w_i(x) \rho_{\tau_k}(y_i - \widetilde{X}_i^T \beta) \right], \tag{1.16}$$

where $w_i(x) = K((x_i - x)/h)$ and $\widetilde{X}_i = (1, (x_i - x), (x_i - x)^2, \ldots, (x_i - x)^p)^T$, $\beta_\tau = (\beta_{0\tau}, \beta_{1\tau}, \beta_{2\tau}, \ldots, \beta_{p\tau})$. The previously introduced method has been used to estimate the function $m(x) = E(Y|X = x)$ and the derivative of the function, $m'(x)$. Zheng et al [45] generalize CQR to allow for optimal data based weights as opposed to the equal weighting scheme of Kai et al [22].

1.6 Bandwidth Selection

Bandwidth is interpreted as a degree of smoothness of a curve. Choosing the optimal bandwidth is highly important in non-parametric regression. There are several methods for bandwidth selection in non-parametric mean regression, namely, plug-in, rule-of-thumbs and cross validation. These procedures, usually find an asymptotic optimal bandwidth by minimizing the Mean Square Error(MSE) or Mean Integrated Squared Error (MISE). However, like other methods, classical techniques for bandwidth selection have been extended to the field of quantile regression.

Abberger (1996) adjusted the cross validation to kernel quantile regression replacing the squared loss criterion by the quantile loss function defined in (1.2). Therein, the following formula was used:

$$CV(h) = \sum_{i=1}^{n} \rho_\tau(Y_i - Q_n^{(-i)}(\tau|x_i)), \tag{1.17}$$

where $Q_n^{(-i)}(\tau|x_i)$ is the leave-one-out estimator for the conditional quantile estimate $Q_n(\tau|x_i)$ defined in (1.5). One defect of the cross validation procedure is that it has a low relative convergence rate, namely $\mathcal{O}\left(n^{-1/10}\right)$ [30].

Yu & Jones [42] presented a rule-of-thumb based on plug-in idea for selecting regression quantile smoothing parameters. They considered minimizing a local linear quantile function according to (1.15) with p = 1. Let f be the marginal density of X , $Q(\tau|x)$ be the conditional quantile estimate and $g(H(Y)|X = x)$ be the conditional density of some function $H(Y)$ based on τ. Then, the optimal bandwidth is

$$h_\tau^5 = \frac{R(K)\tau(1-\tau)}{n\mu_2(K)^2 Q''(\tau|x)^2 f(x) g(Q(\tau|x)|x)^2}, \tag{1.18}$$

where $\mu_2(K) = \int u^2 K(u)\,du$ and $R(K) = \int K^2(u)\,du$. $Q''(\tau|x)$ and $g(Q(\tau|x)|x)$ both are unknown functions.

The authors have proposed the subsequent steps (suitable only for symmetrical distributions) to find the optimal bandwidth.

- Compute the ratio $\left(\frac{h_{\tau_1}}{h_{\tau_2}}\right)^5$ by using the optimal bandwidths at different quantiles τ_1 and τ_2.

$$\left(\frac{h_{\tau_1}}{h_{\tau_2}}\right)^5 = \frac{\tau_1(1-\tau_1)Q''(\tau_2|x)^2 g(Q(\tau_2|x)|x)^2}{\tau_2(1-\tau_2)Q''(\tau_1|x)^2 g(Q(\tau_1|x)|x)^2},$$

- According to their rule-of-thumb set $Q''(\tau_1|x) = Q''(\tau_2|x)$.

- Employ the standard normal distribution for $g(Q(\tau|x)|x)$.

- Then the bandwidth formula becomes,

$$h_\tau^5 = \pi^{-1} 2\tau(1-\tau)\phi(\Phi^{-1}(\tau))^{-2} h_{1/2}^5,$$

  where $h_{1/2}$ is the optimal bandwidth for the median.

- Compute $h_{1/2}$ using the following expression, which can be considered as a combination of a plug-in rule and a rule-of-thumb.

$$\left(\frac{h_{\text{mean}}}{h_{1/2}}\right)^5 = \frac{2}{\pi},$$

  where h_{mean} is the optimal bandwidth for mean regression. Plug-in rule is used to find the optimal choice for h_{mean} (Fan & Gijbels [13] ; Ruppert et.al [38]),

$$h_{\text{mean}}^5 = \frac{R(K)\sigma^2(x)}{n\mu_2(K)^2\{m''(x)\}^2 f(x)},$$

  with the conditional mean function $m(x)$ and the variance $\sigma^2(x)$.

The proposed rule-of-thumb has a relative rate of convergence of $\mathcal{O}\left(n^{-1/7}\right)$ under the normal assumption. A detailed description can be found in Yu & Jones [42] and Yu & Lu [44].

Both Kai et al and Zheng et al, exploit the relationship between asymptotic MSE for OLS and that for CQR to find plug-in bandwidths for CQR.

The optimal bandwidth in the sense of minimizing MISE($\hat{m}_\tau(x)$) is defined as,

$$h = \left(\frac{\nu_0 \int \dfrac{\sigma^2(x)w(x)}{f_X(x)} dx}{n \int m''(x)^2 w(x)\, dx\, \mu_2^2} \right)^{1/5} R_1(q)^{1/5}, \tag{1.19}$$

where $\nu_0 = \int K^2(u)\, du$, $\mu_2 = \int u^2 K(u)\, du$, $f_X(\cdot)$ the marginal density function of the covariate X , $w(x)$ a weight function and

$$R_1(q) = \frac{1}{q^2} \sum_{k=1}^{q} \sum_{k'=1}^{q} \frac{\tau_{kk'}}{f(c_k) f(c_k)}, \tag{1.20}$$

with $\tau_{kk'} = \tau_k \wedge \tau_{k'} - \tau_k \tau_{k'}$ and $c_k = F^{-1}(\tau_k)$.

The function $f(\cdot)$ is estimated with the fitted residuals and $\sigma^2(x)$ is the variance of the residuals.

The plug-in bandwidth for the local linear regression estimator is

$$h_{\text{LS}} = \left(\frac{\nu_0 \int \dfrac{\sigma^2(x)w(x)}{f_X(x)} dx}{n \int m''(x)^2 w(x)\, dx\, \mu_2^2} \right)^{1/5} . \tag{1.21}$$

Then the expressions (1.19) and (1.21) follow that

$$h = h_{\text{LS}} R_1(q)^{1/5}. \tag{1.22}$$

Generally, local quadratic regression decreases the bias of an estimation without increasing the variance (Fan & Gijbels [13]). Therefore, local quadratic regression is preferred for estimating the derivative. The optimal bandwidth is computed by minimizing the MISE ($\hat{m}'_\tau(x)$) ,

$$h = \left(\frac{27\nu_2 \int \dfrac{\sigma^2(x)w(x)}{f_X(x)} dx}{n \int m'''(x)^2 w(x)\, dx\, \mu_4^2} \right)^{1/7} R_2(q)^{1/7}, \tag{1.23}$$

where $\nu_2 = \int u^2 K^2(u)\,du$, $\mu_4 = \int u^4 K(u)\,du$ and

$$R_2(q) = \frac{\left(\sum_{k=1}^{q}\sum_{k'=1}^{q}\tau_{kk'}\right)}{\left(\sum_{k=1}^{q} f(c_k)\right)^2}. \tag{1.24}$$

The plug-in bandwidth for the derivative estimator is

$$h_{\mathrm{LS}} = \left(\frac{27\nu_2 \int \dfrac{\sigma^2(x)w(x)}{f_T(x)} dt}{n\int m'''(x)^2 w(x)\,dt\,\mu_4^2} \right)^{1/7}. \tag{1.25}$$

It follows that,

$$h = h_{\mathrm{LS}} R_2(q)^{1/7}. \tag{1.26}$$

In order to find $\int m'''(x)^2$, consider a global cubic model,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon, \tag{1.27}$$

where $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3$ and $\int (m'''(x))^2\,dx$ is,

$$\int (m'''(x))^2\,dx \approx \lim_{\delta t \to 0} \sum_i (m'''(x_i))^2 \delta x,$$

with $\delta x = (x_i - x_{i-1})$.

It is important to mention that in contrast to the previously discussed methods, the local quadratic CQR estimator for the derivative exhibits the optimal rate of convergence $\mathcal{O}\left(n^{2/7}\right)$.

Many researchers have studied non-parametric regression with correlated errors. A good review on this topic is given in the paper Opsomer et al [35]. They have focused on the non-parametric model for time series data,

$$y_i = m\left(\frac{i}{n}\right) + \epsilon_i, \tag{1.28}$$

with $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_\epsilon^2$ and equally spaced fixed design points $x_i = \dfrac{i}{n}$.
Suppose the errors $\{\epsilon_i\}$ comprise a stationary process with correlation function $\rho(k)$ satisfying

$\sum_{k=1}^{\infty} |\rho(k)| < \infty$. The optimal plug-in bandwidth for estimating $m(x)$ is

$$h_{\text{LS}} = \left(\frac{\nu_0 \, \sigma_\epsilon^2 (1+2R)}{n \, \mu_2^2 \int m''(x)^2 \, dx} \right)^{1/5}, \tag{1.29}$$

where $R = \sum_{k=1}^{\infty} \rho(k)$. In Chapter 2 we consider model

$$\epsilon_t = \phi \epsilon_{t-1} + z_t,$$

where $|\phi| < 1$ and $\sigma_z^2 < \infty$. In this case we estimate

$$\hat{h}_{\text{LS}} = \left(\frac{\nu_0 \, \hat{\sigma}_z^2}{(1-\hat{\phi})^2 n \, \mu_2^2 \int m''(x)^2 \, dx} \right)^{1/5}. \tag{1.30}$$

The corresponding non-parametric quantile model for time series data can be defined as

$$y_i = m_\tau \left(\frac{i}{n} \right) + \epsilon_i^\tau, \tag{1.31}$$

with $x_i = \dfrac{i}{n}$ equally spaced fixed design points. The optimal bandwidth for estimating the function $m_\tau(x)$ is

$$\hat{h}_\tau = \hat{h}_{\text{LS}} \hat{R}_1(q)^{1/5}. \tag{1.32}$$

In Chapter 2 we estimate $m'_\tau(x)$ by smoothing $y_t^* = y_t - \hat{\phi} \hat{\epsilon}_{t-1}$ against t. In this case the optimal h_τ is approximately

$$\hat{h}_\tau = \left(\frac{27 \, \nu_2 \, \hat{\sigma}_z^2}{n \, \mu_4^2 \int m'''(x)^2 \, dx} \right)^{1/7} \hat{R}_2(q)^{1/7}. \tag{1.33}$$

The remainder of this thesis is organized as follows. Chapter 2 applies quantile regression to a data set of autocorrelated temperature anomalies. Our main goal is to estimate rate of change of temperature anomalies. Chapter 3 explains some properties of the tornado dataset which is non-normally distributed; We apply quantile regression method to tornado count and investigate derivative behavior. The conclusions are given in Chapter 4 and the R-code is in the Appendix.

Chapter 2

Quantile Regression with Temperature data

2.1 Parametric Procedure

This work is concerned in estimating the derivatives of mean or quantile regression functions. Estimation of the derivative i.e. rate of change, is important in exploring the structure of regression curves.

Most of the statistical literature on regression analysis focuses on the conditional mean function $\mu(\boldsymbol{X}) = E(Y|\boldsymbol{X} = (x_1, \ldots, x_n))$ and estimating partial derivatives $\dfrac{\partial \mu(\mathbf{X})}{\partial x_i}$ using regression coefficients. In this work we will only consider one explanatory variable. As in Chaudhuri et al [4], quantile regression defines the conditional quantile function and the rate of change in the response variable as $\theta_\tau(\boldsymbol{X}) = Q_\tau(Y|\boldsymbol{X} = (x_1, \ldots, x_n))$, $\dfrac{\partial \theta_\tau(\mathbf{X})}{\partial x_i}$, respectively. The derivatives are estimated by using β_τ, which are the regression estimates at the τ-th quantile as defined in the model (1.6).

We consider a parametric and non-parametric approach applied to a data set of temperature anomalies in order to capture the behavior of the derivative. This section explains the parametric approach. The temperature anomaly is the difference between a particular temperature and the average over a base period. The base period is also called long-term average or reference value. A positive anomaly signifies that the temperature was warmer than the reference value and a negative anomaly means that the temperature was cooler than the reference value. The reason to study the anomalies, instead of the actual values is that they could function as better indicators. For example, a summer month over an area may be cooler than average, both at a mountain top and inhabited valley, but the actual temperatures will be quite different at the two locations. Moreover, it is difficult to collect temperature values in some areas in the world which have few temperature measurement stations so that temperature is measured over large areas such as deserts, mountains and remote forests. Thus, using the departure from an average, compared to the actual temperatures, allows for more accurate interpretations [32].

The next important thing is selecting a proper model for the given data. One of the best criteria used for this purpose is the Schwarz Information Criterion(SIC)(Kohler & Murphree 1988). Since the SIC is derived using Bayesian arguments, it is also known as the Bayesian Information Criterion (BIC) [40].
The Schwarz Criterion takes the general form,

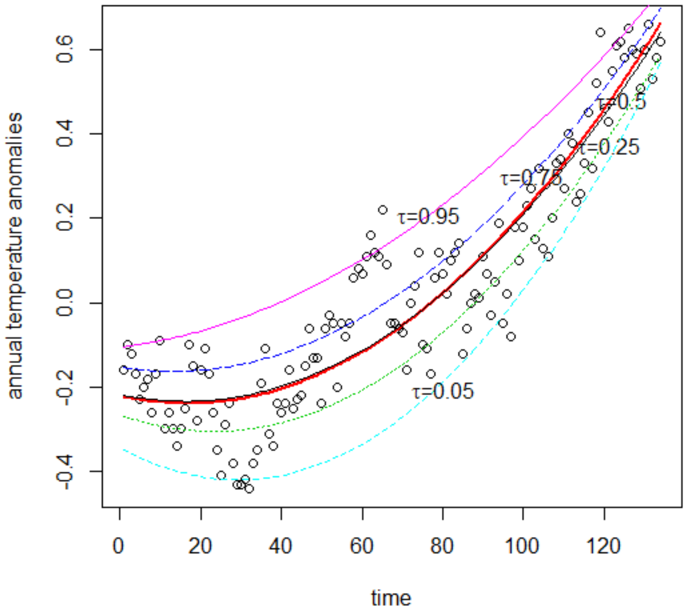
$$\text{SIC} = n \ln \left(\frac{\sum_i \rho(\epsilon_i)}{n} \right) + k \ln (n) \tag{2.1}$$

where k is the number of parameters. As a consequence, the best model will possess the minimum SIC value.

Among the linear, quadratic and cubic models, we applied the SIC criterion to choose the optimal one. The results indicate that the quadratic model has minimal SIC value and therefore it is the one that we consider for our work.
Let $t = (t_1, t_2, \ldots, t_n)$ be time points and $y = (y_1, y_2, \ldots, y_n)$ the n observed responses. Consider the parametric regression model,

$$y_\tau = \beta_{0\tau} + \beta_{1\tau} t + \beta_{2\tau} t^2 + \epsilon_\tau \tag{2.2}$$

where ϵ_τ represents the residual vector, and ϵ_τ satisfies (1.6). We use the model (2.2) to fit regression quantiles for temperature data and observe the behavior of the curves at different quantiles. Figure 2.1 illustrates the fitted models for different quantiles $\tau = 0.05, 0.25, 0.50, 0.75, 0.95$. The package **quantreg** available in R is used to estimate and make inferences about conditional quantile functions [27]. Let us consider the median quantile regression fit. The function **rq** in **quantreg** is used to fit the median regression for the observed data using the quadratic model.



Note: Black Line is the Linear Regression Model

Figure 2.1: Quantile Regression Curves

In most cases time series data inherits autocorrelation, a property which can be verified with the Ljung-Box test [29], known to be robust to outliers. We apply this test to residuals from the fitted parametric quantile regression. The Ljung-Box test statistic for our model is 182.3074, which has a p-value of $<$ 2.2e-16. Since the p-value for the test statistic regarding the median regression is near zero, we can conclude that the temperature data has significant autocorrelation.

The first part of our goal, fitting a parametric model to our data, is complete. The next step is to obtain an estimate of the derivative using a resampling procedure from the autocorrelated residuals. For that matter, we consider the ARIMA(1,0,0) model,

$$\epsilon_t = \phi\epsilon_{t-1} + z_t \tag{2.3}$$

with $|\phi| < 1$. After examining the time series plot of the residuals (not included) and applying the Ljung-Box test to AR(1) residuals $\{\hat{z}_t\}$, we selected model (2.3) to fit the residuals of the parametric model.

Then we apply a non-parametric bootstrap procedure (the algorithm is explained in Section (1.4)) inside the parametric approach to construct the confidence interval for the derivative of the median regression fit.

The bootstrap sample is generated in the following way:

1. Obtain bootstrap residual values $z_1^*, z_2^*, \ldots, z_n^*$ using the **boot** function in R.

2. Compute ϵ_t^* values with the recursive formula $\epsilon_t^* = \hat{\phi}\epsilon_{t-1} + z_t^*$ where $t = 2, \ldots, n$.

3. Construct new data y^* with the resampled residuals ϵ_t^*.

4. Estimate regression coefficents $\beta_{1\tau}^*$ and $\beta_{2\tau}^*$ using the new model $y_\tau^* = \beta_{0\tau} + \beta_{1\tau}t + \beta_{2\tau}t^2 + \epsilon_\tau^*$.

5. Repeat steps 1 through 4, N times to obtain estimates of $\beta_{i\tau}^*$ with $i = 1, 2$.

6. Find the derivative of the bootstrapped model, .i.e $T_n^* = \beta_{1\tau}^* + 2\beta_{2\tau}^* t$.

After running for $N = 999$, we have a collection of estimated coefficients for the derivative. We construct a 95% confidence interval for the estimation using those coefficients with the following formula:

$$(T^*_{(N+1)(\alpha/2)}, T^*_{(N+1)(1-(\alpha/2))}) \tag{2.4}$$

where T^* is the estimate of the bootstrap sample. If $(N+1)(\alpha/2)$ is an integer, the quantile , $T^*_{(N+1)(\alpha/2)}$ is estimated with the $(N+1)(\alpha/2)$ element of the ordered bootstrap sample. If not, interpolation is used between $(\lfloor(N+1)(\alpha/2)\rfloor)$-th and $(\lfloor(N+1)(\alpha/2)+1)\rfloor)$-th elements of the ordered sample, where $\lfloor\cdot\rfloor$ denotes the floor function.

According to the above formula, for 999 estimated values, 95% confidence level would have 25 -th and 975 -th elements as lower and upper limits. The derivative for the first quartile, median and third quartile regression fits and the 95% confidence interval for the estimation are illustrated in Fig.2.2. Based on the figures we have a 95% confidence that the derivative is non-negative after 1913 for the median regression while it is true for first and third quartile after 1918, 1911 respectively.

2.2 Non-Parametric Procedure

Non-parametric kernel smoothing techniques can be applied without making any restrictive assumptions about the form of the unknown function introduced in the model (1.12). Therefore, these techniques have become quite popular based on their flexibility over the parametric methods.

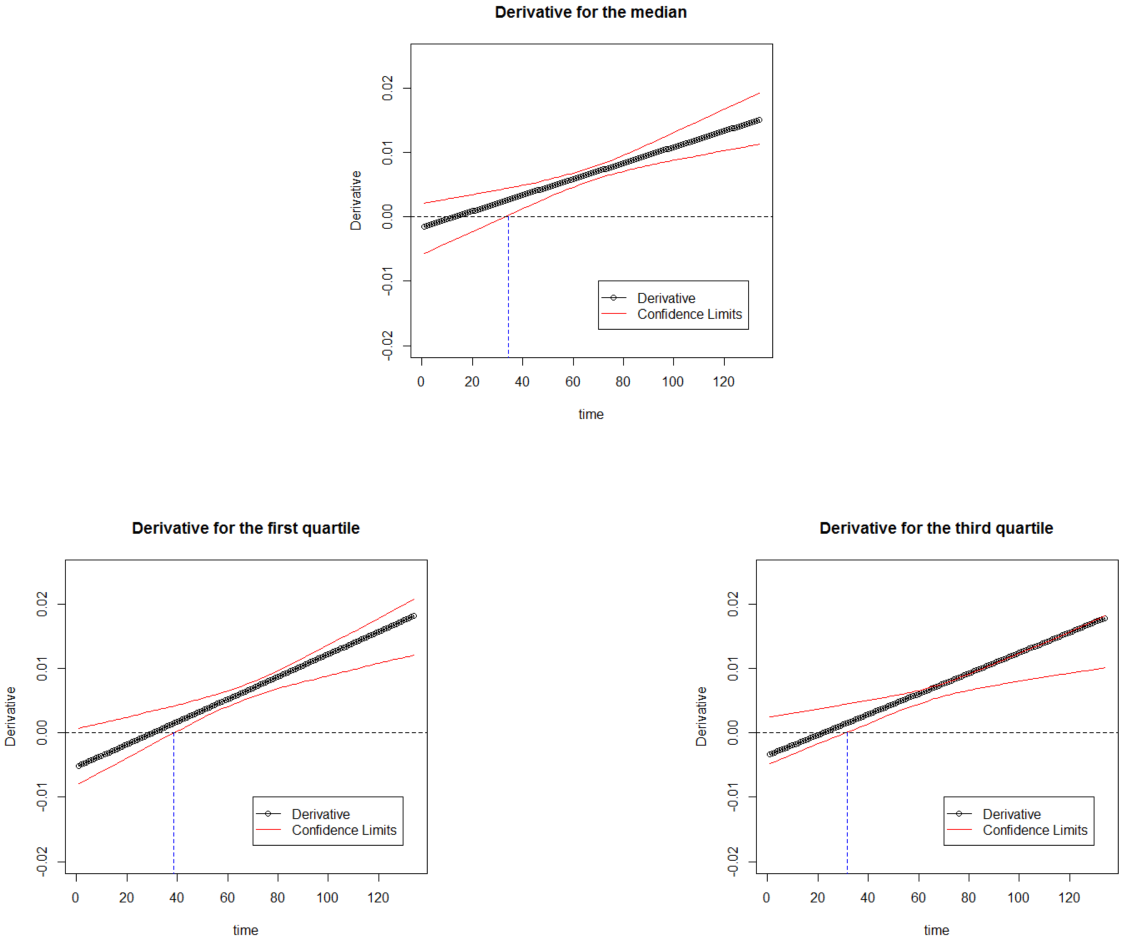


Figure 2.2: Parametric Estimation of the Derivative for different Quantiles

Nadaraya-Watson, local linear and nearest-neighbor are some well-known kernel smoothing methods. The performance of these methods is based on the smoothness of the regression function, the kernel density function and the bandwidth parameter. Here we are interested in local polynomial quantile regression.

Local fitting combined with quantile regression was introduced in the field of non-parametric statistics by Chaudhuri [3]. Yu & Jones [42] worked on local linear quantile regression focusing on bandwidth selection. Yu & Lu [44] studied the estimation of average derivative using a local linear additive quantile regression model. Ghouch & Genton [16] worked on local polynomial quantile regression with parametric features. A vast literature can be found on this non-parametric regression method.

In order to estimate the derivative for the local quadratic median regression we first recall the non-parametric model (1.31)

$$y_t = m_\tau\left(\frac{t}{n}\right) + \epsilon_t^\tau$$

with equally spaced fixed design points $x_i = \dfrac{t}{n}$ and apply the following steps:

1. We consider the bandwidth (1.32) (which is explained in Section (1.6)). Finding the optimal bandwidth for the local linear regression estimator requires the plug-in bandwidth h_{LS} and $R_1(q)$. To get h_{LS}, we do the following:

   (a) Approximate $\int (m''(x))^2\,dx$ in the following way:

      i. We fit a second order polynomial of mean regression.

      ii. Then the derivative $m''(x)$ is given by $2\hat{\beta}_2$.

      iii. Finally, approximate $\int (m''(x))^2\,dx$ with $\sum_i (m''(x_i))^2 \delta x$ (Refer (1.6)). In our case x denotes years and $\delta x = 1$.

   (b) Set $\nu_0 = \dfrac{1}{2\sqrt{\pi}}$ and $\mu_2 = 1$ for the Gaussian kernel.

   (c) The ARIMA coefficient $\hat{\phi}$ is computed by fitting an ARIMA model (2.3) for the residuals of the quadratic model obtained in step (a), and $\hat{\sigma}_z^2$ is the variance for residuals of the ARIMA model.

   (d) Using the above information we now calculate the plug-in bandwidth h_{LS}.

Letting $q = 1$, $R_1(q)$ is computed with

$$R_1 = \frac{(1/2)(1 - 1/2)}{f(F^{-1}(0.5))^2}$$

where $f(F^{-1}(0.5))$ is the density of ϵ_t^τ evaluated at 0.5-th quantile. In here, $f(\cdot)$ is estimated with the fitted residuals of the median regression function. Since kernel density estimation is a well-known non-parametric method for estimating probability density functions, we use the Gaussian kernel density function to infer the error density (see the R-code).

With the values h_{LS} and R_1, we are ready to compute the optimal bandwidth for estimating the function $\hat{m}_\tau(x)$ as explained in (1.32). At this point the first step is complete.

2. We estimate $m_\tau(x)$ using the bandwidth computed in step 1. That is done by fitting the local linear qunatile regression model (1.15) using the package **lprq** in R. The code to produce the local linear fit is:

```
fit=lprq(x, y, h, tau = 0.5, m)
```

where x and y are the explanatory and the response variables, respectively. In our case x is time, y is the annual temperature anomalies, h is the bandwidth parameter, τ is the fixed quantile and m is the number of points where the function is to be estimated. For the smoothing kernel we use the Gaussian kernel which is the default in R.

In order to take into account autocorrelation, we obtain the residuals of the non-parametric fit (ϵ_t^τ) using `fit$residuals`. Once ϵ_t^τ is obtained, as we have done in the parametric approach, we plug it in back into the ARIMA model (2.3) and find the value of the ARIMA coefficient $\hat{\phi}$. Now we consider a new model in terms of the original one which can be written in the following way:

$$y_t = m_\tau(x) + \phi\epsilon_{t-1}^\tau + z_t \tag{2.5}$$

Using the ARIMA coefficient $\hat{\phi}$ and error terms ϵ_{t-1}^τ, we substitute them in the previous equation and get

$$y_t - \hat{\phi}\epsilon_{t-1}^\tau = \tilde{m}_\tau(x) + z_t \tag{2.6}$$

which defines the new model

$$y^* = \tilde{m}_\tau(x) + z_t \tag{2.7}$$

where y^* represents the new observed values. The final step is to estimate $m_\tau(x)$ again based on (x, y^*). The purpose of transforming the original y values to $y*$ values is to remove the autocorrelation by bringing it closer to the iid assumption. This whole procedure is executed as many times as necessary until we get the desired smoothness.

3. With the new $y*$ values we estimate the derivative of $\hat{m}_\tau(x)$ using the optimal bandwidth in (1.33).

   Here, $\nu_2 = 1/(4\sqrt{\pi})$ and $\mu_4 = 3$ for the Gaussian kernel. The way of calculating $\int m'''(x)\,dx$ is similar to the one that was explained in step 1 , but with a cubic model (1.27) and σ^2 is the variance of the residual arising in model (2.7). Computing

$$R_2 = \frac{(1/2)(1-1/2)}{f(F^{-1}(0.5))^2}$$

   and the bandwidth for local quadratic regression, we get the optimal bandwidth, which we use together with the y^* values to estimate the derivative.

The curves shown in the Fig:2.3, Fig:2.4 and Fig:2.5 display the estimate of the derivative regarding the median, first and third quartile, respectively.

We considered three different quantiles, where we expected different behaviors of their curves. However, the results showed something different, namely, we saw a similar pattern, i.e. a growth of the three estimates in the parametric approach in contrast to the corresponding non-parametric estimates. This is because that data is approximately normally distributed (see Fig: 2.6) and appears to have a constant variance, so that $m_\tau(x) = m(x) + q_\tau$ and $m'_\tau(x) = m'(x)$.

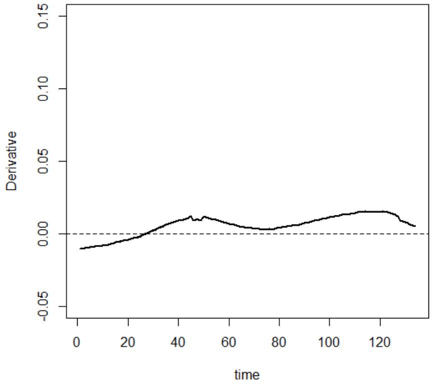


Figure 2.3: Non-Parametric Estimation of the Derivative for the Median

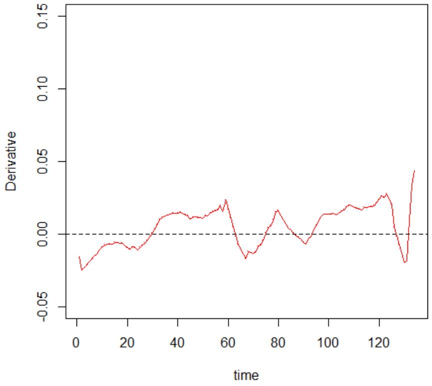


Figure 2.4: Non-Parametric Estimation of the Derivative for First Quartile

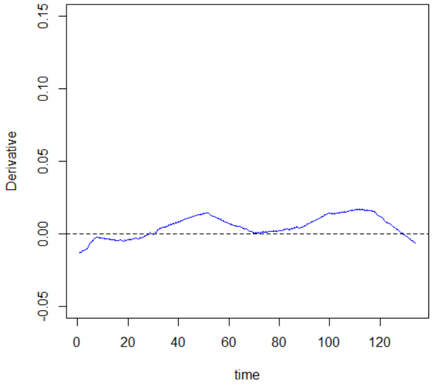


Figure 2.5: Non-Parametric Estimation of the Derivative for Third Quartile

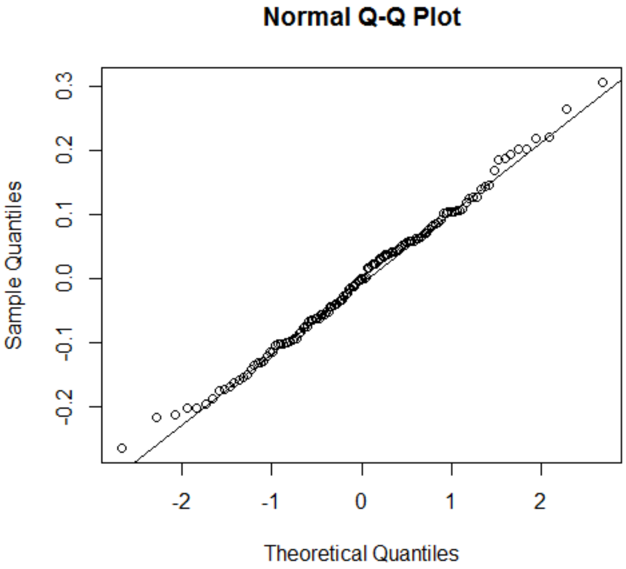


Figure 2.6: Q-Q plot for Temperature Data

Chapter 3

Tornado Climatology with Quantile Regression

Since quantile regression does not make restrictive assumptions on the form of the error distribution, it is able to do statistical analysis on non-normal data. We have studied a data set comprising the number of tornadoes for each month from 1950-2013. The data set is archived from NOAA [32].

Our work consists of two parts. First we split the data by months and observe the profile of the derivative of the number of tornadoes as a function of time (years) and check whether the derivative depends on the month and quantile. Next, for every year, we sum over the months the number of tornadoes, i.e. yearly tornado counts and study the dependency of the derivative on different quantiles.

3.1 Monthly Tornado Analysis

3.1.1 Parametric Approach

Consider the following linear quantile regression model,

$$y_{n,m} = \beta_{0,m}(\tau) + \beta_{1,m}(\tau)t + \epsilon^{\tau}_{n,m} \quad n = 1, \dots, 64 \quad m = 1, \dots, 12 \tag{3.1}$$

where n denotes the year and m denotes the month.

First we fit the linear quantile model introduced in (3.1) for each month. Using the Q-Q plots obtained for the residuals of twelve parametric models, we can conclude that the data follows non-normality. See Fig.3.1.

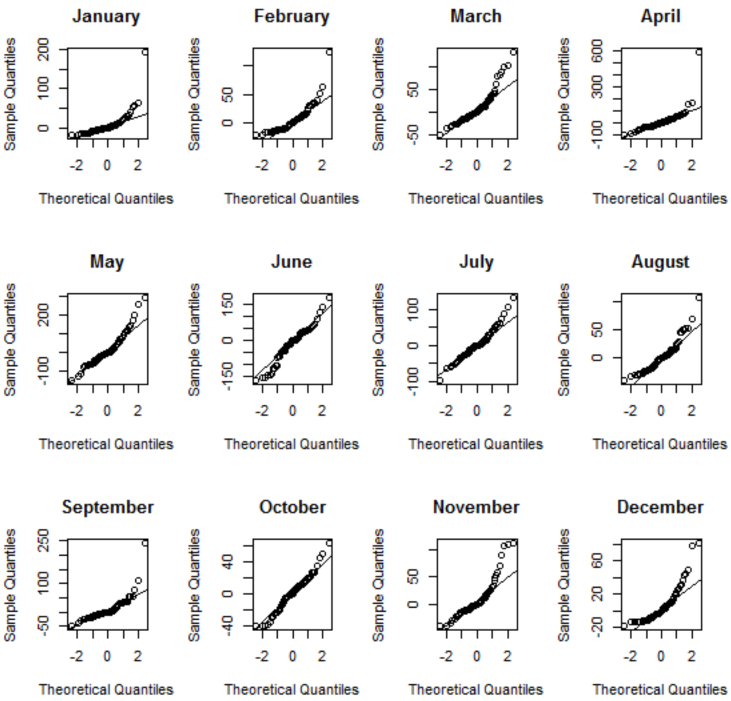


Figure 3.1: Q-Q Plots for Monthly data

The next task was to check the autocorrelation. We applied the Ljung-Box test to assess it. The resultant p-values by the aforementioned test indicate that there is no autocorrelation. As a consequence, we continue our work under the independent and non-normal assumption.

The method that we used to observe the behavior of the derivative is explained as follows:

1. Consider the linear quantile model introduced in (3.1) for each month.

2. The quantile regression estimator for the derivative is given by the coefficients $\beta_{1,m} \quad m = 1, \ldots, 12$.

3. Plot the derivative versus the month.

4. Repeat steps 1-3 for different quantiles.

Fig: 3.2 displays the derivative for each month at three different quantiles.

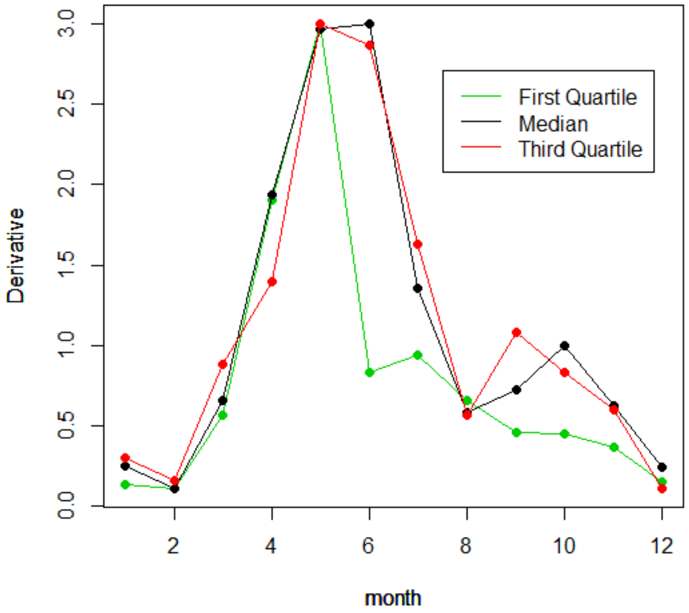


Figure 3.2: The Derivative Estimation by the Month and Quantile

There is a small difference in the derivative for the quantiles in January, February, March, August and December and significant one in June. Moreover, the estimation is almost the same for the month of May.

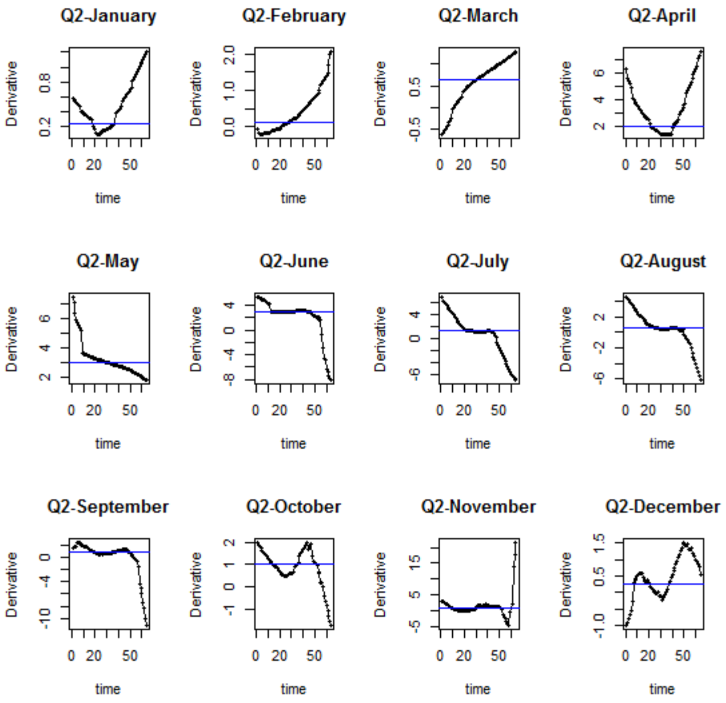
3.1.2 Non-Parametric Approach

We consider the local quadratic quantile regression to estimate the derivative of $m_\tau(x)$ in (1.12). The **lprq** package in R is used to fit the nonparametric model.

Bandwidth plays an important role in nonparametric regression. Thus, we should aim for finding the best possible candidate. In order to do this, we apply the method in the Section 1.6, Kai & Li [22] along with independent observations. More concretely, we estimate the derivative using the optimal bandwidth given by this concise formula explained in (1.26)

$$h = h_{\mathrm{LS}} R_2(q)^{1/7} \tag{3.2}$$

with h_{LS}, the optimal plug-In bandwidth for local least squares regression. In turn, h_{LS} is computed with the function **dpill** in the package **KernSmooth** in R.The following figure (Fig:3.3) illustrates the behavior of the derivative for local median regression.



Note that the blue line is the parametric estimation of the derivative.

Figure 3.3: The Local Median Regression estimation for the derivative by Month

Since we used a local quadratic function, the derivative should be linear. Nevertheless, Fig:3.3 exhibits different patterns. One possibility for this could be that we chose the simplest model, which may be inappropriate. It may be recommendable to introduce a higher order polynomial to obtain a better estimation. By looking at the graphs, a cubic model for January-April, a fourth order polynomial for May-August and a periodic function for September-December might be appropriate. Another reason could be that the non constant nature of the variance might be responsible for the nonlinearity of the derivative.

In addition, we observe the shape of the derivative at different quantiles which are characterized by Fig: 3.4, Fig: 3.5, Fig: 3.6 and Fig:3.7.

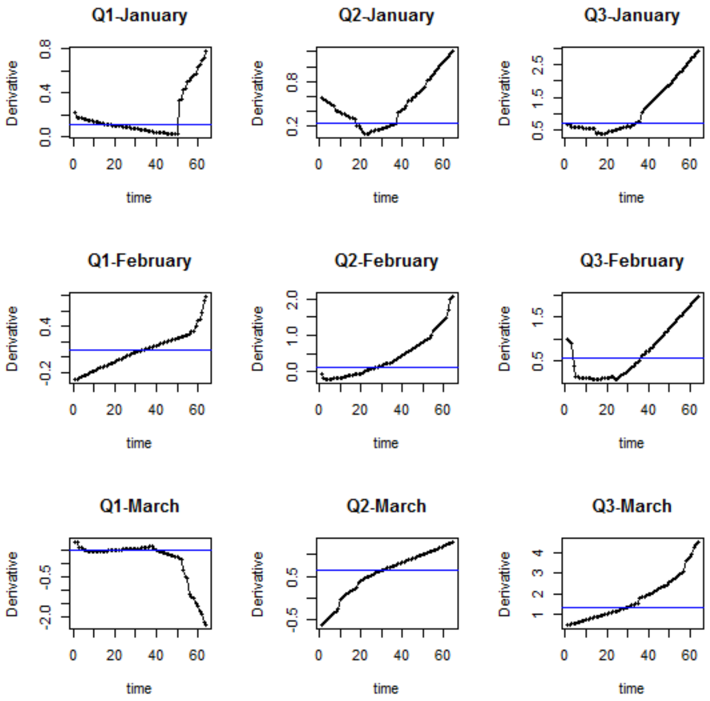


Figure 3.4: The Local Regression estimation for the derivative at $\tau = 0.25$, $0.5\,\&\,0.75$ -th quantiles : January-March

3.2 Yearly Tornado Analysis

One might be interested in estimating the yearly total number of tornadoes and its derivative. This section provides a method to do such analysis and applies it to tornado data going from 1950 to 2013. First we fit a local quadratic median regression and check the normality of the residuals using the Q-Q plot . In our case, Fig:3.8 indicates a departure from normality. Additionally, the Ljung-Box test reveals that there is no significant autocorrelation.

Then, we apply the same bandwidth used in Section 3.1.2 to non-parametrically estimate the derivative. The results are illustrated in Fig:3.9. From the graphs we can conclude that the non-parametric fit for the median is close to the parametric model over the years. As in the monthly analysis, the quadratic model does not appears suitable to estimate the derivative. Based on the results, a global quartic model may be appropriate.

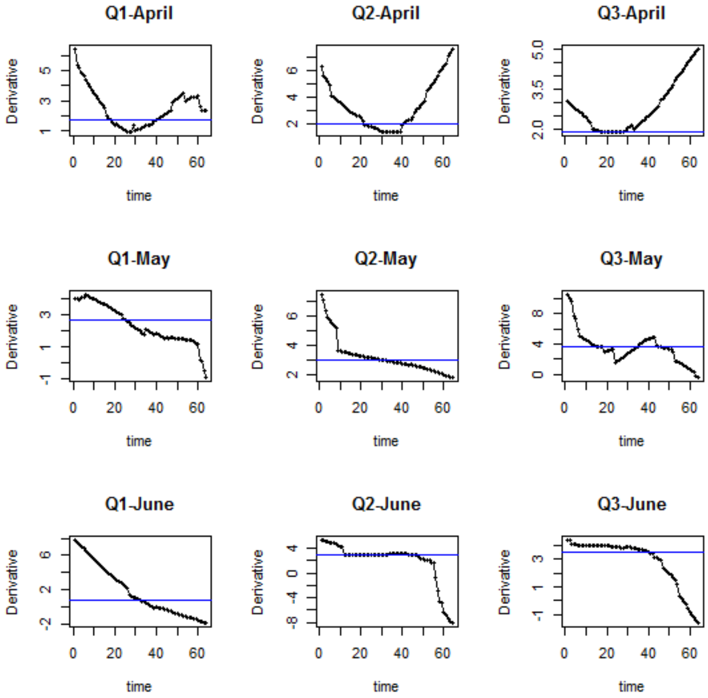


Figure 3.5: The Local Regression estimation for the derivative at $\tau = 0.25$, $0.5 \,\&\, 0.75$ -th quantiles : April-June

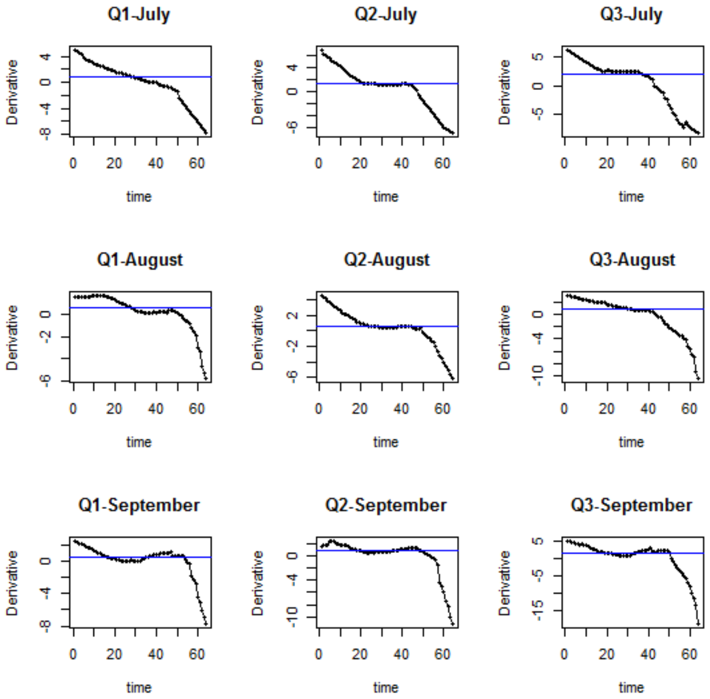


Figure 3.6: The Local Regression estimation for the derivative at $\tau = 0.25$, $0.5 \,\&\, 0.75$ -th quantiles : July-September

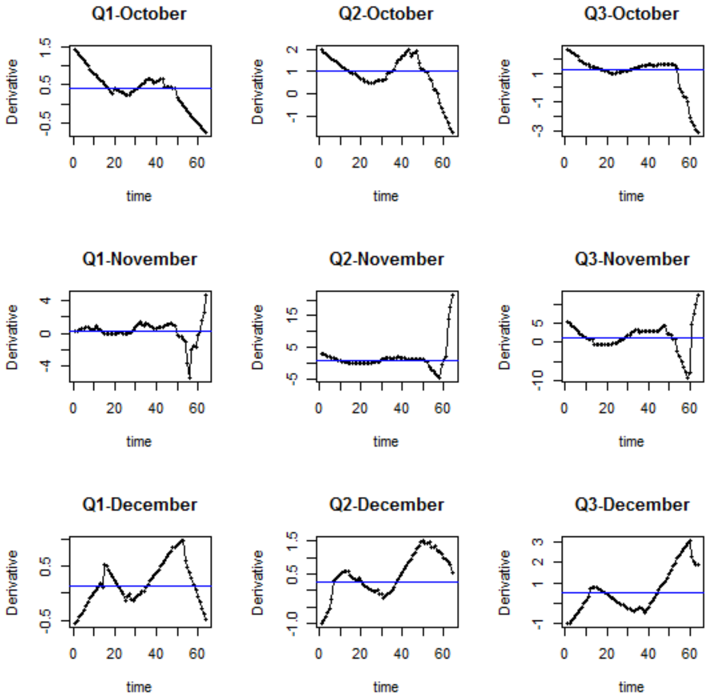


Figure 3.7: The Local Regression estimation for the derivative at $\tau = 0.25,\ 0.5\,\&\,0.75$ -th quantiles : October-December

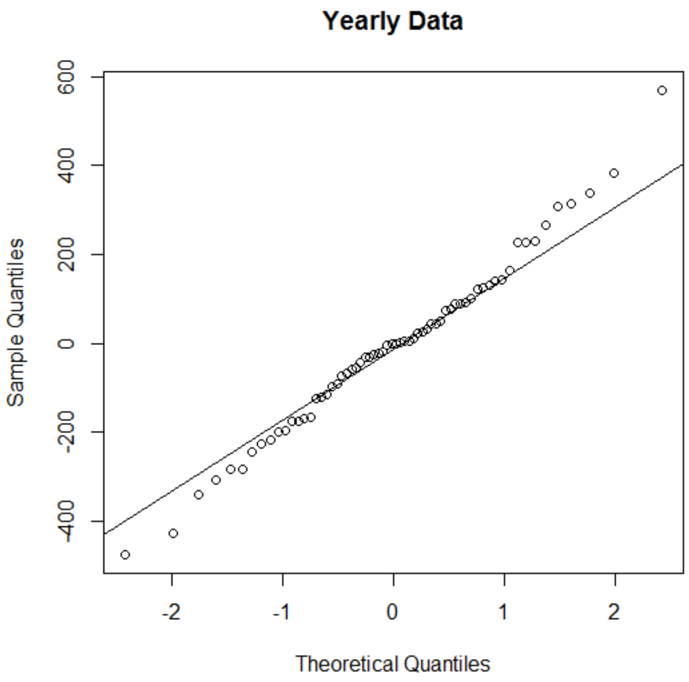


Figure 3.8: The Q-Q plot for Yearly Data

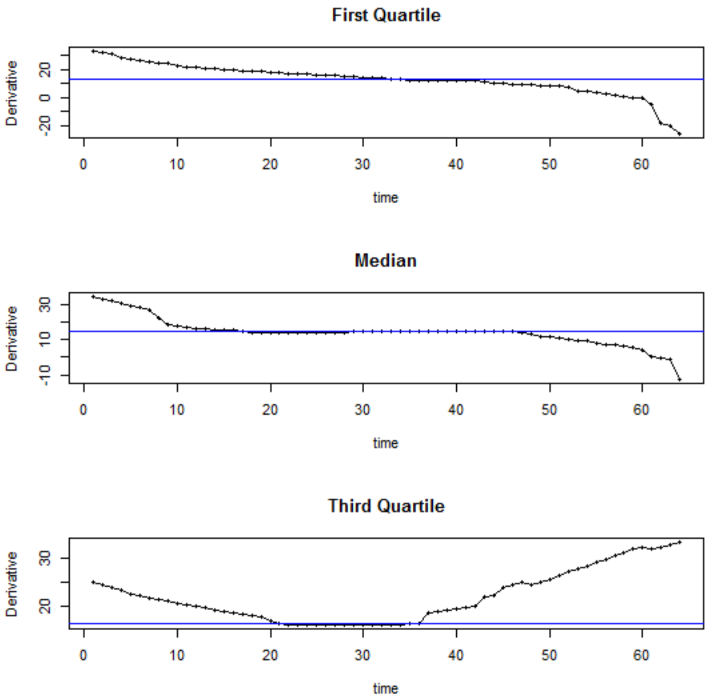


Figure 3.9: The Derivative for Yearly Data at $\tau = 0.25,\, 0.5\,\&\,0.75$ -th Quantiles

Chapter 4

Conclusions and Discussion

This thesis illustrates two applications of climatology using quantile regression in several settings. In particular, we introduced a method to handle autocorrelation in the framework of quantile regression and used it with the temperature data. Our results illustrate that the method works well for the parametric model since the increasing nature of the temperature is captured by the model in that the derivative is mostly positive (see Fig:2.2). In fact the sieve bootstrap method which we used inside the parametric approach was successful with quantile regression. In the non-parametric case, selecting the bandwidth parameter was of utmost importance. However, it can be shown that using the proper bandwidth parameter, the non-parametric model approaches the parametric one. The reason for this is the normality and the constant variance . By looking at the results we can see there is an issue in the smoothness of the first quartile compared to the median and third quartile (Fig:4.1). Since the optimal bandwidth for mean regression (h_{LS}) (Refer (1.30)) provided us fair results, the only quantities that may be the cause of not attaining the desired output are $R_1(q)$ and $R_2(q)$ in (1.20) and (1.24). In turn, $R_1(q)$ and $R_2(q)$ depend on $f(\cdot)$ which is the unknown distribution function estimated with the fitted residuals using a kernel density function. Thus, estimating $f(\cdot)$ is the crucial step. However, we will not pursue this idea further.

Therefore, in order to get a better estimation for the derivative our attention focused on local least square regression. We followed the same method in section 2.2 with the local quadratic least square function and the bandwidths (1.30), (1.33) (see R code in Appendix C). Using this procedure we obtained the result shown in Fig:4.2 which has the same pattern as the parametric models that were obtained for quantile regression (Fig:2.2).

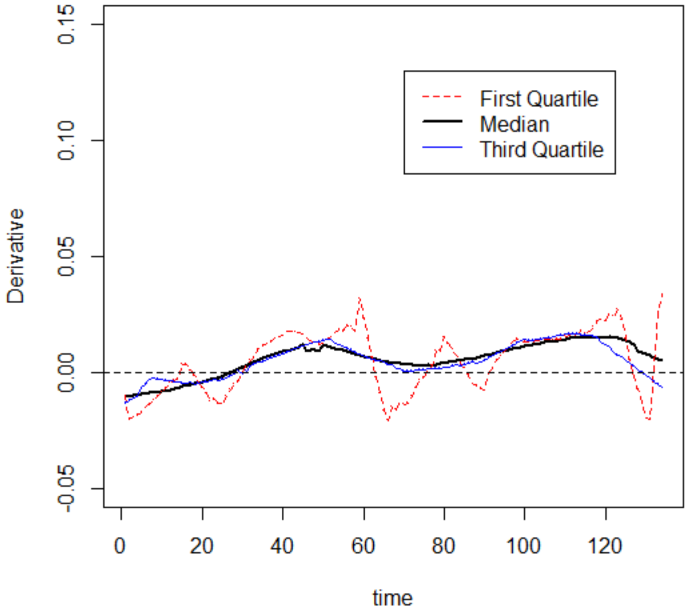


Figure 4.1: The Derivative Estimation for the Local Quadratic Quantile Regression

Moreover, we compared the parametric and non-parametric model for least square regression, and both generated very similar results as illustrated in Fig.4.3 and Fig.4.2. Now, the question of whether we need the non-parametric model can be easily addressed. The answer is simply no, since there is no significant difference in the estimated derivative.

Based on the previous discussion, after analyzing the temperature data, we can conclude that local least square regression performs better than quantile regression. Another important remark regarding the temperature analysis is that the quadratic model with first order autocorrelation provides a reliable estimate of the derivative function.

Since quantile regression does not require any distributional assumptions, we were able to handle non-normal data as it was done with the tornado data set. We chose the simplest model to estimate the derivative. However, figures (Fig:3.4 -Fig:3.7) emphasize that a more general model has to be considered in order to do a better analysis. A possible continuation of this work is the introduction of a global cubic model and a quartic model for our monthly and yearly tornado data, respectively, which we expect will allow for comparisons between months and years.

Through out this thesis we have discussed the flexibility of both parametric and non-

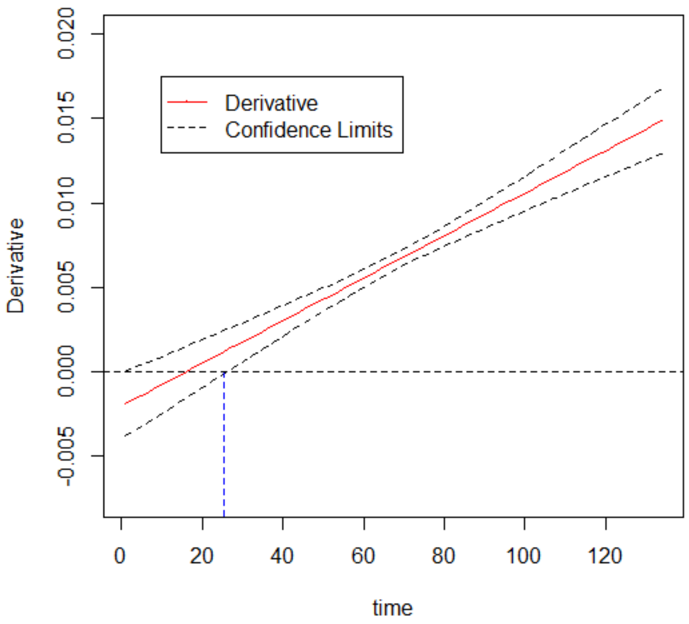


Figure 4.2: The Derivative Estimation for the Local Quadratic Least Square Regression

parametric statistical methods with quantile regression. We discussed as well the approach of statistical analysis of linear and nonlinear models with quantile regression. Quantile Regression provides us with a comprehensive picture of the sampling distribution compared to least square regression. Least square regression focuses on the mean of the response given x variables without accounting for the full conditional distributional properties. In contrast, quantile regression uses several regression curves at various quantiles of the distribution, providing us those very same distributional properties (see Fig:2.1). Even though quantile regression is more versatile, when talking about residuals with the normality and the constant variance assumption, we would prefer least square regression for our temperature analysis. On the other hand, when dealing with the non-normality and non constant variance assumption, quantile regression is a better candidate for the estimation of the derivative.

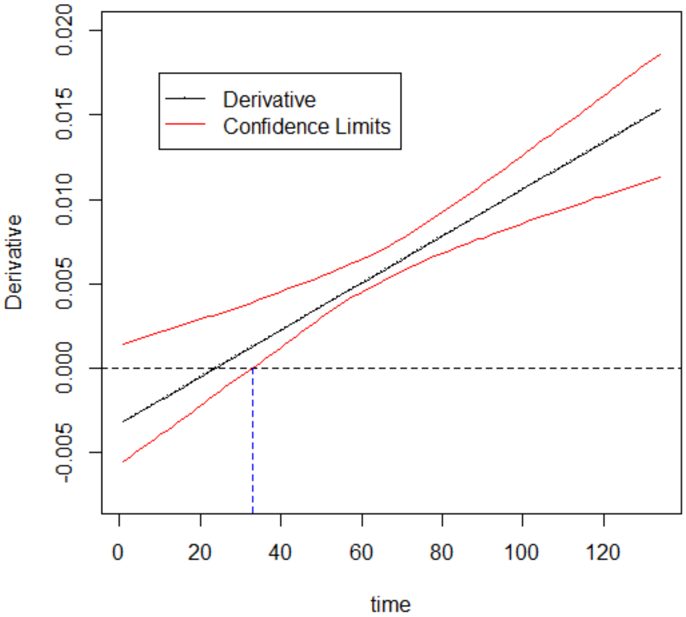


Figure 4.3: The Parametric Estimation of the Derivative for the Quadratic Least square Regression

Appendices

Appendix A R Code for Temperature Data

Parametric Approach

```
annualtemp <- scan()#Read Land and Ocean Data
times <- 1:(length(annualtemp))
plot(times,annualtemp,ylab ="annual temperature anomalies")
x1<- times^2
fitmed <- rq(annualtemp~times+x1,tau =.5)
lines(times,fitmed$fitted,col=2)

beta0 <-summary(fitmed)$coefficients[1,1]
beta1 <-summary(fitmed)$coefficients[2,1]
beta2 <-summary(fitmed)$coefficients[3,1]

res_P <-fitmed$residuals ###Residulas of parametric Model
R_P <-as.matrix(res_P)

acf(res_P)
pacf(res_P)

####Fitting ARIMA model

arimafit <-arima(res_P,order=c(1,0,0))
p <-arimafit$coef[1]


####Z_t values in ARIMA model

res_A <-arimafit$residuals
Z <-matrix(nrow=134,ncol=1)
Z <-res_A
```

```
Box.test(Z,type ="Ljung",lag=24,fitdf=0)

T <-beta1+(2*beta2*times)##Derivative for the Parametric Model

###Bootstraping T values

eps_star <- matrix(nrow =134,ncol= 1)
ystar <-matrix(nrow=134,ncol=1)

boot1 <-function(index,M){
subscripts1 <-sample((1:134),134,replace=TRUE)
Zstar <-M[subscripts1,]

eps_star[1] <-Zstar[1]
ystar[1] <-beta0+(beta1*times[1])+(beta2*x1[1])+eps_star[1]

for(i in 2:134){

eps_star[i] <-Zstar[i]+(p*eps_star[i-1])
ystar[i] <-beta0+(beta1*times[i])+(beta2*x1[i])+eps_star[i]
}
#ystar
Boot_fitmed <-rq(ystar~times+x1,tau=.5)

betastar1 <-summary(Boot_fitmed)$coefficients[2,1]
betastar2 <-summary(Boot_fitmed)$coefficients[3,1]
Tstar <-betastar1 + (2*betastar2*times)
Tstar
}
Tstar_B <-sapply(1,boot1,M=Z)
```

```
Tstar_Boot <-matrix(nrow=134,ncol=999)
Tstar_Boot <-sapply(1:999,boot1,M=Z)

qup <-apply(Tstar_Boot,1,quantile,probs =0.975)
qlow<-apply(Tstar_Boot,1,quantile,probs =0.025)

plot(1:134,Tstar_B,xlab="times",ylab="Derivatives",
ylim=c(-2e-2,2.5e-2))
lines(1:134,Tstar_B)
lines(1:134,qup,col=2)
lines(1:134,qlow,col=2)
abline(h=0,lty=2)
legend(70,-0.01,c("Derivative","Confidence Limits"),
col=c(1,2),lty=c(1,1),pch=c(1,NA))
title("Derivative for the median")
lines(x=c(34.3,34.3),y=c(0,-0.03),lty=2,col="blue")
```

Non-Parametric Approach

```
annualtemp<-scan()#Read Land and Ocean Data
times<-1:(length(annualtemp))
x1<-times^2

##########Step 1-Computing the Bandwidth

fitmedian <-lm(annualtemp~times+x1)
beta2 <-summary(fitmedian)$coefficients[3,1]

T <-2*beta2
m_dprime <-(T^2)
```

```
PR <-fitmedian$residuals


######ARIMA coefficient


arimafit <-arima(PR,order=c(1,0,0))
p <-arimafit$coef[1]


#######variance
sigmaZ <-var(arimafit$residuals)


######compute R1


fitmedian2 <-rq(annualtemp~times)
res2<-fitmedian2$residuals


######kernel density
den<-density(res2,kernel = "gaussian")
n <-length(den$y)
y.cs <-cumsum(den$y)
i.med <- length(y.cs[2*y.cs <= y.cs[n]])
y.med <- den$y[i.med]


v0 <-1/(2*sqrt(pi))
mu2<-1


h_LS <-((v0*sigmaZ)/(134*((1-p)^2)*m_dprime*(mu2^2)))^(1/5)
R1 <-((1/4)/((y.med)^2))
hoptimal <-h_LS*(R1^(1/5))



##########Step 2-Estimating the function
```

```
annualtemp_c <- annualtemp
for(k in 1:10){
Nfit <-lprq(times,annualtemp_c,hoptimal,m=134,tau=0.5)
res <-annualtemp-Nfit$fv
arimafit <-arima(res,order=c(1,0,0))
phi<-arimafit$coef[1]

#ystar[1]<-annualtemp[1]
#annualtemp[1] never changes
#res[n] never used
for(i in 2:134){
annualtemp_c[i]<-annualtemp_c[i]-(phi*res[i-1])
}
}

R <-arimafit$residuals
Box.test(R,type="Ljung",lag=24,fitdf=0)

sig <-var(R)
newy <-annualtemp_c

########### Step 3-Estimating the derivative

x1 <-times^2
x2 <-times^3
fitmed1 <-lm(newy~times+x1+x2)

beta3 <-summary(fitmed1)$coefficients[4,1]
T2 <-6*beta3
T_tprime <-(T2^2)
```

```
fit <-rq(newy~times,tau=0.5)

RR <-fit$residuals

D1 <-density(RR,kernel = "gaussian")

n <-length(D1$y)

y.cs <-cumsum(D1$y)

i.med <- length(y.cs[2*y.cs <= y.cs[n]])

y.med <- D1$y[i.med]



v2 <-1/(4*sqrt(pi))

mu4 <-3

h_LS <-((v2*sig)/(134*T_tprime*(mu4^2)))^(1/7)

R2 <-((1/4)/((y.med)^2))

hopt <-h_LS*(R2^(1/7))



lprq<-function (x, y, h, tau = 0.5, m = 50)

{

    xx <- seq(min(x), max(x), length = m)

    fv <- xx

    dv <- xx

    for (i in 1:length(xx)) {

     z <- x - xx[i]

       u<-z^2

        wx <- dnorm(z/h)

        r <- rq(y ~ z+u, weights = wx, tau = tau, ci = FALSE)

        fv[i] <- r$coef[1]

        dv[i] <- r$coef[2]

        }

    list(xx = xx, fv = fv, dv = dv)
}


Nfit1 <-lprq(times,newy,hopt,m=134,tau=0.5)

plot(Nfit1$xx,Nfit1$dv,pch =".",ylim=c(-0.05,0.05),lwd=2)

lines(Nfit1$xx,Nfit1$dv,lwd =2)

abline(h=0,lty=2)
```

Appendix B R Code for Tornado Data

```
#################### MONTHLY ANALYSIS

tornado<-read.csv(file.choose())
freq<-tornado$Freq
times<-1:64

##SPLITING THE DATASET BY MONTHS

jan<-freq[seq(1,768,by=12)]
feb<-freq[seq(2,768,by=12)]
mar<-freq[seq(3,768,by=12)]
apr<-freq[seq(4,768,by=12)]
may<-freq[seq(5,768,by=12)]
jun<-freq[seq(6,768,by=12)]
jul<-freq[seq(7,768,by=12)]
aug<-freq[seq(8,768,by=12)]
sep<-freq[seq(9,768,by=12)]
oct<-freq[seq(10,768,by=12)]
nov<-freq[seq(11,768,by=12)]
dec<-freq[seq(12,768,by=12)]

###### PARAMETRIC APPROACH

fitjan<-rq(jan˜times,tau=.50)
res_jan<-fitjan$resid

qqnorm(res_jan)
qqline(res_jan)
```

```
betajan<-summary(fitjan)$coefficients[2,1]
Beta<-cbind(betajan,betafeb,betamar,betaapr,betamay,
betajun,betajul,betaaug,betasep,betaoct,betanov,betadec)
B<-matrix(Beta,ncol=1,nrow=12,byrow=FALSE)
month<-1:12
plot(month,B,xlab="month",ylab="Derivative",pch=19)
lines(month,B)

###### NONPARAMETRIC APPROACH

lprq=function (x, y, h, tau, m )
{
    xx <- seq(min(x), max(x), length = m)
    fv <- xx
    dv <- xx
    for (i in 1:length(xx)) {

         z <- x - xx[i]
     u <- zˆ2
     wx <- dnorm(z/h)
        r <- rq(y ˜ z+u, weights = wx, tau = tau, ci = FALSE)
        fv[i] <- r$coef[1]
        dv[i] <- r$coef[2]
    }
    list(xx = xx, fv = fv, dv = dv)
}



fitjanorg<-rq(jan˜times,tau=.50)
betajan<-summary(fitjanorg)$coefficients[2,1]
```

```
D<-density(res_jan,kernel = "gaussian")

n<-length(D$y)
y.cs<-cumsum(D$y)
i.med <- length(y.cs[2*y.cs <= y.cs[n]])
y.med <- D$y[i.med]

year<-1:64
h_LS<-dpill(year,jan)
hLAD<-((1/4)/((y.med)^2))^(1/7)
h<-h_LS*hLAD

Nfit<-lprq(year,jan,h,m=64,tau=0.50)
plot(Nfit$xx,Nfit$dv,pch=20,main="Q2-January")
lines(Nfit$xx,Nfit$dv)
abline(h=betajan,col="blue")



#################### YEARLY ANALYSIS

yearly<-1:64
for(i in 0:63){
yearly[(i+1)]=sum(freq[(12*i+1):(12*i+12)])
}

times<-1:64
plot(times,yearly,xlab="year",ylab="Number of Tornadoes")
fitmed<-rq(yearly~times,tau=.50)
res<-fitmed$resid
```

```
qqnorm(res,main="Yearly Data")
qqline(res)

Box.test(res, lag = 24, type = c("Ljung-Box"))
### pvalue=0.8168

beta<-summary(fitmed)$coefficients[2,1]

x2<-times^2
fitmed1<-rq(yearly~times+x2,tau=.50)
res1<-fitmed$resid

D<-density(res1,kernel = "gaussian")
n<-length(D$y)
y.cs<-cumsum(D$y)
i.med <- length(y.cs[2*y.cs <= y.cs[n]])
y.med <- D$y[i.med]

h_LS<-dpill(times,yearly)
hLAD<-((1/4)/((y.med)^2))^(1/7)
h<-h_LS*hLAD

Nfit<-lprq(times,yearly,h,m=64,tau=0.50)
plot(Nfit$xx,Nfit$dv,pch=20,main="Median")
lines(Nfit$xx,Nfit$dv)
abline(h=beta,col="blue")
```

Appendix C R Code for Temperature Data : Least Square Regression

Parametric Approach

```
annualtemp<-scan()#Land and Ocean
times<-1:(length(annualtemp))
x1<-times^2

fit<-lm(annualtemp~times+x1)
beta0<-summary(fit)$coefficients[1,1]
beta1<-summary(fit)$coefficients[2,1]
beta2<-summary(fit)$coefficients[3,1]



res_P<-fit$residuals
arimafit<-arima(res_P,order=c(1,0,0))
p<-arimafit$coef[1]



Z <-arimafit$residuals

Box.test(Z,type="Ljung",lag=48,fitdf=0)

T<-beta1+(2*beta2*times)##Derivative for the Parametric Model
T_prime<-2*beta2####Second derivative

###Bootstraping T values

eps_star<-matrix(nrow=134,ncol=1)
ystar<-matrix(nrow=134,ncol=1)
```

```
boot1<-function(index,M){
subscripts1<-sample((1:134),134,replace=TRUE)
Zstar<-M[subscripts1,]

eps_star[1]<-Zstar[1]
ystar[1]<-beta0+(beta1*times[1])+(beta2*x1[1])+eps_star[1]

for(i in 2:134){

eps_star[i]<-Zstar[i]+(p*eps_star[i-1])
ystar[i]<-beta0+(beta1*times[i])+(beta2*x1[i])+eps_star[i]
}
#ystar
Boot_fit<-lm(ystar˜times+x1)
#summary(Boot_fit95)
betastar1<-summary(Boot_fit)$coefficients[2,1]
betastar2<-summary(Boot_fit)$coefficients[3,1]
Tstar<-betastar1+(2*betastar2*times)
Tstar
}
Tstar_B<-sapply(1,boot1,M=Z)

Tstar_Boot<-matrix(nrow=134,ncol=999)
Tstar_Boot<-sapply(1:999,boot1,M=Z)

#Tstar_Boot

qup<-apply(Tstar_Boot,1,quantile,probs=0.975)
qlow<-apply(Tstar_Boot,1,quantile,probs=0.025)
plot(1:134,Tstar_B,xlab="times",pch=’.’,
ylab="Derivatives",ylim=c(-0.0075,0.02))
lines(1:134,Tstar_B)
lines(1:134,qup,col=2)
lines(1:134,qlow,col=2)
abline(h=0,lty=2)
legend(10,0.0175,c("Derivative","Confidence Limits"),col=c(1,2),lty=c(1,1),pch=c(’.’,NA
title("Derivative for the median")
lines(x=c(33.0825,33.0825),y=c(0,-0.03),lty=2,col="blue")
```

Non-Parametric Approach

```
######Step 1
annualtemp<-scan()#Land and Ocean
times<-1:(length(annualtemp))
x1<-timesˆ2

fitmedian<-lm(annualtemp˜times+x1)
beta2<-summary(fitmedian)$coefficients[3,1]

#######second derivative

T<-(2*beta2)
m_dprime<-(Tˆ2)

######variance of the parametric model
par_res<-fitmedian$residuals

#####ARIMA coefficient

arimafit<-arima(par_res,order=c(1,0,0))
p<-arimafit$coef[1]
```

```
#######variance
sigmaZ<-var(arimafit$residuals)


v0<-1/(2*sqrt(pi))
mu2<-1

h_LS<-((v0*sigmaZ)/(134*((1-p)^2)*m_dprime*(mu2^2)))^(1/5)

####### Step 2

annualtemp_c <- annualtemp
for(k in 1:2){
Nfit1<-locfit(annualtemp_c~times,alpha=h_LS)
N<-as.vector(fitted(Nfit))
res<-annualtemp-N
arimafit<-arima(res,order=c(1,0,0))
phi<-arimafit$coef[1]
for(i in 2:134){
annualtemp_c[i]<-annualtemp_c[i]-(phi*res[i-1])
}
}

R<-arimafit$residuals
Box.test(R,type="Ljung",lag=24,fitdf=0)

sig<-var(R)
newy<-annualtemp_c

########Step 3
```

```
fitmed1<-lm(newy˜times+x1+x2)

beta3<-summary(fitmed1)$coefficients[4,1]


T2<-6⋆beta3

T_tprime<-(T2ˆ2)



v2<-1/(4⋆sqrt(pi))

mu4<-3

h_LS<-((v2⋆sig)/(134⋆T_tprime⋆(mu4ˆ2)))ˆ(1/7)

Nfit1<-locfit(newy˜times,deriv=1,alpha=h_LS)

NF<-fitted(Nfit1)

plot(times,NF,pch=".",lwd=2)

lines(times,NF,lwd=2)

abline(h=0,lty=2)


plot(Nfit1,band="local",ylab="Derivative",

ylim=c(-0.0075,0.02),col=2)

lines(x=c(25.54,25.54),y=c(0,-0.03),lty=2,col="blue")

legend(10,0.0175,c("Derivative","Confidence Limits"),col=c(2,1),lty=c(1,2),pch=c(’.’,NA
```

Bibliography

[1] Bhattacharya,P. K. and Gangopadhyay, A. K. (1990): *Kernel and Nearest-Neighbor Estimation of a Conditional Quantile*, The Annals of Statistics,Vol. 18, No. 3 , pp. 1400-1415.

[2] Bühlmann, P. (1997) : *Sieve Bootstrap for Time Series*, Bernoulli, Vol. 3, No. 2, pp. 123-148.

[3] Chaudhuri, P. (1991): *Non-parametric estimates of regression quantiles and their local Bahadur representation*, The Annals of Statistics, Vol. 19, No. 2 , pp. 760-777.

[4] Chaudhuri, P. , Doksum, K. and Samarov, A.(1997): *On Average Derivative Quantile Regression*, The Annals of Statistics, Vol. 25, No. 2, pp. 715-744.

[5] Cleveland, W. (1979) : *Robust Locally Weighted Regression and Smoothing Scatterplots*, Journal of the American Statistical Association, Vol. 74, No. 368 , pp. 829-836.

[6] Cleveland, W. and Devlin, S. (1988) : *Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting* , Journal of the American Statistical Association, Vol. 83, No. 403 , pp. 596- 610.

[7] Cleveland, W.S. and Loader, C. (1996): *smoothing by local regression:principles and methods*, Statistical Theory and Computational Aspects of Smoothing, Springer, New York, 10-49.

[8] Davison,A. C and Hinkley,D.V (1997): *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics.

[9] Dunbar B. (2005): *What's the Difference Between Weather and Climate?*, Available: http://www.nasa.gov/mission-pages/noaa-n/climate/climate-weather.html , Last accessed 17th June 2014.

[10] Efron, B. (1979): *Bootstrap Methods: Another Look at the Jackknife*, The Annals of Statistics ,Vol. 7, No. 1, pp. 1-26.

[11] Efron, B. (1987): *Better Bootstrap Confidence Intervals*, Journal of the American Statistical Association, Vol. 82, No. 397, pp. 171-185.

[12] Efron B. and Tibshirani, R.J (1993) : *An Introduction of the Bootstrap*, Chapman and Hall/CRC: New York.

[13] Fan, J.Q. and Gijbels, P.J.(1992): *Variable bandwidth and local linear regression smoothers*, The Annals of Statistics, Vol. 20,No.4 , pp. 2008-2036.

[14] Freedman, D. (1981) : *Bootstrapping Regression Models*, The Annals of Statistics, Vol. 9, No. 6 , pp. 1218-1228.

[15] Friedman, J.H. and Stuetzle, W. (1981): *Projection Pursuit Regression*, Journal of the American Statistical Association, Vol. 76, No. 376, pp. 817-823.

[16] Ghouch, A. and Genton, M. (2009) : *Local Polynomial Quantile Regression With Parametric Features*, Journal of the American Statistical Association, Vol. 104, No. 488, pp. 1416-1429.

[17] Hahn, J. (1995) : *Bootstrapping Quantile Regression Estimators*, Econometric Theory, Vol. 11, No. 1 , pp. 105-121.

[18] Hall,P. , Sheathers, J., Jones, M .C. and Marron,J. S. (1991): *On optimal data-based bandwidth selection in kernel density estimation*, Biometrika , Vol. 78, No. 2, pp. 263-270.

[19] Härdle, W., Hart, J., Marron, J. S. and Tsybakov, A. B. (1992): *Bandwidth choice for average derivative estimation*, Journal of the American Statistical Association, Vol. 87, No. 417, pp. 218-226.

[20] Hastie, T. and Tibshirani, R.(1986) : *Generalized Additive Models*, Statistical Science, Vol. 1, No. 3 , pp. 297-310.

[21] Horowitz, J. (1998) : *Bootstrap Methods for Median Regression Models*, Econometrica, Vol. 66, No. 6 , pp. 1327-1351.

[22] Kai, B. and Li, R. (2010) : *Local composite quantile regression smoothing:an efficient and safe alternative to local polynomial regression*, Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 72, No. 1, pp. 49-69.

[23] Koenker, R., and G. Bassett (1978): *Regression Quantiles*, Econometrica, Vol. 46, No. 1, pp. 33-50.

[24] Koenker,R. and D'Orey, V. (1987): *Algorithm AS 229: Computing Regression Quantiles*, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 36, No. 3, pp. 383-393.

[25] Koenker, R. , Ng, P. and Portnoy, S. (1994): *Quantile Smoothing Splines*, Biometrika, Vol. 81, No. 4, pp. 673-680.

[26] Koenker, R. and Hallock, K.F (2001) : *Quantile Regression* , The Journal of Economic Perspectives, Vol. 15, No. 4, pp. 143-156.

[27] Koenker, R. (2005): *Quantile Regression*, Cambridge University Press, Cambridge.

[28] Leider,J.(2012): *A Quantile Regression Study of Climate Change in Chicago, 1960-2010*, Department of Mathematics, Statistics and Computer Science, University of Illinois,Chicago.

[29] Ljung, G.M. and Box, G.E.P.(1978): *On a measure of lack of fit in time series models*, Biometrika , Vol. 65, No. 2, pp .297-303.

[30] Loader, C.R.(1999): *Classical or Plug-in?*, The Annals of Statistics, Vol. 27, No. 2, pp. 415-43.

[31] Met office,.(2011):*Global surface temperature*, Available: http://www.metoffice.gov.uk/research/monitoring/climate/surface-temperature, Last accessed 17th June 2014.

[32] National Climatic Data Center (NCDC): *Global Surface Temperature Anomalies — Monitoring References*, Available: http://www.ncdc.noaa.gov/monitoring-references/faq/anomalies.php, Last accessed 18th June 2014.

[33] National Climatic Data Center (NCDC), (2013): *Global Analysis - Annual 2013—State of the Climate*, Available: http://www.ncdc.noaa.gov/sotc/global/2013/13, Last accessed 17 th June 2014.

[34] National Climatic Data Center (NCDC), (2013): *Tornadoes - Annual 2013 —State of the Climate*, Available: http://www.ncdc.noaa.gov/sotc/tornadoes/2013/13, Last accessed 17 th June 2014.

[35] Opsomer, J. , Wang, Y. and Yang, Y. (2001) : *Nonparametric Regression with Correlated Errors*, Statistical Science,Vol. 16, No. 2 , pp. 134-153.

[36] Parzen, E.(1962): *On the estimation of a probability density and the mode*, The Annals of Mathematical Statistics, Vol. 33, No. 3 , pp. 1065-1076.

[37] Rosenblatt, M. (1956) : *On the Estimation of Regression Coefficients of a Vector-Valued Time Series with a Stationary Residual*, Annals of The Institute of Statistical Mathematics, Vol. 27, No. 1 , pp. 99-121.

[38] Ruppert,D. , Sheather,S. J. and Wand M. P.(1995): *An Effective Bandwidth Selector for Local Least Squares Regression* , Journal of the American Statistical Association,Vol. 90, No. 432 , pp. 1257-1270.

[39] Stone, C. (1977) : *Consistent Nonparametric Regression*, The Annals of Statistics, Vol. 5, No. 4, pp. 595-620.

[40] Wikipedia : *Bayesian information criterion*, Available:http://en.wikipedia.org/wiki/Bayesian-information-criterion,Last accessed 17 th June 2014.

[41] Xiao,Z., Linton O. B. , Carroll,R. J. and Mammen, E.(2003):*More efficient local polynomial estimation in nonparametric regression with auto-correlated errors*, Journal of the American Statistical Association, Vol. 98, No. 464, pp. 980-992.

[42] Yu, K. and Jones, M. C. (1998): *Local linear quantile regression*, Journal of the American Statistical Association, Vol. 93, No. 441, pp. 228-237.

[43] Yu, K. ,Lu, Z. and Stander, J. (2003) : *Quantile Regression: Applications and Current Research Areas*, Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 52, No. 3 , pp. 331-350.

[44] Yu, K. and Lu, Z. (2004) : *Local Linear Additive Quantile regression*, Scandinavian Journal of Statistics, Vol. 31, No. 3, pp. 333-346.

[45] Zheng, Q. ,Gallagher, C. and Kulasekera, K.B. (2013) : *Adaptively weighted kernel regression*, Vol. 25, No. 4, pp. 855-872.