Modeling electricity demand and fuel prices using nonparametric methods and extreme value theory

Levi Ng'ang'a Mbugua

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy in Applied Statistics in the Jomo Kenyatta University of Agriculture and Technology

2013

DECLARATION

This thesis is my original work and has not been presented for a degree in any other university.

Signature: ……………………………….. .   Date: …………………..

Levi Ng'ang'a Mbugua

This thesis has been submitted for examination with our approval as University Supervisors:

Signature: …………………………….......   Date: …………………….

Prof. Peter N. Mwita

JKUAT, KENYA

Signature: ………………………………......   Date: ………………….

Dr. Samuel M. Mwalili

JKUAT, KENYA

DEDICATION

To Pillen and Ann

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my supervisors: Professor Peter Mwita for introducing nonparametric quantiles to me and accepting to supervise me. Together with Dr. Samuel Mwalili, they sacrificed much of their valuable time to discuss this thesis with me at various stages, offering constructive criticisms, advice and suggesting amendments. Subsequently, their encouragement has been valuable towards the success of this thesis. Beside their tight schedules, they still had time with me throughout this period. My sincere thanks also goes to Professor Verdiana Grace Masanja of National University of Rwanda for her reviewing comments towards my work, Professor Cuthbert Mhilu of University of Dar es Salaam and Professor Adoniya B. Sebitosi of Stellenbosch University, South Africa for their fruitful conversations during various workshops and for continuously reviewing my papers.

I'm very grateful too to the National Council of Science and Technology for funding this undertaking. This financial support came at a much deserved time. I wish also to enlist the support of my colleagues: Dr. J. Mung'atu, O. Ngesa, J. Akinyi and others, without their encouragement this goal would have been an impossible task. The effort of The Chairperson Department of Statistics and Actuarial Sciences, Dr. G. Orwa should also be acknowledged and all the members of Statistics and Actuarial Science Jomo Kenyatta University of Agriculture and Technology. To all other friends of mine, I thank you so much for the varied support you gave me in the challenging period of my work. Finally, I remain indebted to my parents and To God the Almighty.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND ACRONYMS

AIC	Akaike Information Criterion
AICc	Improved Akaike Information Criterion
AR	Autoregressive
ARCH	Autoregressive conditional Heteroscedasticity
AMSE	Asymptotic Mean Square Error
ARIMA	Autoregressive Integrated Moving Average
Cdem	Change in electricity demand
CDF	Cumulative Distribution Function
Cp	Mallows bandwidth selection criterion
CV	Cross Validation
df	Degrees of freedom
EVT	Extreme Value Theory
ES	Expected Shortfall
ESc	Conditional Expected Shortfall
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
GCV	Generalized Cross Validation
GEV	Generalized Extreme Value
GPD	Generalized Pareto Distribution
iid	Independent and identically distributed

LL	Local Linear
MA	Moving Average
MLE	Maximum Likelihood Estimate
NEW	New Weighted Estimator
NW	Nadaraya-Watson
PDF	Probability Density function
POT	Peak over Threshold
pdem	Previous day demand of electricity
pprice	Previous day fuel price
RCP	Rice bandwidth determination criterion
VaR	Value at Risk
VaRc	Conditional Value at Risk
WNW	Weighted Nadaraya-Watson
WTI	West Texas Intermediate

ABSTRACT

Consumer behaviour towards different forms of utility varies over time. The variation can be so large that the estimated relationship between the response variable and its associated explanatory variables is seriously affected. In this study, kernel smoothing based conditional quantile approach, a nonparametric procedure is used to model volatile demand data. Nevertheless, quantile regression procedures work well in non extreme parts of a given data but poorly on extreme levels therefore we apply the threshold model of extreme value in order to circumvent the lack of observation problem at the tail of the distribution. It is shown that nonparametric estimation method has less bias relative to other standard methods when the underlying distribution is not known. Various kernel estimation methods and extreme value theory are discussed and the asymptotic properties of the estimators given. The methods are applied to model extremes in electricity demand and fuel price data. The underlying dynamics in the data inform of volatility clustering is also estimated using a standard Generalised Autoregressive Conditional Heteroscedastic (GARCH) model. A combination of nonparametric approach and extreme value theory will be shown as a method for estimation of value at risk. Value at risk is chosen in this work as it is extensively used in practice. The results indicate that electricity demand formation is influenced by time, behavioral variables and also by the forces of the market mechanism. It is also found that fuel prices play a crucial role in

influencing electricity demand. From the extreme value methods it is found that the goodness of fit depends on the estimated parameters that define the shape and behavior of the fitted distribution function. This indicates that the extreme value methods are case specific, which emphasizes the role of result validation. From these methods, it is found that maximum possible information can be extracted from the data and the threshold can be determined by calculation instead of subjective judgment. It is also easy to implement these methods by a complete program. With Generalized Pareto Distribution our estimates of value at risk and the expected shortfall for negative rate of change of fuel prices indicate that with probability 1% the daily rate of change of fuel prices could go as low as -1.3818% and given this rate of change, the average rate of change value will be 2.187%. Also with probability 5% the price daily rate of change could drop to -0.624% and that when it does the average fall is 1.404%. These results can be used to estimate risk measures in the energy related sectors as well as providing insights to producers of energy and also as a reference for actual or potential investors in the energy industry.

CHAPTER ONE

1.0 INTRODUCTION

1.1 Introduction to modeling energy demand and prices

Most of the developed countries have higher energy utilization rate while least developed countries are characterized by low energy demand and consumption. This means that energy availability accelerates development. Every countries energy mix involves a range of national preferences and priorities that are reflected in national policies. These policies represents a compromise between expected energy shortages, environmental quality, energy security, cost, public attitudes, safety and security, available skills and production and service capabilities. Relevant national stake holders must take all of these into account when formulating an energy strategy. In Kenya there have been many programs aimed at improving energy availability although all these efforts have not guaranteed Kenya any sustainable energy sufficiency. Ironically, due to increased electricity consumption and the extension of its power grid to rural constituencies, energy supply has led to severe environmental pressure and different energy utilization patterns which have not ensured sustainable resources development. Most of the energy planning carried out is on the short and medium term, within the framework of 10-15 years. However predictions put it that most of the petroleum fuel resources enjoyed now will last for no more than 30-40 years to come; Sokolov and McDonald (2006). Although renewable energy resources are at the disposal of most of the developing countries, higher costs and limited intensive applications put them to the shelf for some reasons; International Energy Agency (2010).

Kenya companies lose 9.5 percent of production because of power outages and fluctuations; Angelica et al. (2005). Efficiency in energy use is also another factor which has impeded the competitiveness of the country products in the international market. This has made the government to be under intense pressure to be a model player in the control of the energy prices. Similarly the private sector which deals with energy issues seems to be reluctant of the high change in prices. Nevertheless the pursuit of these goals could have a serious impact on the efficiency of the market. Given these differences, many researchers have sort to ascertain whether the trend in energy demand can be projected with a high level of certainty and accuracy.

Estimation and projection of energy demand in the current global economy has become a crucial task in the risk management market. This is due to the unstable regulatory and political environment, lack of information and inefficiency in the market. To this end, mathematical modeling offers the solution to this type of projection. Models are useful because of their interpretability and secondly in modeling theory supports inference and interpretation is easy. Mathematical models describe the pattern of change and researchers gain more insight on the underlying casual relationship related to pattern and structures of measurements as a result of being taken at different times.

The daily rate of change in electricity usage at the macro level can be attributed to the extreme changes in its production cost as a result of ever changing requirements. As a result, electricity is seen effectively more like a service than a commodity by nature. The physical limitations related to the production, consumption and delivery make it an “instantaneous” product, due to the fact that electricity can not be stored in bulk. Understanding electricity demand needs has thus become a significant element of utmost necessity of the planning exercise in the energy sector. The advent of the concept smart grid, which entails minimum losses of energy from development to transmission and finally use, makes it crucial for the synchronization of electricity production and consumption.

A lot of studies have been done on electricity development in Africa. Mhilu (2007) applies the multiple regression techniques for the development of regional equations for the estimation of low flow regimes of small catchments. On power transmission, Sebitosi, (2010), Sebitosi and Okou, (2010) discusses the transmission of energy in sub-Saharan Africa and the wastage associated with the transmission and came up with electricity tariff models on energy efficient. They conclude that there is need to understand electricity demand characteristic over time.

There are several reasons for studying change in utility demand over time, but two main areas of interest can be distinguished. First, one may be interested in obtaining the structural change reference curves over time and variation of change patterns based on observed data. The second motivation is mainly scientific and relates to having some knowledge about the relationship between the explanatory variable and the response variable which includes interesting special features like monotonicity, unimodality and other characteristics including the size of the extreme. There are also governmental implications which include the qualitative understanding of Influxes with time.

Regression analysis is one of the most commonly used techniques in statistics to study rates of change. The aim of the analysis is to explore the relationship between the explanatory and the response variables with the aim of understanding the contribution of the explanatory variables and their impact on the response variable. The simplest regression models, which are linear models, are desirable because they are easy to work with both analytically and computationally. The results from linear models are easy to interpret and there is a wide variety of useful techniques for testing the assumptions involved. None the less, there are cases where the linear models should not be applied because of an intrinsic nonlinearity in the data. To attenuate these challenges, nonparametric regression provides a means of modelling such data.

Nonparametric methods are statistical techniques that do not require a researcher to specify the distribution of the function being estimated. Instead the data itself informs the resulting model in a particular manner. In regression framework, this approach is also known as “nonparametric smoothing”. The methods we survey are known as kernel methods. They are best suited to situations involving large data sets for which the number of variables involved is manageable. These methods are often employed when common parametric specification is found to be unsuitable for the problem at hand. Though kernel methods are popular, they are but one of the many approaches towards construction of flexible models. Other approaches to flexible modelling include Spline, nearest neighbour, neural network and a variety of series methods. One of the most popular methods of nonparametric kernel regression was proposed by Nadaraya (1964) and Watson (1964), also known as the local constant which forms the basis of this study.

When the distribution is skewed and the conditional variance is not constant, a more flexible approach, the conditional quantiles becomes appropriate. The estimation of extreme conditional quantiles has become an increasingly important issue in effective risk management. This is because extreme events often lead to failure and losses and secondly it may be appropriate to be equipped with hedge models due to the nature of unobservable extra ordinary occurrences. Since data sparseness is more severe in extreme quantiles, fully nonparametric methods do not yield reliable estimates. Therefore nonparametric quantile regression methods need to be refined with extreme value theory so as to proficiently model extreme quantiles accurately, more specifically when quantifying risk measures.

The use of quantitative risk measures has become an essential management tool to be placed parallel with models of rates of change and returns in finance. These measures are used for investment and supervisory decisions, risk capital allocation and external regulation when there are extra ordinary occurrences. Value at risk has become a standard measure of risk employed by financial and economic institutions and their regulators. This is essentially due to its conceptual simplicity. Value at risk summarizes many complex bad and good outcomes to a single number, naturally representing a compromise between the needs of different users. This compromise has received a blessing of a wide range of users and regulators.

Measurement of risk associated with commodity markets is a relatively new field of research and surprisingly smaller number of papers deals with this topic. Oil price risk management has not been extensively studied but oil volatility and dynamics have been studied to some extent. Thus this study also investigates the distribution which best fit the extreme electricity and fuel data. This information is relatively important for risk management purposes as well as for pricing of structured commodity derivatives.

1.2 Statement of the problem

In the recent past the energy market in Kenya has been hit by a series of crisis ranging from electricity fluctuations to fuel shortage. Some believe that this is due to manipulation of imports of energy related products with the aim of pushing up prices. It is also believed that companies which have the largest share in Kenya's fuel market have intentionally been importing less fuel than they have committed themselves. This has led to a high cost of energy which is one of the biggest bottlenecks to the country economic advancement. Given these differences, many researchers have sort to ascertain whether the trend in energy demand can be projected with a high level of certainty and accuracy. Thus we are justified to understand the energy demand characteristics over time so that we can be able to project future demand.

The extreme rate of change of energy production factors expose the electricity producers and retailers to significant risks, which they can hedge provided that they have good modeling tools of which they can forecast these changes well enough. There exists a wide range of mathematical approaches to modeling and forecasting change, but often do perform poorly in case of extreme events; Byström (2005). Modeling the changes by distributions with finite variance is known to be inappropriate, since changes in prices in financial markets do not follow a Gaussian distribution, but are rather modeled better by stable distributions. Thus we need a model that captures the complete underlying structure of electricity demand and the extreme behaviour of the rate of change of fuel prices. We need a technique to analyze a portfolio and make forecasts of the likely losses that would be incurred in the market. Particularly, quantifying potential losses due to negative fluctuations of a portfolio market value is of particular relevance. This will help managers to asses the amount of capital reserves to maintain and to help guide their purchases and sales of various classes of financial assets. Failure to sufficiently capture the underlying patterns of the probable risks can have very serious implications.

In the past, risk analysis was done qualitatively but now with the advent of powerful computing software, quantitative risk management can be done quickly and effortlessly. When extreme observations occur in a data set and one understands its pattern and distribution, its analysis is straight forward. However in many scenarios, the distribution of extremes is not well known. To overcome this challenge, we need to come up with methods which do assume that the rate of change of the observed has a distribution which is not known.

1.3 Objectives of the study

The aim of this study is to develop statistical models that define the structural change of a time series data. The specific objectives are to

(a) Analyse the underlying Kenya electricity demand distribution, finding out interesting special features like monotonicity, unimodality and the size of extrema in the time series data.

(b) Determine the optimal method of nonparametric smoothing parameter selection for electricity demand data among the different methods available in literature.

(c) Determine the Kenya electricity consumers response to change in fuel prices on the fast growing electricity demand.

(d) Model extreme fuel price changes by detecting any special sort of dependency over time, quantifying potential losses due to negative price fluctuations.

This study shall answer the questions:

(i) Is there any pronounced trend in the electricity demand and fuel prices?

(ii) What is the nature of volatility in the electricity demand and fuel prices?

(iii) What are the risks associated with investing in the fuel industry?

1.4 Significance of the study

Electricity demand varies substantially from time to time while its supply tends to be relatively stable. This makes electricity as a commodity to behave quite differently from most other commodities. Over the last couple of years, electricity demand and fuel prices have exhibited extreme changes. These sudden changes have contributed to a climate of uncertainty for energy companies and investors and a climate of distrust among consumers and regulators. The unexpected changes are fundamentally determined by supply and demand imbalances of which commodity market participants strongly focus on economic models which relate supply and demand to fundamental market variables. Since our motivation is mainly scientific we expect this study to give insight on the underlying change process over certain phases and get a clear quantitative feature in electricity demand fluctuations overtime, we also expect the study to help us identify entry points for planning electricity management, savings and conservation interventions.

For risk management and regulatory reporting purposes, a business may need to estimate the lower bound on the changes in the value of a portfolio which will hold with high probability so as to help determine whether interventions are required. With some hidden information about market movements, the challenge has been to find a suitable model of the extreme conditional time varying statistics for risk measurement that is able to adapt to the rates of change distribution. Potential investors need to understand the dynamics in risks and hence arrive at a more cautious approach when making decisions. Since parametric curve estimation often does not meet the need of flexibility in data analysis, the nonparametric approach makes it possible to estimate functions of greater complexity and suggests other distributions. Nonparametric approach also provides a versatile method of exploring a general relationship between two variables, giving predictions of observations yet to be made without reference to a specific parametric model. It provides a tool for finding spurious observations and constitutes a flexible method of interpolating between adjacent explanatory variables. In addition to the further knowledge that will be created by this study, the findings will be a stepping stone for future research.

1.5 Literature Review

Nonparametric methods have attracted a great deal of attention from statisticians in the past few decades as evidenced by the arrays of text for the methods which are best suited to situation in which one knows little about the functional form of the parameter being estimated and also when the number of covariates in the model are small and the researcher has a reasonably large data set. In particular, the estimation of conditional quantiles has gained particular attention because of their useful applications in various fields such as econometrics, finance and other related fields. The first published paper on kernel estimation, a special type of nonparametric method appeared in

Rosenblatt (1956). Modeling both continuous and discrete time series phenomenon has been a basic analytical tool in modern statistics since the seminal papers by Sharpe (1964), Black and Scholes (1973). The rationale behind these papers is that news arrives at the market in both continuous and discrete manners and end up having different impact. Several authors have studied the asymptotic properties of nonparametric estimation, such as kernel and nearest neighbor. They include Stone (1977) and Bhattacharya and Gangopadhyay (1990). Some finite-sample properties of regression quantiles were discussed by Koenker and Basset (1978) and their asymptotic behavior were further developed by Ruppert and Carroll (1980). Koenker and Basset (1982) used regression quantile techniques to test heteroscedasticity and Powell (1986) applied the idea to censored data in econometrics.

It is well known that kernel type procedures have serious draw backs namely boundary effects and the asymptotic bias involves the design density. To eliminate these drawbacks, Fan et al. (1994) proposed the use of the "check function" such as a robustified local linear smoother. This was further extended by Yu and Jones (1998) "double-kernel" procedure. An alternative procedure is first to estimate the conditional distribution function by using double kernel local linear technique of Fan *et al.* (1996) and then to invert the conditional distribution estimator to produce an estimator of a conditional quantile which is called the Yu and Jones estimator. Fan and Gijbels (1996), used quantiles to quantify the extent to which the poor got poorer and the rich got richer during the Reagan administration (1981-1988). Although Local linear methods have some attractive properties, they have a disadvantage of producing conditional distribution function estimators that are not constrained either to lie between zero and one or to be monotone increasing. Hall, *et al.* (1999) proposed a reweighted version of the Nadaraya-Watson estimator which is designed to posses the superior properties of local linear methods such as bias reduction and no boundary effect and to preserve the property of the Nadaraya-Watson estimator that it is always a distribution function. Pagan and Ullah (1999) and Horowitz (2001) provided examples on how conclusions drawn from a convenient but incorrectly specified model more especially parametric models can be very misleading leading to misspecification error.

According to Parzen (1962), the mean can be a misleading summary of a distribution; one should always plot a quantile function to check for skewness and tails of outliers. Cai (2002), studied nonparametric estimation of regression quantiles for time series data by inverting the Reweighted Nadaraya-Watson estimator of conditional distribution. He established the asymptotic normality and weak consistency. A disadvantage of nonparametric estimation is the low frequency observations at the tails which lead to the estimation which exhibits a very high volatility. The variance is very high in some cases even infinite. This will result to poor estimates of the tails which are very crucial for risk measures estimation. Due to the sparseness of data in extreme regions, the nonparametric kernel methods do not guarantee reliable description of the tails.

The theory of extreme values mitigates this problem by introducing a parametric distribution functions at the tails.

The extreme quantiles can be estimated by using ideas from Extreme Value Theory (EVT). The use of EVT in financial market calculations is a fairly recent innovation. Embrechts *et al.* (1997) surveys the mathematical theory of EVT and discusses its applications. The EVT can be used to characterize the behaviour of the extreme returns and extreme rate of change or the extreme tails distribution without tying the analysis down to a single parametric family fitted to a whole distribution. Because of the presence of stochastic volatility and some distinct important stylized facts such as persistent volatility clustering, heavy tails, strong serial dependence and occasionally sudden but large jumps in financial and econometric data, it is inappropriate to apply these models directly since they are nested in a frame work of identical and independent distributed variables which is not consistent with the aforementioned characteristics. Danielsson and de Vries (1997) have shown that these models do not work well in the common low probabilities such as 0.95. Attempts have been made to extend extreme value methodology to take into account volatility. They include Barone-Adesi et al. (1988), McNeil and Frey (2000), and Mwita, (2003). Their approach revolves around the $GARCH$, with heavy tailed innovation. Engle and Manganelli (2004) used regression quantile methodology to determine the unknown parameter estimates of Value at Risk (VaR) under the assumption that the quantile process is correctly specified. Application of extreme value modeling have been published in the field of alloy strengthen prediction, Tryon and Cruse (2000), ocean wave modeling; Dawson (2000) memory cell failure; McBulty et al. (2000), wind engineering; Harris (2001), thermodynamics of earthquakes; Lavenda (2001), nonlinear beam vibration; Dunne and Ghanbari (2001). Other areas include management strategy and biomedical data processing. The distinguishing feature of an extreme value analysis is the objective to quantify the stochastic behavior of a process at usually large or small levels.

In the study of electricity demand, Lindley and Smith (1972) introduced some variant of hierarchical linear models in the research of household electricity demand. An excellent exposition of the statistical foundations of such hierarchical models from Bayesian standpoint may be found in Smith (1973). A classical survey of the studies on the demand for electricity was given by Taylor (1975) and it was later on updated and extended to natural gas, heating fuels and gasoline by Taylor (1977). Taylor concluded that the price elasticity of demand for electricity for all classes of consumers is much larger in the long run than in the short run. Hendricks and Koenker (1991) used hierarchical spline parameterization of the conditional quantiles for household electricity demand using data from Chicago metropolitan and found that there was a very strong periodic component and weather impact on base load (25% quantile), while estimates at the 95% quantile had a strong periodic shape. Hausman and Newey (1995) used kernel estimates of demand functions to estimate the equivalent variation for changes in gasoline prices and the dead weight losses associated with increases in gasoline taxes. Blundell *et al.* (2003) used kernel estimate of Engel curves in the investigation of the consistency of household-level data and came up with the preference theory. Beirlant *et al.* (2004) used the American electric utility company's data to study the relationship between input and output of firms in a productivity analysis. They proposed a flexible nonparametric two stage procedure for estimating extreme quantiles in a regression setting. They combined the merits of local polynomial quantile regression (first step) and recent extreme value methods (second step).

Oil price risk management has not been extensively studied but oil volatility and dynamics have been studied to some extent among others, Birol (2001). The literature on measuring financial risk and volatility via VaR models in financial industry is vast and is discussed in details by Jorion (2001). Giot and Laurent (2003) investigated commodity futures including US benchmark oil West Texas Intermediate (WTI) returns in the period 1987-2002. They found that WTI returns are characterized by negative skewness and leptokurtosis and tested the performance of ARCH and Risk Metrics parametric models. In their study, Risk Metrics performed rather poorly at confidence levels above 99%. Zikovic and Fatur (2007) investigated WTI oil returns over the period 2000-2006 and also found negative asymmetry and leptokurtosis. They found that parametric normally distributed VaR provided correct unconditional coverage at 90%, 95% and 99% confidence levels both for long and short positions. These findings can probably be attributed to the fact that their out-of-sample period was relatively tranquil.

In this study we extend Beirlant *et al.* (2004) ideas by developing statistical models which investigate what type of distribution best fit the extreme tails of Kenya’s electricity demand and fuel prices rate of change. Non parametric methods will be used in order to find the quantiles. Secondly we shall apply the VaR model in measuring the risk occurring in the far negative tail of the rate of change distribution of fuel prices. VaR Models are calculated for a one-day holding period at 95%, 99% and 99.9% risk coverage.

1.6 Outline of the Thesis

This thesis is outlined as follows: In this chapter we introduce the work by giving the background information in nonparametric regression and extreme value theory. We also state the problem, objectives and the significance of the study. In the closing of the chapter a brief literature review on nonparametric methods and extreme value theorem is discussed. An overview of one dimensional smoothing tools as well as the key ideas on optimal smoothing parameters are given in Chapter 2. Nonparametric regressions for time series and models with exogenous variables are also explained. In Chapter 3, quantiles are introduced and their asymptotic properties are derived. In Chapter 4 different approaches of modeling extreme values and risk measures are discussed. In Chapter 5, we present the results and discussions. In Chapter 6, we conclude this study by citing some of the unresolved modeling issues and suggestions for further research in the market risk management.

CHAPTER TWO

2.0 NONPARAMETRIC REGRESSION ESTIMATION

2.1 Nonparametric design models

The general nonparametric regression models are either of fixed or of random design, such that if n data points, $(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$ have been collected, the relationship is modeled as

$$Y_i = m(X_i) + \varepsilon_i, \quad 1 \le i \le n \tag{2.1.0.1}$$

where, X_i is the predictor variable also known as the regressor, Y_i is the response variable and m is the unknown regression function with observation error ε_i of which $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. The fixed design model is concerned with controlled non stochastic regressor X variable, this implies that the regressors are controlled by the researcher and are simply assumed to be measured without error. In fixed design, for any given observation (x_i, Y_i), $x_i \in \Re^d$ and Y_i is an independent random variable with $E(Y_i) = m(x_i)$.

Random design models are used in observational studies and are common in non experimental science. The observed predictor variables are independent and identically distributed (iid) random variables so that

$$m(X) = E(Y|X) = \int_y y \frac{f(x,y)}{f(x)} \tag{2.1.0.2}$$

$f(x, y)$ is the joint density of (X, Y) and $f(x) = f_X(x)$ is the marginal probability function of x. Although the stochastic mechanism is different, the idea of fitting the mean function is the same in both cases.

2.2 Smoothing methods

The term smoothing can be defined as the approximation of the function $m(x)$. Every smoothing method is of the form $\widehat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} w_{ij}Y_{ij}$, where w_{ij} are the weights. The common smoothing methods include the kernel, the nearest neighbor, the orthogonal series and the spline method. We concentrate on kernel techniques.

2.2.1 Kernel estimation

Kernel estimators are linear estimators in that we can express the value of the estimator at any point x as the weighted sum of the responses. Let us define a weight function,

$$K_h(x - x_i) = \frac{1}{h}K\left(\frac{x - x_i}{h}\right) \qquad (2.2.1.1)$$

with K function supported on $[-1,1]$ that has a maximum at zero. This support holds with the exception of the Gaussian kernel. K , determines the shape of the weights and satisfies the moment conditions

$$\int_{-1}^{1} K(u)du = 1 \qquad (2.2.1.2)$$

$$\int_{-1}^{1} uK(u)du = 0 \tag{2.2.1.3}$$

$$M_2 = \int_{-1}^{1} u^2K(u)du \neq 0 \quad and \quad V = \int_{-1}^{1} K(u)^2du < \infty \tag{2.2.1.4}$$

Condition (2.2.1.2) is roughly equivalent to having the weights sum to one and condition (2.2.1.3) is a type of symmetry condition that is automatically satisfied if K is symmetric about zero. Kernel function include Gaussian kernel given by

$$K(u) = \sqrt{2\pi}^{-1} exp(-u^2/2) \tag{2.2.1.5}$$

and the "symmetric Beta family", which are of the form

$$K(u) = \frac{1}{\beta(1/2, \gamma + 1)}(1 - u^2)^\gamma, \ \gamma = 0,1, \dots, \tag{2.2.1.6}$$

When $\gamma = 0$ in (2.2.1.6) we have uniform kernel, when $\gamma = 1$ we have the Epanechnikov kernel, and when $\gamma = 3$ we have the triweight kernel

Table 1 *The different types of Beta Kernels*

γ	Name of the kernel	$K(u)$
0	Uniform	$\frac{1}{2}I(\|u\| \leq 1)$
1	Epanechnikov	$\frac{3}{4}(1 - u^2)I(\|u\| \leq 1)$
2	Biweight	$\frac{15}{16}(1 - u)^2I_{[-1,1]}(u)$
3	Triweight	$\frac{70}{81}(1 - \|u\|^3)^3I(u)$

There is no loss in assuming that K has support on $[-1,1]$. This is because any kernel with finite support can be rescaled to have support on$[-1,1]$. Kernels, which have infinite support on the entire line, result to estimators with global bias difficulties. Parameter (h) in (2.2.1.1) is called the bandwidth or smoothing parameter which determines the size of the weights. Small h leads to wigglier (rougher) estimators while larger h leads to a more averaging (horizontal) estimator. From (2.2.1.1), we obtain the estimator

$$m_h(x) = (nh)^{-1}\sum_{i=1}^{n} K\big(h^{-1}(x - x_i)\big)y_i \qquad (2.2.1.7)$$

If we generalize (2.2.1.7) and replace n^{-1} by $x_i - x_{i-1}$ which is more appropriate for equally spaced data, we obtain

$$m_h(x) = \sum_{i=1}^{n}(x_i - x_{i-1})h^{-1}K(h^{-1}(x - x_i))y_i \quad with\ x_0 = 0 \qquad (2.2.1.8)$$

and with expectation,

$$h^{-1}\sum_{i=1}^{n}(x_i - x_{i-1})K\big(h^{-1}(x - x_i)\big)m(x_i)$$

If we modify (2.2.1.7) to ensure that the weights sum to one, we have

$$m_h(x) = \sum_{i=1}^{n} K\big(h^{-1}(x - X_i)\big)y_i \Big/ \sum_{j=1}^{n} K\big(h^{-1}(x - X_i)\big) \qquad (2.2.1.9)$$

The use of estimator (2.2.1.9) when x_i is random was suggested by Nadaraya (1964) and Watson (1964).

If we replace $(x_i - x_{i-1})h^{-1}K\big(h^{-1}(x - x_i)\big)$in (2.2.1.8) with $h^{-1}y_i \int_{\gamma_{i-1}}^{\gamma_i} K\big(h^{-1}(x - \gamma)\big)d\gamma$ for large n we obtain

$$m_h(x) = \sum_{i=1}^{n} \left(h^{-1} \int_{\gamma_{i-1}}^{\gamma_i} K\big(h^{-1}(x-\gamma)\big)d\gamma \right) y_i \qquad (2.2.2.0)$$

where, $\gamma_0 = 0$, $\gamma_{i-1} \leq x_i \leq \gamma_i$, $i = 1, \dots, n-1$, $\gamma_n = 1$. Rather than a piecewise constant approximation as used in estimator (2.2.2.0) we use a piecewise linear approximation $\tau_y(\gamma)$ for the estimate $m(\gamma)$,$x_i < \gamma \leq x_{i+1}$, to obtain

$$m_h(x) = \int_0^1 h^{-1}K\big(h^{-1}(x-\gamma)\big)\tau_y(\gamma)d\gamma \qquad (2.2.2.1)$$

where,

$$\tau_y(\gamma) = \begin{cases} y_i & \gamma \leq x_i, \\ y_i\left(\dfrac{x_{i+1}-\gamma}{x_{i+1}-x_i}\right) + y_{i+1}\left(\dfrac{\gamma - x_i}{x_{i+1}-x_i}\right), & x_i < \gamma \leq x_{i+n} \\ y_n, & \gamma > x_n \end{cases}$$

Estimators (2.2.1.7) and (2.2.1.8) stem from the work of Priestley and Chao (1972). Estimator (2.2.2.0) was originally studied by Gasser and Muller (1979) and Cheng and Lin (1981). The closely related estimator (2.2.2.1) was proposed by Clark (1977).

2.2.2 Nadaraya Watson Estimator

Assume the observation of some variable Y have been taken n times for some utility at times $t_1 \cdots, t_n$. Let y_i, be decomposed into two parts, $m(\cdot)$ the regression function which represents the true underlying change curve following the economic and physical potential and ε_i as defined in (2.1.0.1). The errors ε_i may not depend on time. These errors not only stand for observational error but also for economic random variation due to seasonal and other exogenous factors. We assume that $0 \le t_1 \le t_2 \le \cdots \le t_n \le 1$ where $t_1, t_2, \dots, t_n$ is the explanatory variable analogous to $x_1, x_2, \dots, x_n$ for ease of notation. A kernel estimator $\widehat{m}(t_0)$ for $m(t_0)$ can be written as:

$$\widehat{m}(t_0) = \sum_{i=1}^{n} w_i(t_0; t_1 \cdots, t_n : h) y_i \qquad (2.2.2.3)$$

where w_i, are the weights given by

$$w_i = (t_0; t_1, \cdots, t_n; h) = \sum_{i=1}^{n} K_h(t - t_i) / \sum_{j=i}^{n} K_h(t - t_j) \qquad (2.2.2.4)$$

The weights do not depend on $\{Y_i\}$ and therefore $\widehat{m}(t_0)$ is a linear estimator which can be expressed as a minimiser of the locally weighted least squares

$$\widehat{m}(t_0) = \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (t_1 - t_0)^j \right\}^2 K_h(t_i - t_0) \qquad (2.2.2.5)$$

Where the squared part of the right hand side of (2.2.2.5) represents the polynomial part and the other part represents the local constant. While the sum ranges from 1 to n, only those y_i lying in the interval $(t_0 - h, t_0 + h)$ contribute to $\widehat{m}(t_0)$. For simplicity, we define (2.2.2.4) as $\hat{g}(t)/\hat{f}(t)$ which are the finite sample approximation to

$$g(t) = \int_{-\infty}^{\infty} y f(t, y) dy \qquad \text{and} \qquad f(t) = \int_{-\infty}^{\infty} f(t, y) dy$$

We first examine the behaviour of the Nadaraya-Watson (NW) estimator when the observed pairs of data $(t_1, y_1), \ldots, (t_n, y_n)$, are identical and independent. In order to understand the estimator as a whole, we begin with lemmas for the asymptotic performance of the numerator and denominator, $\hat{g}(t)$ and $\hat{f}(t)$ as they approximate $g(t)$ and $f(t)$. We first impose some restrictions on the behaviour of the bandwidth h as the sample size grows large and on the conditional distribution of the errors.

Assumption 2.2.1 *As the sample size* $n \to \infty$ *the bandwidth* $h \to 0$ *in such a way that* $nh \to \infty$

Assumption 2.2.2 $E[\varepsilon_i | t = t_i] = 0, and\ E[\varepsilon_i^2 | \boldsymbol{t}_i = t] = \sigma^2(t) < \infty$

Under these assumptions, $\hat{f}(t)$ and $\hat{g}(t)$ are consistent estimators of $f(t)$ and $g(t)$. This is quantified in the following lemma

Lemma 2.1 *If* t *is contained in an open interval of which* $f(t)$ *has* p *bounded continuous derivatives and* $m(t)$ *has* q *bounded continuous derivatives, then under Assumption 2.2.1 and Assumption 2.2.2 we have*

$$E[\hat{f}(t)] - f(t) = o(h^p) \qquad (2.2.2.6)$$

$$E[\hat{g}(t)] - g(t) = o(h^k) \qquad (2.2.2.7)$$

where, $k = min\{p, q\}$. If both $f(t)$ and $g(t)$ are infinitely differentiable, then each of these biases become $o(h^m)$ for all positive real m.

If we impose the additional assumptions that the observed pairs of data are identical and independent random variables and that $f, g \; and \; \sigma^2(t)$ are infinitely differentiable, then the variance of $\hat{f}$ and $\hat{g}$ are also differentiable.

The next assumption is necessary to ensure that the asymptotic approximations we use are valid and to avoid division by zero.

Assumption 2.2.3 *The point t is a continuity point of $\sigma^2(t), f(t) > C$ for some $C > 0$ and m and f are each differentiable in a neighbourhood of t*

Lemma 2.2 *Under Assumption 2.2.1-2.2.4*

$$var[\hat{f}(t)] = \frac{f(t)}{nh} \int_{-\infty}^{\infty} K^2(u)du + o\left(\frac{1}{nh}\right) + O\left(\frac{1}{n}\right) \tag{2.2.2.8}$$

$$var[\hat{g}(t)] = \frac{\left(m^2(t) + \sigma^2(t)\right)f(t)}{nh} \int_{-\infty}^{\infty} K^2(u)du + o\left(\frac{1}{nh}\right) + O\left(\frac{1}{n}\right) \tag{2.2.2.9}$$

and

$$cov[\hat{f}(t), \hat{g}(t)] = \frac{m(t)f(t)}{nh} \int_{-\infty}^{\infty} K^2(u)du + o\left(\frac{1}{nh}\right) + O\left(\frac{1}{n}\right) \tag{2.2.3.0}$$

The joint asymptotic normality of $\hat{f}$ and $\hat{g}$ will be established via the Liapunov condition, which requires a uniform bound on the $2 + \epsilon \, th$ moments of y_i, for some $\epsilon > 0$, such that

$$r_n^{2+\epsilon} = \sum_{i=1}^{n} (|t_i - \mu_i|^{2+\epsilon})$$

Then if $\lim_{n\to\infty} \frac{r_n}{S_n} = 0$, where $S_n^2 = \sum_{i=1}^{n} \sigma_i^2$ and for all $a < b$, we have

$$\lim_{n\to\infty} P\left(a < \frac{S_n}{s_n}\right) = \Phi(b) - \Phi(a)$$

where $\Phi(\cdot)$ is the normal distribution function.

Assumption 2.2.4 *There exists positive constant M and ϵ such that for all t*

$$E[y_i|^{2+\epsilon}|\boldsymbol{t}_i = t] < M$$

The final assumption forces the conditional variance of the errors to be bounded above and below for all t. The bound from below is assumed for technical simplicity.

Assumption 2.2.5 *There exist positive constant b and B such that $b < \sigma^2(t) < B$ for all t*

Lemma 2.3 *Under Assumptions 2.2.1-2.2.5, for all real c_1 and c_2 (not both zero),*

$$\sqrt{nh}\left[c_1\left(\hat{f}(t) - E\left[\hat{f}(t)\right]\right)\right] + c_2\left(\hat{g}(t)\right) - E[\hat{g}(x)] \xrightarrow{D} N\left(0, \omega(t)\right)$$

where,

$\omega(t) = \left(c_1^2 + 2c_1c_2m(t) + c_2^2[m^2(t) + \sigma^2(t)]f(t)\int_{-\infty}^{\infty} K^2(u)du\right).$

This implies the joint asymptotic normality of $\hat{f}$ and $\hat{g}$. Once this has been shown, a Taylor's series argument can be employed to show that $\widehat{m}$ has an asymptotic normal distribution.

The performance of the estimator $\widehat{m}(t)$ of the regression function and its vth derivative; $m^{(v)}(t)$ is assessed via its mean squared error (MSE) and mean integrated squared error $(MISE)$. An estimator is an (MSE) consistent iff both the bias and variance of the estimator, approach zero. Convergence in mean square implies convergence in probability. Define

$$MSE(t) = E\{\widehat{m}(t) - m(t)\}^2$$

and mean integrated squared error

$$(MISE) = \int MSE(t)w(t)dt$$

with $w \geq 0$ a weight function. When MSE is used the main objective is to estimate the function $m(\cdot)$ at the point t and when $MISE$ is used the main goal is to estimate the whole curve. It can be seen that MSE has the following bias-variance decomposition:

$$MSE(t) = E\{\widehat{m}(t) - m(t)\}^2 + Var\{\widehat{m}(t)\}$$

This is shown in the following theorem by Cai (2002);

Theorem 2.1 *If t is contained in an open interval on which $f(t)$ has p bounded continuous derivatives and $m(t)$ has q bounded continuous derivatives, then under Assumptions 2.2.1-2.2.6*

$$\sqrt{nh}\Big(\widehat{m}(t) - m(t) + o(h^k)\Big) \xrightarrow{D} N\left(0, \frac{\sigma^2(t)}{f(t)} \int_{-\infty}^{\infty} K^2(u)du\right)$$

Where, $k = min\{p, q\}$. To make this intuitive argument more precise, the MSE of $\widehat{m}(t)$ can be approximated by

$$MSE\{\widehat{m}(t)\} \cong \mu_2^2 \left\{\frac{m'(t)f'(t)}{f(t)} + \frac{1}{2}m''(t)\right\}^2 h^4 + \frac{1}{nh}R(K)\frac{\sigma^2(t)}{f(t)} \qquad (2.2.3.1)$$

where $f(t)$, is the marginal probability distribution function (PDF) of t, $\mu_2 = \int u^2 K(u)du$, $R(K) = \int K^2(u)du$ and $\sigma^2(t) = E(\varepsilon^2|t = t)$

From (2.2.3.1), while the bias increases as a square of the bandwidth h, the variance decreases hyperbolically with h. Moreover, the variance depends on the data only via the residual variance $\sigma^2(t)$, and the bias only on the first and the second derivative of the underlying regression function m. Quantitatively, bias leads to a flattening of peaks and valleys and little distortions in flat parts of the curve. Bias is “conservative” in that it dampens the true structure but does not generate artificial structure. Ideally from statistical perspective, the choice of K and h should lead simultaneously to low bias and low variability.

We see that Nadaraya-Watson locally uses one parameter less compared to local linear without reducing the asymptotic variance. It suffers from large bias and it does not adapt to non uniform designs (the bias can be very large when $f'(x)/f(x)$ is large), it has zero minimax efficiency unless $f'(x_0)$ = 0. It also possesses a larger bias when estimating a curve at a boundary region and is not design adaptive; it assigns symmetric weights to both sides for non equispaced design.

Despite these shortcomings, this estimator has the property of positivity where the conditional distribution is constrained to lie between 0 and 1 and monotonicity (monotone increasing). These properties are advantageous if the method of inverting conditional distribution estimator is applied to obtain an estimator of a conditional quantile. To overcome these difficulties, Hall *et al*. (1999) proposed a weighted version of the Nadaraya-Watson estimator.

2.2.3 The Weighted Nadaraya-Watson Estimator

Let $p_i = p_i(t)$ for $1 \leq i \leq n$, denote weight functions of the data $\boldsymbol{t}_1, \dots, \boldsymbol{t}_n$ with the property that each $p_i \geq 0$, $\sum_i p_i = 1$ and

$$\sum_{i=1}^{n} p_i(t)(\boldsymbol{t}_i - t)K_h(\boldsymbol{t}_i - t) = 0 \qquad (2.2.3.2)$$

Of course, $p_i's$ satisfying these conditions are not uniquely defined and we specify them by maximizing $\prod_i p_i$ subject to the constraints $\sum_{i=1}^{n} p_i(t) = 1$ and (2.2.3.2). Computation of the $p_i's$ is simplified by the fact that

$$p_i(t) = n^{-1}\{1 + \lambda(t - \boldsymbol{t}_i)K_h(\boldsymbol{t}_i - t)\}^{-1},$$

Where λ, a function of the data and of t is uniquely defined by (2.2.3.2). This is computed by using the successive approximation $t_{n+1} = t_n - \frac{f(t_n)}{f'(t_n)}$, where t_n is the approximation root of $f(t)$. The Weighted Nadaraya-Watson estimator (WNW) is therefore

$$\widehat{m}(t) = \{\textstyle\sum_{i=1}^{n} w_i(t) I(Y_i \leq y)\}/\{\sum_{i=1}^{n} w_i(t)\}, \qquad (2.2.3.3)$$

where, $w_i(t) = p_i(t)K_h(t_i - t)\sum_{j=1}^{n} p_j(t)\left(t - t_j\right)^2 K_h\left(t - t_j\right)$

When $\lambda = 0$, then $\widehat{m}(t)$ becomes the classical Nadaraya-Watson estimator. Thus $p_i(t)$ is a constant for all i making $\widehat{m}$ to be monotone in y and $0 \leq \widehat{m}(t) \leq 1$. Hall *et al*. (1999) showed that $\widehat{m}$ is first order equivalent to a local linear estimator.

2.2.3.1 Sampling properties of Weighted Nadaraya-Watson Estimator

We now impose the following regularity conditions:

Condition 2.2.1

(A1). For fixed y *and* t, $g(t) > 0$, $0 < m(t) < 1$, $g(\cdot)$ *is continuous at* t *and* $m(t)$ *has continuous second order derivative, in the neighbourhood of* t.

(A2). The kernel $K(\cdot)$ *is symmetric density satisfying* $C_1 = sup_u|uK(u)| < \infty$

(A3). As $n \to \infty, h \to 0$ *and* $nh^3 \to \infty$

(A4). Let $g_{1,i}(\cdot,\cdot)$ *be the joint density of* t_1 *and* t_i *for* $i \geq 2$. *Assume that*

$\left|g_{1,i}(u,v) - g(u)g(v)\right| \leq M < \infty$, for all u and v

Theorem 2.2 *Suppose Condition 2.2.1 holds, then, as* $n \to \infty$,

$$\hat{m}(t) - m(t) = \frac{1}{2}h^2\mu_2 m''(t) + o_p(h^2) + O_p\left((nh)^{-1/2}\right) \quad (2.2.3.4)$$

In addition,

$$\sqrt{nh}\left[\hat{m}(t) - m(t) - B(y|t) + o_p(h^2)\right] \xrightarrow{D} N\left(0, \sigma^2(y|t)\right) \quad (2.2.3.5)$$

where $B(y|t)$ is the bias term. Also, it can be seen that the asymptotic MSE is

$$MSE_{WNW} = \frac{1}{4}h^2\mu_2^2\{m''(t)\}^2 + \frac{v_0}{nh}\frac{\sigma^2(t)}{f(t)} \quad (2.2.3.6)$$

Remark: From the WNW estimator, $\hat{m}(t) \to m(t)$ in probability with a rate, which of course implies that, $\hat{m}(t)$ is consistent. Also the estimator does not depend on the asymptotic bias in the design density $f(\cdot)$. Its dependence is on the simple conditional distribution curvature $m''(t)$.

Table 2 The bias and variance of Nadaraya Watson, Local Linear and Weighted Nadaraya Watson estimators.

Estimator	Bias	Variance
(NW)	$\frac{1}{2}h_n^2\left[m''(t)+m'(t)\frac{f'(t)}{f(t)}\right]+\mu_2$ $+\,o(h_n^2)$ $+\,O((nh_n)^{-1})$	$(nh_n)^{-1}v_0\sigma^2+o\{(nh_n)^{-1}\}$
(LL)	$\frac{1}{2}h_n^2m''(t)\mu_2+o(h_n^2)+O(n^{-1})$	$(nh_n)^{-1}v_0\sigma^2+o\{(nh_n)^{-1}\}$
(WNW)	$\frac{h^2\mu_2m''(t)}{2}$	$\frac{v_o\sigma^2(t)}{nhf(t)}$

2.3 Smoothing Parameter Selection

In nonparametric kernel estimation, the smoothing parameter effectively controls the model complexity. When $h=\infty$, local modeling becomes a global modeling, when $h=0$ the estimate essentially interpolates the data and the modeling bias will be small. This can be seen in (2.2.3.6) where the consistency of the estimator is basically based on the sum of the bias and variance. Since the bias is proportional to h^2 and the variance proportional to $\frac{1}{h}$, the bandwidth has to be taken neither too large nor too small so as not to increase the bias and variance of the estimates. The problem can be solved theoretically by choosing a bandwidth that balances the trade-off between the bias and the variance components. The positive value h that minimizes any of the selection criteria is selected as an optimal smoothing parameter.

(i) Cross-Validation (CV): The basic idea of CV is to leave the points $\{x_i, y_i\}_{i=1}^{n}$ out one at a time and select the smoothing parameter h that minimizes the residual sum of squares and to estimate squared residuals for a smooth function at x_i based on the remaining $(n-1)$ points. The CV score function to be minimized is given by

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{f}_h^{(-1)}(x_i)\right\}^2 \equiv CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_h(x_i)}{1-(S_h)_{ii}}\right\}^2 \quad (2.3.0.1)$$

where $\hat{f}_h$ is the fit (smoother) for n pairs of measurements $\{x_i, y_i\}_{i=1}^{n}$ with smoothing parameter h and $\hat{f}_h^{(-1)}$ is the fit calculated by leaving out the $i\,th$ data point and $(S_h)_{ii}$ is the $i\,th$ diagonal element of the smoother matrix S_h. S_h is an $n \times n$ hat matrix depending on the x variate and the smoothing parameter. This method is useful for assessing the performance of an estimator via estimating its prediction error.

(ii) Generalized Cross Validation (GCV): It is an improved version of (CV) in terms of computation. The main idea of GCV is to replace the factors $1-(S_h)_{ii}$ in (2.3.0.1) with the average score $1-n^{-1}tr(S_h)$, where $tr(\cdot)$ is the trace of the hat matrix. Thus the GCV approach selects the bandwidth h that minimizes:

$$GCV(h) = \frac{1}{n}\frac{\sum_{i=1}^{n}\{y_i - \hat{f}_h(x_i)\}^2}{\{1 - n^{-1}tr(S_h)\}^2} = \frac{n^{-1}\|(1 - S_h)y\|^2}{[n^{-1}tr(1 - S_h)]^2} \qquad (2.3.0.2)$$

A drawback of the cross-validation type method is its inherent variability. Further, it cannot be directly applied to select bandwidths for estimating derivative curves. As pointed out by Fan, *et al.* (1995), the cross-validation type method performs poorly due to its large sample variation, even worse for dependent data. Plug-in methods avoid these problems. The basic idea is to find a bandwidth h minimizing estimated mean integrated square error. To overcome the over-fitting or under-fitting tendency and fit the time series data discussed in the next section, we propose the nonparametric version of the Akaike Information criterion.

(iii) Improved Akaike Information Criterion (AIC_c): The basic idea is described as follows: by adopting the classical AIC for linear models under likelihood setting

$$-2(maximized\ loglikelihood) + 2(number\ of\ estimated\ parameters)$$

select h minimizing

$$AIC_c = log\frac{\sum\{y_i - \hat{f}_h(x_i)\}^2}{n} + 1 + \frac{2\{tr(S_h) + 1\}}{n - tr(S_h) - 2}$$

$$= log\frac{\|(S_h - I)y\|^2}{n} + 1 + \frac{2\{tr(S_h) + 1\}}{n - tr(S_h) - 2}$$

$$= log(\hat{\sigma}^2) + \varphi(tr(S_h), n) \qquad (2.3.0.3)$$

where, $\varphi(\cdot)$ is chosen particularly to be the form of the bias corrected version of the AIC. $tr(\cdot)$ is the trace of the smoothing matrix regarded as the nonparametric version of degrees of freedom, called the effective number of parameters.

when $\varphi(tr(\cdot)) = -2log(1 - tr(S_h)/n)$, then (2.3.0.3) becomes the generalized cross validation criterion.

(iv) When $\varphi(tr(\cdot)) = -log(1 - 2tr(S_h)/n)$, (2.3.0.3) becomes the T criterion (RCP) proposed and studied by Rice (1984) for identically and independent variabbles. When $tr\,(S_h)/n \to 0$, then the nonparametric AIC, the GCV, and RCP are asymptotically equivalent. However the RCP requires $tr\,(S_h)/n < 1/2$ and when $tr\,(S_h)/n$ is large, the GCV has relatively weak penalty

Other bandwidth selection criterion includes

(v) Mallows' C_p criterion: when σ^2 is known, an unbiased estimate of the residual sum of squares is given by C_p criterion of Mallows (1973):

$$C_p(h) = \frac{1}{n}\{\|(S_h - I)\|^2 + 2\sigma^2 tr(S_h) + \sigma^2\}$$

$$= \frac{1}{n}\left\{\left\|y - \hat{f}_h\right\|^2 + 2\sigma^2 tr(S_h) + \sigma^2\right\} \qquad (2.3.0.4)$$

Unless σ^2 is known in practice σ^2 is approximated by

$$\hat{\sigma}^2 = \hat{\sigma}^2_{\hat{h}} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{f}_h(x_i)\right)^2}{tr(1 - S_h)} = \frac{\|(S_h)y\|^2}{tr(1 - S_h)} \qquad (2.3.0.5)$$

where $\hat{h}$ is pre-chosen with any of the CV, GCV, AIC_c or RCP criterion

2.4 Nonparametric Regression for Time Series

2.4.1 Introduction

Since most economic and financial data are time series, we discuss our methodologies and theory under the framework of time series. For linear models, the time series structure can often be assumed to have some well known forms such as an autoregressive integrated moving average (ARIMA) model. However, under nonparametric setting, this assumption might not be valid. Such structures can be modeled in various contexts. We concentrate on the following scenarios, for which there exist a large body of literature. The first scenario is a stationary sequence of random variables $\{(X_i, Y_i), i \geq 1\}, X \in \mathfrak{R}^d, Y \in \mathfrak{R}$ is observed. In econometrics, non-stationarity may be due to evolution of the economy, legislative changes, technological changes, political events and changes in climatic conditions among others. The observations may be dependent via the time index $t = 1,2, \ldots$ and it is desired to estimate a functional of the conditional distribution $L(Y|X)$ like the mean function or the median function $m(x)$. That is,

$$m(x) = E(Y|X = x) \qquad (2.4.0.1)$$

The second scenario is of a nonlinear autoregressive time series

$$Y_t = m(Y_{t-1}, \ldots, Y_{t-d}) + e_t, \quad t = 1,2, \ldots \qquad (2.4.0.2)$$

with independent innovation shocks $e_t = s_t . \xi_t$. One is interested in predicting new observations and in estimating the nonparametric autoregressive function m or the conditional variance function

$$V_t = s_t^2 = VaR(e_t|(Y_{t-1}, \dots Y_{t-d}) = x) \qquad (3.1.0.3)$$

The first scenario is typical for dynamic economic systems, which are modeled as multiple time series. The explanatory variable x may denote an exogenous variable and the function of interest is to predict the response variable Y for a given value of X. The second scenario is widely used in the analysis of financial and economic time series. In that context the variance function of the innovations is of great interest. In a parametric context the variance function is often estimated via the autoregressive conditional heteroscedasticity model family $(ARCH)$. Engle (1982) introduced this model class and Gouri´eroux (1992) gave an overview. Mathematically the second scenario can be mapped into the first one. Let us assume a more general time series dependence, which is commonly used in the literature, described as follows.

2.4.2 Stationarity

A process is said to be strictly stationary if the joint distribution of $X_1, X_2, \dots, X_k$ is the same as the joint distribution of $X_{t+1}, X_{t+2}, \dots, X_{t+k}$ evaluated at the same set of points $x_1, x_2, \dots, x_k$, i.e.

$$F_{X_1,X_2,\dots,X_k}(x_1, x_2, \dots, x_k) = F_{X_{t+1},X_{t+2},\dots,X_{t+k}}(x_1, x_2, \dots, x_k) \qquad (2.4.2.1)$$

for all t and for all k.

Let $\{X_t\}$ be a strictly stationary time series for $n \geq 1$. The stationary process is called strongly mixing if

$$\sup_{A \epsilon \mathcal{F}_1^n, B \epsilon \mathcal{F}_{n+k}^\infty} |P(A \cap B) - P(A)P(B)| \leq \alpha_k \qquad (2.4.2.2)$$

Where $\alpha_k \to 0$ and $\mathcal{F}_i^j$ is the α-field generated by $x_i, \dots, x_j$. The random variable X here may also stand for the pair (X, Y) so that the σ- fields are defined appropriately for the regression problem. Mixing dependence is commonly used to characterize the dependent structure and it is often referred to as short range dependence or weak dependence. This means that as the distance between two observations increases, the dependence becomes weaker and weaker very faster. The short term dependence does not have much effect on the local smoothing method since for any two given random variables X_i and X_j and a point x, the random variables $K_h(X_i - x)$ and $K_h(X_j - x)$ are nearly uncorrelated as $h \to 0$. With K being a kernel function assigning weights to each datum point. K is supported on $[-1,1]$ and determines the shape of the weights and satisfies the moment conditions; $\int_{-1}^{1} K(u)du = 1$, $\int_{-1}^{1} uK(u)du = 0$, $\int_{-1}^{1} u^2K(u)du \neq 0$ and $\int_{-1}^{1} K(u)^2du < \infty$. The parameter h is the smoothing parameter which determines the size of the weights.

It is well known that α-mixing includes many time series models as a special case. The key usage of mixing conditions is contained in the following lemma due to Volkonskii and Rozanov (1959);

Lemma 2.4 *Let $V_1, \dots, V_L$ be random variables measurable with respect to the σ-algebras $\mathcal{F}_{i1}^{ji}, \dots, \mathcal{F}_{iL}^{jl}$ respectively with $i_{l+1} - j_l \geq w \geq 1$ and $|V_j| \leq 1$ for $j = 1, \dots, L$. then*

$$\left| E\prod_{j=1}^{L} V_j - \prod_{j=1}^{L} E(V_j) \right| \leq 16(L-1)\alpha(w)$$

Lemma 2.4 shows that the dependent random variables can be approximated by a sequence of independent random variables having the same marginal distribution. This can be seen by taking $V_j = exp(it_j X_j)$. Thus this lemma becomes a statement about the characteristic function of the random variables.

2.4.3 Local polynomial fitting

Consider observations $(X_1, Y_1), \dots, (X_n, Y_n)$ that can be thought of as a realization from a stationary process. Of interest is to estimate $m(x)$ in (2.4.0.1) and its derivatives $m^{(j)}(x)$. $m(x)$, can be approximated by a weighted least squares regression problem. That is minimize

$$\sum_{i=1}^{n} \{Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j\}^2 K_h(X_i - x_0) \qquad (2.4.3.1)$$

Under certain mixing conditions for local polynomial estimators lemma (2.3) holds. Let $f(x)$ be the density of X_1 and $\sigma^2(x) = Var(Y_1|X_1 = x)$. Let S, S^* and cp denote some moment matrices and vector, then we have the following results by Fan and Gijbels (1996);

Theorem 2.3: *If* $h_n = O(n^{1/(2p+3)})$, *then as* $n \to \infty$,

$$\sqrt{n}h_n \left[diag(1, \dots, h_n^p)\{\hat{\beta}(x) - \beta(x)\} - \frac{h_n^{p+1} m^{(p+1)}(x)}{(p+1)!} S^{-1} cp \right] \xrightarrow{D} N\{0, \sigma^2(x) S^{-1} S^* S^{-1} / f(x)\}$$

(2.4.3.2)

at x, a continuity point, whenever $f(x) > 0$. An immediate consequence of Theorem (2.3) is that derivative estimator $\hat{m}_v(x)$ based on the local polynomial fitting is asymptotically normal;

$$\sqrt{nh_n^{2v+1}}\left\{\hat{m}_v(x) - m^{(v)}(x)\int t^{p+1}K_v^* dt \frac{v!m^{(p+1)}(x)}{(p+1)!} h_n^{p+1-v}\right\} \xrightarrow{D} N\left\{0, \frac{(v!)^2\sigma^2(x)\int K_v^{*2}(t)dt}{f(x)}\right\}$$

(2.4.3.3)

where K_v^* is the equivalent kernel. When $v = 0$, (2.4.3.3) gives the asymptotic normality of $\hat{m}(x)$

2.4.4 Asymptotic properties of nonparametric estimators for time series

We derive the asymptotic properties of the nonparametric estimator for the time series. Note that the mathematical derivations are different for the iid case and time series situations since $E[Y_t|X_1 \dots, X_n] \neq E[Y_t|X_t] = m(X_t)$, which is true for the identical and identically distributed case. We consider a simple case where $p = 1$ in (2.4.3.1) and the Nadaraya-Watson estimate. Then;

$$\hat{m}_{nw}(x) = \underbrace{\frac{1}{n}\sum_{t=1}^{n} m(X_t)K_h(X_t - x)/f_n(x)}_{I_1} + \underbrace{\sum_{t=1}^{n} W_t\varepsilon_t}_{I_2} \qquad (2.4.4.1)$$

I_1, contributes only to the bias and I_2 gives the asymptotic normality. First we derive the asymptotic bias for the boundary points. By Taylors' expansion, when X_t is in $(x - h, x + h)$, we have

$$m(X_t) = m(x) + m'(x)(X_t - x) + \frac{1}{2}m''(x_t)(X_t - x)^2$$

where $x_t = x + \theta(X_t - x)$ with $-1 < \theta < 1$. Then

$$I_1 = \frac{1}{n}\sum_{t=1}^{n} m(X_t)K_h(X_t - x).$$

If $p > 1$ (multivariate case), the asymptotic bias ($B_{nw}(x)$) becomes

$$B_{nw}(x) = \frac{h^2}{2} tr\left[\mu_2(K)\left\{m''(x) + 2f'(x)m'(x)^T / f(x)\right\}\right] \quad (2.4.4.2)$$

where $\mu_2(K) = \int uu^T K(u)du$. Under some regularity conditions it can be shown that for x being an interior grid point,

$$nh^p Var\,(I_2) \to V_0(K)\sigma_\varepsilon^2(x)/f(x) = \sigma_\varepsilon^2(x) \quad (2.4.4.3)$$

where; $\sigma_\varepsilon^2(x) = Var(\varepsilon_t | X_t = x)$. The asymptotic normality is

$$\sqrt{nh^p}\left[\hat{m}_{nw}(x) - m(x) - B_{nw}(x) + o_p(h^2)\right] \to N\{0, \sigma_m^2(x)\} \quad (2.4.4.4)$$

which as been proved by Cai (2002) as shown in appendix ($A4$).

2.5 Bias-corrected confidence bands

Consider a nonparametric regression model

$$y_t = m(x_t) + \epsilon_t, \quad t = 1,2, \ldots \quad (2.5.0.1)$$

where $\{\epsilon_t\}$ is a sequence of iid random variables with $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) = \sigma^2$; $\{(x_t, y_t)\}$ is a sequence of observations and for simplicity we assume that $x_t \in [0,1]$ with density function $f(x)$. Suppose that $\hat{m}(x)$ is an estimator of $m(x)$; then a $100(1 - \alpha)\%$ confidence bands is of the form

$$\lim_{n\to\infty}[Pr\{|\hat{m}(x) - m(x)| \le l_\alpha(x), for\ all\ x \in [0,1]\}] \ge 1 - \alpha \quad (2.5.0.2)$$

for some $l_\alpha(x)$. To find a solution to (2.5.02) we need the asymptotic distribution of $\sup_{0\le x\le 1}|\hat{m}(x) - m(x)|$. Secondly the supremum over [0,1] makes the distribution very sensitive to the estimator $\hat{m}(x)$. Since there is a bias in the estimator of $m(x)$ using Nadaraya-Watson estimation and local polynomial smoothing we introduce a bias correction term in the estimator when x_t is random designed and under dependence. The method can be generalized for fixed non uniform design.

We assume that $\{(x_t, y_t)\}$ is a strongly mixing sequence, which includes independent and identically distributed observations case and many time series models. Also under some conditions, autoregressive moving average models, autoregressive conditional heteroscedastic models and other nonlinear time series models are strongly mixing sequences with mixing coefficients of geometric rates. We use the local linear smoother with a correction to a bias term to provide confidence band which works well for randomly designed x_t.

We start by considering the simplest case, the Nadaraya-Watson estimator of $m(x)$ as discussed in section 2.2.2. Using condition (2.2.1) we have shown that

$$E\{\hat{m}_{NW,h}(x)\} = m(x) + m'(x)f'(x)\,h^2/f(x) + m''(x)\,h^2/2 + o(h^2) \qquad (2.5.0.3)$$

Uniformly for $x \in [0,1]$. When x_t is randomly designed on [0,1] then $f'(x) \equiv 0$ and the main part of the bias is $m''(x)\,h^2/2$. An effective way of removing this bias is by using the local linear smoother. According to Fan (1993), the local linear smoother of $m(x)$ is

$$\hat{m}_h(x) = \sum_{t=1}^{n} w_{t,h}(x) y_t \Big/ \sum_{t=1}^{n} w_{t,h}(x) \qquad (2.5.0.4)$$

where

$$w_{t,h}(x) = K\left(\frac{x_t - x}{h}\right)\left(s_{n,h,l} - \frac{x_t - x}{h} s_{n,h,l}\right)$$

with

$$s_{n,h,l} = \sum_{t=1}^{n} K\left(\frac{x_t - x}{h}\right)\left(\frac{x_t - x}{h}\right)^l, \quad l = 0,1, \ldots \qquad (2.5.0.5)$$

$\hat{m}_h(x)$ can then be expanded as

$$E\{\hat{m}_h(x)\} = m(x) + m''(x)\, h^2/2 + o(h^2) \qquad (2.5.0.6)$$

To obtain an estimator of $m''(x)/2$ denoted by $m_2(x)$ we use local third order polynomial fitting of Fan and Gijbels (1996), whereby to estimate $m^{(v)}(\cdot)$ on the interval $[a, b]$ using the p^{th} order polynomial fit, the most appropriate and simple choice is $p = v + 1$. Thus the estimator of $m_2(x)$ is the third element of

$$\left(\hat{m}_{0,b}(x), \hat{m}_{1,b}(x), \hat{m}_{2,b}(x), \hat{m}_{3,b}(x)\right)^T = S_{xx}^{-1} S_{xy} \qquad (2.5.0.7)$$

where $S_{xx} = \left(s_{n,b,i+j}\right)_{0 \le i,j \le 3}, s_{n,b,i+j}$ is defined as in (2.5.0.5) with h and $K(\cdot)$ replaced by b and $K_0(\cdot)$, another bandwidth and kernel function;

$$S_{xy} = (a y_t, a(x_t - x) b^{-1} y_t, a(x_t - x)^2 b^{-2} y_t, a(x_t - x)^3 b^{-3} y_t)^T$$

where

$$a = \sum_{t=1}^{n} K_0\{(x_t - x) b^{-1}\}$$

We shall then construct the confidence band for $m(x)$ in the form

$$\{m(x): |\hat{m}_h(x) - \hat{m}_{2,b}(x)h^2 - m(x)| \le l_\alpha(x), for\ all\ x \in [0,1]\} \quad (2.5.0.8)$$

For the $k\ th$ order polynomial fitting the bias term is $m^{(k+1)}(x)h^{k+1}/(k+1)!$ and the optimal bandwidth is of order $O\left(n^{-1/(2k+3)}\right)$ as shown by Fan and Gijbels (1996). Most of the data driven band widths also have this rate. However for the non-corrected confidence bands, we can only allow the bandwidth to achieve a rate of $O\left(n^{-\delta}\right), 1/(2k+3) < \delta < 1/3$ as shown by Härdle (1989). The left hand limit for δ is due to the bias term

$$m^{(k+1)}(x)h^{k+1}/(k+1)!$$

The bias term needs to be estimated separately to use the data driven bandwidth. This is why we use the local linear smoother and local third order polynomial fitting to estimate $m(x)$ and $m_2(x)$ respectively.

2.6 Modeling dynamics using nonparametric methods

Modeling the dynamics of macroeconomic factors is one of the most important aspects of short rate movement description. The underlying process of interest $\{Y_t, t \ge 0\}$ is often modeled as a time homogenous diffusion (volatility) process or stochastic differential equation

$$Y_t = \mu(Y_{t-1})dt + \sigma(Y_{t-1})dw_t \quad (2.6.0.1)$$

where Y_t is a stationary transformed series, and the smooth functions:

$\mu(\cdot)$; is the percentage drift of Y_{t-1} which is the instantaneous mean and

$\sigma(\cdot)$; is the percentage volatility or the instantaneous variance.

$\mu(\cdot)$ and $\sigma(\cdot)$ are constants which determine the dynamics of the model.

w_t, is a stochastic process where, $\{w_t\}_{t\geq 0}$ satisfies the following conditions:

Condition 2.2.2

i. *For each $s \geq 0$ and $t > 0$ the random variable $w_{t+s} - w_s$ has the normal distribution with mean zero and variance σ^2.*

ii. *For each $n \geq 1$ and any times $0 \leq t_0 \leq t_1 \leq \cdots \leq t_n$, the random variables $\{w_{t_r} - w_{t_{r-1}}\}$ are independent.*

iii. $w_0 = 0$

iv. *w_t is continuous in $t \geq 0$*

There are two basic approaches of identifying $\mu(\cdot)$ and $\sigma(\cdot)$. The first is the parametric approach which assumes some parametric forms of $\mu(\cdot,\beta)$ and $\sigma(\cdot,\beta)$ and estimates the unknown model parameters, say β. The second approach is nonparametric which does not assume any restrictive functional form for $\mu(\cdot)$ and $\sigma(\cdot)$ beyond Condition 2.2.2. Since the time series sequence is observed at equally spaced time points, we use the infinitesimal generator, the first order approximation of moments of Y_{t-1}, a discretized version of the Ito's process. The drift and diffusion are respectively the first two moments of the infinitesimal conditional distribution of Y_t:

$$\left.\begin{aligned} \mu(Y_{t-1}) &= \lim_{\Delta\to 0} \Delta^{-1} E(Y_t | Y_{t-1}) \\ \sigma^2(Y_{t-1}) &= \lim_{\Delta\to 0} \Delta^{-1} E(Y_t^2 | Y_{t-1}) \end{aligned}\right\} \qquad (2.6.0.2)$$

The drift describes the movement of Y_{t-1} due to time changes, while the diffusion term measures the magnitude of random fluctuations around the drift.

According to Bandi and Nguyen (2000), the approximations of the drift and diffusion of any order display the same rate of convergence and limiting variance so that the asymptotic argument in conjunction with computational issues suggests simply using the first order approximation of Santon (1987) that is

$$\mu(Y_t)^{(1)} = \frac{1}{\Delta}E\{Y_{t+\Delta} - Y_t|Y_t\} = O(\Delta)$$

Suppose we observe Y_{t-1} at $t = \tau\Delta, \tau = 1, \dots, n$ in a fixed time interval $[0, T]$ Denoting the random sample as $\{Y_{(t-1)\tau\Delta}\}_{\tau=1}^{n}$ then it follows from (2.6.0.2) that the first order approximations of $\mu(Y_{t-1})$ and $\sigma(Y_{t-1})$ lead to

$$\mu(Y_{t-1}) \approx \frac{1}{\Delta}E\left(Y_\tau \middle| Y_{(t-1)\tau\Delta} = y_{t-1}\right) \ and$$

$$\sigma^2(Y_{t-1}) \approx \frac{1}{\Delta}E\left(Y_\tau^2 \middle| Y_{(t-1)\tau\Delta} = y_{t-1}\right) \quad (2.6.0.3)$$

for all $1 \leq \tau \leq n-1$, where $Y_\tau = Y_{(t-1)(\tau+1)\Delta} - Y_{(t-1)\tau\Delta}$. $\mu(Y_{t-1})$ and $\sigma^2(Y_{t-1})$ becomes classical nonparametric regression problem and a nonparametric smoothing approach can be applied to estimate them. The WNW estimators for $\mu(Y_{t-1})$ and $\sigma^2(Y_{t-1})$ are given for any grid point y_{t-1}, respectively by

$$\hat{\mu}(y_{t-1}) = \frac{1}{\Delta}\frac{\sum_{\tau=1}^{n-1} Y_\tau K_h\left(y_{t-1} - Y_{(t-1)\tau\Delta}\right)w_t(y_{t-1})}{\sum_{\tau=1}^{n-1} K_h\left(y_{t-1} - Y_{(t-1)\tau\Delta}\right)w_t(y_{t-1})}$$

$$and \qquad \hat{\sigma}(x) = \frac{1}{\Delta}\frac{\sum_{\tau=1}^{n-1} Y_\tau^2 K_h\left(y_{t-1} - Y_{(t-1)\tau\Delta}\right)w_t(y_{t-1})}{\sum_{\tau=1}^{n-1} K_h\left(y_{t-1} - Y_{(t-1)\tau\Delta}\right)w_t(y_{t-1})} \quad (2.6.0.4)$$

Since

$$\partial Y_t = \mu(Y_{t-1})\partial + \sigma(Y_{t-1})\epsilon\sqrt{\partial}$$

where $\partial Y_t = Y_{t-1+\partial} - Y_{t-1}$, $\epsilon \sim N(0,1)$, Y_{t-1} and ϵ_t are independent, therefore since the drift is of order dt and diffusion has a lower order $\sqrt{dt}$ as $(dw_t)^2 = dt + O((dt)^2)$ for infinitesimal changes in time, and the local time dynamics of the sampling paths reflects more of the diffusion than those of the drift term. Therefore where Δ is very small, it becomes much easier for the identification of the diffusion term than the drift term. Therefore using (2.6.0.4) to estimate $\sigma^2(y_{t-1})$ we observe that the drift

$$\mu(Y_{t-1}) = \frac{1}{2\pi(Y_{t-1})}\frac{\partial[\sigma^2(Y_{t-1})\pi(Y_{t-1})]}{\partial Y_{t-1}}$$

where $\pi(Y_{t-1})$ is the stationary density of $\{Y_{t-1}\}$.

Therefore

$$\hat{\mu}(y_{t-1}) = \frac{1}{2\hat{\pi}(y_{t-1})}\frac{\partial\{\hat{\sigma}^2(y_{t-1})\hat{\pi}(y_{t-1})\}}{\partial y}$$

If the process Y_t in (2.6.0.1) represents a rate of change then we have

$$Y(t + \Delta t) - Y(t) = \mu\big(t, Y(t)\big)\Delta t + \sigma\big(t, Y(t)\big)\Delta w(t)$$

where, $\Delta w(t) = w(t + \Delta t) - w(t)$. To make the process precise, by letting $\Delta t \to dt$ we obtain the following stochastic differential equation

$$dY(t) = \mu\big(t, Y(t)\big)dt + \sigma\big(t, Y(t)\big)dw(t) \qquad (2.6.0.5)$$

With the initial condition $Y(0) = a$, (2.6.0.5) can be expressed as

$$Y(t) = a + \int_0^t \mu\big(s, Y(s)\big)dS + \int_0^t \sigma\big(s, Y(s)\big)dw(s)$$

In order to guarantee the existence of the stochastic integral $\int_0^t g(s)dw(s)$ we have to impose some kind of integratebility conditions on g and the class of $\mathcal{L}^2$ turns out to be natural.

Condition 2.2.3

1. *The process g belongs to the class $\mathcal{L}^2[a,b]$ if the following conditions are satisfied.*

   (A1) $\int_a^b \mathbb{E}[g^2(s)]ds < \infty$

   (A2) The process g is adapted to the $\mathcal{F}_t^w$- filtration.

2. *The process g belongs to the class $\mathcal{L}^2[0,t]$ for all $t > 0$.*

Assume $g(t) \in \mathcal{L}^2[a,b]$, assume also that $g(t)$ is a simple function hence we can define a partition $a = t_0 < t_1 \ldots < T_n = b$ such that g is constant in the time intervals $t_i - t_{i-1}$. We can then define the stochastic integral

$$\int_a^b g(s)dw(s) = \sum_{i=0}^{n-1} g(t_i)[w(t_{i+1}) - w(t_i)] \qquad (2.6.0.6)$$

where we have evaluated $g(s)$ at $s = t_i$ over the time interval $[t_i, t_{i+1})$. Note that the time interval includes the point t_i and excludes t_{i+1}.

If $f(S,t)$ is a deterministic function of S and time t we approximate the change in f due to a change in both S and t as

$$df(S,t) = \frac{\delta f}{\delta t}dt + \frac{\delta f}{\delta S}dS \qquad (2.6.0.7)$$

Assume f only depends on S, then

$$\Delta f \equiv f(S + \Delta S) - f(S)$$

$$= f'(S)\Delta S + \frac{1}{2!}f''(S)(\Delta S)^2 + \frac{1}{3!}f'''(S)(\Delta S)^3 + \cdots$$

If S is stochastic and $dS = dw$. Then

$$\Delta f(S) = f'(S)\Delta S + \frac{1}{2!}f''(S)(\Delta S)^2 + \frac{1}{3!}f'''(S)(\Delta S)^3 + \cdots$$

$$= f'(S)\Delta w + \frac{1}{2!}f''(S)(\Delta w)^2 + \frac{1}{3!}f'''(S)(\Delta w)^3 + \cdots$$

$$= f'(S)\emptyset\Delta t^{1/2} + \frac{1}{2!}f''(S)\emptyset^2\Delta t^{2/2} + \frac{1}{3!}f'''(S)\emptyset^3\Delta t^{3/2} + \cdots \quad (2.6.0.8)$$

where $\Delta w = \emptyset\Delta t^{1/2}, \emptyset \sim N(0,1)$.

Assume that the process S satisfies the following stochastic differential equation

$$dS = \mu(S,t)dt + \sigma(S,t)dw(t) \quad (2.6.0.9)$$

where $\mu(S,t)$ and $\sigma(S,t)$ are adapted processes. Let f be a twice continuous differentiable function then

$$df(S,t) = \left(\frac{\partial f}{\partial t} + \mu(S,t)\frac{\partial f}{\partial S} + \frac{1}{2}\sigma^2(S,t)\frac{\partial^2 f}{\partial S^2}\right)dt + \sigma(S,t)\frac{\partial f}{\partial S}dw$$

In modeling rates of change

$$\frac{dS}{S} = \mu dt + \sigma dw(t) \quad (2.6.1.0)$$

Naïve solution of the stochastic differential equation would be:

$$\int_0^t \frac{dS}{S} = \int_0^t \mu ds + \sigma\int_0^t dw(s)$$

$$\ln\frac{S(t)}{S(0)} = \mu t + \sigma w(t) \quad .$$

Hence

$$S(t) = S(0)e^{\mu t+\sigma w(t)} \qquad (2.6.1.1)$$

Let f = ln$S$, rewriting (2.6.1.0) as $ds = S\mu dt + S\sigma dw(t)$ with $\mu(S,t) = \mu S$ and $\sigma(S,t) = \sigma S$, we can show that this leads to the naïve solution (2.6.1.1).

Let

$$d(\ln S) = \left(S\mu\frac{1}{S} - \frac{1}{2}S^2\sigma^2\frac{1}{S^2}\right)dt + S\sigma\frac{1}{S}dw$$

$$= \left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma dw$$

Hence by integrating both sides we obtain

$$S(t) = S(0)e^{\left(\mu-\frac{1}{2}\sigma^2\right)t+\sigma w(t)}$$

which is the naïve solution with an extra term $\frac{1}{2}\sigma^2 t$.

In the application, let the integrand be a Wiener process representing how much stock an institution holds and the integrator be the movement of the prices. The Itos' stochastic integral represents the pay off of a continuous time trading strategy consisting of holding an amount H_t of the stocks at time t, which is how much money an institution has in total including what out stock is worth at any given movement. We chose a sequential partitions of the interval from 0 to t then we construct the Riemann sums. The limit when computing the Riemann sum is taken in probability as the mesh of the partition goes to zero. The main insight is that the integral can be defined as long as the integrand is adapted, which loosely means that its value at time t can only depend on the information available up until this time. This prevents the possibility of unlimited gains through high frequency trading.

2.7 Models with exogenous variables

Most of the characteristics outlined in the previous section have been univariate, relating to the series of only information contained in that series history. But it is believed that demand and financial asset prices may contain relevant information of the market around them. It is possible that deterministic events also have an impact. They include among others company announcements, macroeconomic announcements and even deterministic time-of-day effects which may all have an influence on the process. Andersen *et al.* (2003), found that the volatility of the deutche mark-dollar exchange rate increases markedly around the time of the announcement of U.S. macroeconomic data, such as the employment report, the Producer Price Index or the quarterly Gross Domestic Product. Glosten *et al.* (1993) found that indicator variables for October and January assist in explaining some of the dynamics of the conditional volatility of equity returns.

Let $\{Y_t, X_t, Z_t\}_{t=-\infty}^{\infty}$ be jointly stationary process, where X_t and Z_t take values in $\Re^p$ and $\Re^l$ with $p, l \geq 0$ respectively. The regression surface is defined by

$$m(x, z) = E\{Y_t | X_t = x, Z_t = z\} \qquad (2.7.0.1)$$

Here, Y_t is the dependent variable measurable on the real line and it is assumed that $E|Y_t| < \infty$. $X_t \in \Re^p$, $(p \geq 1)$, is a vector of possibly endogenous explanatory variables and $Z \in \Re^l$, $(l \geq 1)$, is a vector of exogenous explanatory variables. If $l = 0$, then Z is not in the model. The regression function $m(\cdot,\cdot)$ defined in (2.7.0.1) can be decomposed as the sum

$$m(x, z) = \mu + g_1(x) + g_2(z) \qquad (2.7.0.2)$$

If we assume that $E\{g_1(X_t)\} = 0$ and $E\{g_2(Z_t)\} = 0$, then the projection of $m(x, z)$ on the $g_1(x)$ direction is defined by

$$E\{m(x, Z_t)\} = \mu + g_1(x) + E\{g_2(Z_t)\} = \mu + g_1(x) \qquad (2.7.0.3)$$

To estimate the unknown components in (2.7.0.2), the first initial estimated values of all components will be obtained. Local linear method will be used to estimate directly the high dimension regression surface and then these averages of regression surfaces over the rest of the variables is obtained to stabilize the variance. At the second stage the local polynomial technique is used to estimate any additional components by using the initial estimated values of the rest of the components.

To get the estimate of $g_2(\cdot)$, we use a small bandwidth h_s so that the bias can be asymptotically negligible. Let the additional components have continuous second partial derivatives so that $m(u, v)$ can be locally approximated by a linear term in a neighbourhood of (x, z) namely, $m(u, v) \approx \beta_0 + \beta_1^T(u - x) + \beta_2^T(v - z)$, with $\{\beta_j\}$ depending on x and z, where β_j^T denotes the transpose of β_j. Let $K_0(\cdot)$ and $K_1(\cdot)$ be symmetrical kernel functions in $\Re^p$ and $\Re^l$ and $h_0(n) > 0$ and $h_s = h_s(n) > 0$ be bandwidths in the step of estimating the regression surface. Let β_j be the minimiser of the following locally weighted least squares

$$\sum_{t=1}^{n}\{Y_t - \beta_0 - \beta_1^T(X_t - x) - \beta_2^T(Z_t - z)\}^2 K_{h_0}(X_t - x)K_{h_1}(Z_t - z)$$

where $K_{h_0}(\cdot) = K(\cdot\,|h)/h^p$ and $K_{h_1}(\cdot) = K(\cdot\,|h)/h^l$ then the local linear estimator of the regression surface $m(x,z)$ is $\widehat{m}(x,z) = \hat{\beta}_0$. Using (2.7.0.3) and computing the sample average of $\widehat{m}(\cdot,\cdot)$ the estimators of $g_1(\cdot)$ and $g_2(\cdot)$ become

$$\hat{g}_1(x) = \frac{1}{n}\sum_{t=1}^{n}\widehat{m}(x, Z_t) - \hat{\mu} \quad and \;\; \hat{g}_2(z) = \frac{1}{n}\sum_{t=1}^{n}\widehat{m}(X_t, z) - \hat{\mu}$$

where $\hat{\mu} = n^{-1}\sum_{t=1}^{n} Y_t$. Then using the partial residuals $Y_t^* = Y_t - \hat{\mu} - \hat{g}_2(Z_t)$ we apply the local linear regression technique to the regression model $Y_t^* = g_1(X_t) + \varepsilon_t^*$. To estimate $g_1(\cdot)$ we solve the weighted least squares problem

$$\sum_{t=1}^{n}\{Y_t^* - \beta_1 - \beta_2^T(X_t - x)\}^2 J_{h_2}(X_t - x) \qquad (2.7.0.4)$$

where $J(\cdot)$ is the kernel function in $\Re^p$ and $h_2 = h_2(n) > 0$ is the bandwidth at the second stage. Minimizing (2.7.0.4) with respect to β_1 and β_2 gives the estimate of $g_1(x)$ denoted by $\widetilde{g}_1(x) = \hat{\beta}_1$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the minimiser of (2.7.0.4).

2.7.1 Sampling properties of models with exogenous variables

To establish the asymptotic normality it is assumed that the initial estimator satisfies a linear approximation;

$$\hat{g}_2(Z_t) - g_2(Z_t) \approx \frac{1}{n}\sum_{i=1}^{n} K_{1s}(Z_i - Z_t)\tau(X_i, Z_t)\delta_i + \frac{1}{2}h_s^2 tr\{\mu_2(K_1)g_2''(Z_t)\} \quad (2.7.1.1)$$

where $\delta_t = Y_t - m(X_t, Z_t)$ and $\tau(x, z) = p_1(x)/p(x, z)$

Theorem 2.4 *If bandwidth* h_1 *and* h_2 *are chosen such that* $h_1 \to 0$, $nh_s^{lq} \to \infty$, $h_2 \to 0$ *and* $nh_2^p \to \infty$ *as* $n \to \infty$ *then according to Masry and Fan (1997):*

$$\sqrt{nh_2^p}\left[\tilde{g}_1(x) - g_1(x) - \kappa - o_p(h_1^2 + h_2^2)\right] \xrightarrow{D} N\{0, v_0(x)\}$$

where κ is the asymptotic bias given by

$$\kappa = bias(x) = \frac{h_2^2}{2} tr\{\mu_2(J)g_1''(x)\} - \frac{h_1^2}{2} tr\left(\mu_2(K_1)E\left(g_2''(Z_t)|X_t = x\right)\right)$$

and the asymptotic variance is $v_0(x) = v_0(J)B(x)$. Here

$$B(x) = p_1^{-1}(x)E\{\sigma^2(X_t, Z_t)|X_t = x\} \quad (2.7.1.2)$$

If X_t and Z_t are correlated when $\sigma^2(x, z)$ is a constant then it follows from Cauchy Schwarz inequality that

$$B(x) = \frac{\sigma^2}{p_1(x)}\int p^{1/2}(z|x)\frac{p_2(z)}{p^{1/2}(z|x)}dz \le \frac{\sigma^2}{p_1(x)}\int \frac{p_2^2(z)}{p(z|x)}dz = A(x)$$

$p(x, z)$, denotes the joint density of X_t and Z_t, $p_1(x)$ denotes the marginal density of X_t, $p_2(z)$ is the marginal density of Z_t, $v_0(K) = \int K^2(u)du$ and $\mu_2(K) = \int uu^T K(u)du$.

Masry and Fan (1997) showed that the optimal estimate of $g_1(x)$ denoted by $g_1^*(x)$ is asymptotically normally distributed. This is by using (2.7.0.4) in which the partial residuals Y_t^* is replaced by the partial error $\tilde{Y}_t = Z_t - \mu - g_2(Z_t)$. Thus

$$\sqrt{nh_2^p}\left[g_1^*(x) - g_1(x) - \frac{h_2^2}{2} tr\{\mu_2(J) g_1^{''}(x)\} + o_p(h_2^2)\right] \xrightarrow{D} N\{0, v_0(x)\}$$

This is in agreement with Theorem 2.4, which shows that the estimator share the same asymptotic bias and variance if $h_1 = o(h_2)$.

2.7.2 Consistency

This section gives conditions under which $E\|\hat{g} - g\|^2 \to 0$ as $n \to \infty$ in model (2.7.0.2).

Let

$$K_{p,h}(v) = \prod_{j=1}^{l} K_h\left(v^j \middle| h\right)$$

where K denote a continuously differentiable kernel function whose support is $[-1,1]$ and is symmetrical about 0, $h > 0$ denote a bandwidth parameter. Define, $K_{p,h}(v) = K(v|p, h)$ for any $v \in [-1,1]$. Likewise define $K_{l,h}(v)$ for a l –vector. Let f_{XZ} and f_{YXZ}, respectively denote the probability density functions of (X, Z) and (Y, X, Z). Define

$$F_{YXZ}(y, x, z) = \int_{-\infty}^{y} f_{YXZ}(v, x, z) dv$$

Let the kernel estimators of f_{XZ}, f_{YXZ} and F_{YXZ} be as follows;

$$\hat{f}_{XZ}(x, z) = \frac{1}{nh^{p+l}} \sum_{i=1}^{n} K_{p,h}(x - X_i) K_{l,h}(z - Z_i)$$

$$\hat{f}_{YXZ}(v,x,z) = \frac{1}{nh^{p+l+1}}\sum_{i=1}^{n} K_h(y-Y_i)K_{p,h}(x-X_i)K_{l,h}(z-Z_i)$$

and

$$\hat{F}_{YXZ}(y,x,z) = \int_{-\infty}^{y} \hat{f}_{YXZ}(v,x,z)dv$$

For each $z \in [0,1]^l$, define the operators Γ_z and $\hat{\Gamma}_z$ on $L_2[0,1]^p$ by

$$(\Gamma_z\varphi_z)(x) = \int_{[0,1]^p} \hat{F}_{YXZ}[\varphi_z(x),x,z]dx$$

where φ_z is any function on $L_2[0,1]^p$. The function g satisfies

$$(\Gamma_z g)(x,z) = q\, f_{XZ}(x,z) \qquad (2.7.2.1)$$

where q is an unknown constant satisfying $0 < q < 1$. The function $g(\cdot,z)$ is defined if (2.7.2.1) is unique solution up to a set of x values of Lebesgue measure 0 for the specified z. Define $\Omega = \{\varphi \in L_2[0,1]^p : \|\varphi\|^2 \le M\}$ for some constant $M < \infty$. For each $z \in [0,1]^l$, the estimator of $g(\cdot,z)$ is any solution to the problem

$$\hat{g}(\cdot,z) = \underset{\varphi_z\in\Omega}{\arg\min}\left\{\int_{[0,1]^p}\left[\left(\hat{\Gamma}_z\varphi_z\right)(x) - q\hat{f}_{XZ}(x,z)\right]^2dx + a_n\int_{[0,1]^p}\varphi_z(x)^2dx\right\}$$

We define $\delta_n = h^{2r} + (nh^{p+l})^{-1}$, and we make the following assumptions

Assumption 2.2.6

(A1) *(a) The function g is identified*

(b) $\int_{[0,1]^p} g(x,z)^2dx \le M$ for each $z \in [0,1]^l$ and some constant $M < \infty$

(A2) f_{YXZ} *has* $r > 0$ *continuous derivatives with respect to any combination of its arguments. These derivatives are bounded in absolute value by* M.

(A3) *As* $n \to \infty$, $a_n \to 0$, $\delta_n \to 0$ *and* $\delta_n / a_n \to 0$

(A4) *The kernel function* K *is supported on* $[-1,1]$ *and is continuously differentiable and*

Symmetrical about 0

This leads to the following theorem proved by Masry and Fan (1997):

Theorem 2.5 *Let assumption 2.2.6 (A1-A4) hold, then for each* $z \in [0,1]^l$,

$$\lim_{n\to\infty} E\left(\int_{[0,1]^p} [\hat{g}(x,z) - g(x,z)]^2 dx \right) = 0$$

2.7.3 Rate of Convergence

As when X and Z are scalars, the rate of convergence of $\hat{g}$ in the multivariate model depends on the rate of convergence of the singular values of the Fréchet derivative of Γ_z. Accordingly, let Γ_{gz} denote the Fréchet derivative of T_z at g and let, T_{gz}^* denote the adjoint of T_{gz}. T_{gz} and T_{gz}^*, respectively are the operators defined by

$$(T_{gz}\varphi_z)(x) = \int_{[0,1]^p} f_{YXZ}[g(x,z),x,z]\varphi_z(x)dx \qquad (2.7.3.1)$$

and

$$\left(T^{*}{}_{gz}\varphi_z\right)(x) = \int\limits_{[0,1]^p} f_{YXZ}[g(x,z)x,z]\varphi_z(x)dx \qquad (2.7.3.2)$$

Assume that for each $z \in [0,1]^p$, $T_{gz}^{*}\ T_{gz}$ is non-singular. Let $\{(\lambda_{z,j}, \varphi_{zj}): j = 1,2,\dots\}$ denote the Eigen values and Eigen vectors of $T_{gz}^{*}\ T_{gz}$ ordered so that $\lambda_{z1} \geq \lambda_{z2} \geq \cdots > 0$. The Eigen vectors $\{\varphi_{zj}\}$ form a complete, orthonormal basis for $L_2[0,1]^p$. Thus for each $z \in L_2[0,1]^p$, we can write

$$g(x,z) = \sum_{j=1}^{\infty} b_{zj}\varphi_{zj}(x) \qquad (2.7.3.3)$$

where

$$b_{zj} = \int\limits_{[0,1]^p} g(x,z)\varphi_z(x)dx$$

Assumption 2.2.7

(A5) *(a) There are constants* $\alpha > 1$, $\beta > 1$ *and* $C_0 < \infty$, *such that*

$$\beta - 1/2 \leq \alpha < 2\beta - 1,$$

$$j^{-\alpha} \leq C_0\lambda_{zj}, \text{ and } |b_{zj}| \leq C_0 j^{-\beta} \text{ uniformly in } z \in [0,1]^l \text{ for all } j \geq 1$$

(b) Moreover, $r \geq \max\{[m(2\beta + \alpha - 1) - l]/2, r^*\}$, *where* r^* *is the largest root of the equation*

$$4[(\alpha + 1)/(2\beta + \alpha)]r^2 - 2[(p + l)(2\beta - 1)/(2\beta + \alpha) + p + 1 - l]r - (p + 1)l = 0$$

(A6) *There is a finite constant* $L > 0$ *such that*

$$\|\Gamma_z(g_1) - \Gamma_z(g_2) - \Gamma g_2 z(g_1 - g_2)\| \leq 0.5L\|g_1 - g_2\|^2$$

for any $g_1, g_2 \in L_2[0,1]$ *uniformly in* $z \in [0,1]^l$ *and*

$$\sum_{j=1}^{\infty} \frac{b_{zj}^2}{\lambda_{zj}} < \frac{1}{L}$$

uniformly in $z \in [0,1]^l$

(A7) *The tuning parameters* h *and* a_n *satisfy* $h = C_h n^{-1/(2r+p+l)}$ *and* $a_n = C_a n^{-\tau\alpha/(2\beta+\alpha)}$*, where* $\tau = 2r/(2r + l)$ *and* C_h *and* C_a *are positive finite constraints.*

Let $\mathcal{H}_M = \mathcal{H}_M(M, C_0, \alpha, \beta, L, r, p, l)$ *be the set of distributions of* (Y, X, Z) *that satisfy Assumption 2.2.6 (A1 and A2), Assumption 2.2.7 (A5 and A6) with fixed values of* $M, C_0, \alpha, \beta, L, r, p$ *and* l

Theorem 2.6 *Let Assumptions 2.2.6 (A1, A2, A4), and Assumption (2.2.7) hold. Then for each* $z \in [0,1]^m$

$$\lim_{D\to\infty} \limsup_{n\to\infty} \sup_{H\in\mathcal{H}_m} \boldsymbol{P}_H \left\{ \int_{[0,1]^p} [\hat{g}(x,z) - g(x,z)]^2 dx > D_n^{-\tau(2\beta-1)/2\beta+\alpha} \right\} = 0 \quad (2.7.3.4)$$

If in addition

$$[p(2\beta + \alpha - 1) - l]/2 \geq r^* \quad (2.7.3.5)$$

then, for each $z \in [0,1]^m$

$$\liminf_{n\to\infty} \sup_{H\in\mathcal{H}_M} \boldsymbol{P}_H \left\{ \int_{[0,1]^l} [\hat{g}(x,z) - g(x,z)]^2 dx > D_n^{-\tau(2\beta-1)/(2\beta+\alpha)} \right\} > 0 \quad (2.7.3.6)$$

If $l = 0$, then (2.7.3.5) simplify to $\alpha \geq 1 + \frac{2}{p}r^* - 2\beta$. The rate of convergence in Theorem 2.5 is the same as that in Theorem 2.6. If $p = 1$ and $l = 0$, Theorem 2.5 shows that increasing l decreases the rate of convergence of $\hat{g}$ for any fixed r. Assumption 2.2.7 *(A 5)* implies that as l increases, r must also increase to maintain the rate of convergence $n^{-\tau(2\beta-1)/(2\beta+\alpha)}$. This is a form of the curse of dimensionality that is associated with the endogenous explanatory variable X.

CHAPTER THREE

3.0 REGRESSION QUANTILES

3.1 Introduction to regression quantiles

When data means and variance are non-constant, typically skewed or contains some outliers, it is understood that the observations come from different distributions over time. This creates difficulties in empirical modelling. Standard asymptotic distribution theory often does not apply to regression involving variables of this nature since inferences may be misleading. In such data, median regression a special case of quantile regression is more robust than the mean regression. More of interest is the case where the data pattern shows heteroscedasticity and asymmetries. Estimation of linear models for conditional quantiles function has been considered by Koenker and Bassett (1978). That is, for any $\theta - th$ quantile of a scalar random variable, Y can be viewed as a solution to the problem

$$argmin_{a\in\Re}E\{\rho_\theta(Y_t - a)|X_t = x\} = q_\theta(x) \qquad (3.1.0.1)$$

where $\rho_\theta(\cdot)$ is the "check" function, and $\rho_\theta(u) = \theta|u|^+ + (1-\theta)|u|^-$.

For fixed X, a nonparametric estimator is defined by setting the value $\hat{a}$ in (3.1.0.1) that minimize

$$\sum_{i=1}^{n} \rho_\theta (Y_i - a)^2 K\left(\frac{x - X_i}{h}\right) \qquad (3.1.0.2)$$

where $K(\cdot)$ is the kernel function and h is the bandwidth. $\rho_\theta(\cdot)$ is the loss function given by

$$\rho_\theta(u) = \theta I_{\{u \geq 0\}}(u).u + (\theta - 1) I_{\{u<0\}}(u).u$$

Similarly, in a regression setting we might hypothesize a linear relationship between the conditional quantiles of Y and a vector of covariates $x \in \Re^p$, that is

$$F_Y^{-1}(\theta|x) = x'\beta$$

we may define the θth regression quantiles of the sample $\{(y_i, x_i)\}_{i=1}^{n}$ as solution to

$$\underset{b \in \Re^p}{\text{argmin}} \sum_{i=1}^{n} \rho_\theta (y_i - x_i b)$$

Since our motivation is a convenient method of detecting conditional heteroscedasticity, we assume that $\{Y_t, X_t\}_{t=-\infty}^{\infty}$ is a stationary sequence as described in Chapter 2. Denote $F(Y|X)$ the conditional distribution of Y given $X = x$ and $X_t = (X_{t1}, \dots, X_{td})'$ be the associated covariate vector in $\Re^d$ with $d \geq 1$ and might be a function of exogenous (covariate) variables or some lagged (endogenous) variables or a function of time t. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the order statistics of $\{X_t\}_{t=1}^{n}$. Define the inverse of $F(x)$ as

$$F^{-1}(\theta) = inf\{x \epsilon \Re; F(x) \geq \theta\} \qquad (3.1.0.3)$$

where $\Re$ is the real line, inf denotes the smallest real number x, satisfying $F(x) \geq \theta$. So we define the quantile estimate as

$$\hat{q}_\theta(x) = inf\{y \in \Re: F(y|x) \geq \theta\} = \hat{F}^{-1}(\theta|x) \qquad (3.1.0.4)$$

We remark that $\hat{q}_\theta(x)$ always exists because $\hat{F}(y|x)$ is between zero and one and monotone to y and involves only one bandwidth, so it makes practical implementation more appealing.

3.1.1 The asymptotic properties of the conditional quantile estimator

Define the probability distribution function of X_t at x_i as g, estimated by

$$\hat{g}(x_i) = \left(\frac{1}{nh}\right)\sum_{t=1}^{n} K_h(x_i - X_t) \qquad (3.1.1.1)$$

As pointed out in Parzen (1962) and Mwita (2003), the joint probability density function $f(y,x)$ of (Y_t, X_t) at (y,x) will be estimated by

$$\hat{f}(y,x) = \frac{1}{nh}\sum_{t=1}^{n} \varphi\left(\frac{y - Y_t}{h^*}\right) K_h(x - X_t) \qquad (3.1.1.2)$$

Where the function φ is the kernel and h^* is the bandwidth for Y_t at point y. The conditional probability density function $f_x(y)$ of Y_t given that $X_t = x$ will be estimated by

$$\hat{f}_x(y) = \frac{\hat{f}(y,x)}{\hat{g}(x)} \qquad (3.1.1.3)$$

The conditional cumulative density function $F_x(y)$ of Y_t given $X_t = x$ at distinct points y can be obtained by the Collomb (1980) empirical estimator,

$$\hat{F}_x(y) = \frac{\sum_{t=1}^{n} K_h(x - X_t) I_{\{Y_t \leq y\}}}{\sum_{t=1}^{n} K_h(x - X_t)}$$

$$= \frac{\hat{r}_x(y)}{\hat{g}(x)} \tag{3.1.1.4}$$

The conditional estimator of $q_\theta(x)$ is then obtained by inverting (3.1.1.4) at θ to obtain (3.1.0.4). Because $0 \leq F_x(y) \leq 1$ and is strictly monotone in y, $q_\theta(x)$ exist and is unique. To show that our estimator $\hat{q}_\theta(x)$ maintains the aforementioned advantages of $\widehat{F}(\cdot|x)$, we introduce the following additional conditions, to Condition 2.2.1

Condition 3.1.1

(A1) Assume that $F(y|x)$ has conditional density $f(y|x)$, which is continuous at x

(A2) $f(q_\theta(x)|x) > 0$

Condition 3.1.2

With h and $|h|$ denoting the bandwidth and its determinant respectively the Euclidean norm, of the diagonal bandwidth matrix $h = diag(h_j), j = 1, \dots, d$:

(B1) $h_j > 0, h_j \to 0$ *and* $nh \to \infty$ *for* $n \to \infty$

(B2) $\frac{1}{nh^2} \to \infty, as\ n \to \infty$

Condition 3.1.3

(C1) The process $\{(Y_t, X_t)\}$ is α mixing with coefficient satisfying $\alpha(s) = o(s^{-(2+\delta)})$, for some $\delta > 0$

Condition 3.1.4

(D1) (Y_t, X_t) has joint density $f(x, y)$.Then the density $g(x)$ of X_t exists too.

(D2) For fixed (y,x), $F_x(y) \in (0,1)$ *and* $g(x) > 0$ *are continuous in the neighborhood of* x *where we want to estimate the quantile function. Then the conditional density* $F_x(y)$ *exists at* x.

In order to establish the order of the bias and the variability of the estimator, let $\Omega = \frac{1}{nh}\sum_{t=1}^{n}\gamma_t$, where $\gamma_t = K_h(x - X_t)\left(I_{\{Y_t \le y\}} - F_x(y)\right)$. Then we have the following lemma

Lemma 3.1 *Under regulatory conditions 2.2.1, 3.1.1, 3.1.2 and 3.1.3*

$$Var[\Omega] = \frac{1}{nh}V^2(y)g^2(x_i) + o\left(\frac{1}{nh}\right) \tag{3.1.1.5}$$

where $V^2(y) = \frac{1}{g(x_i)}\left(F_{x_i}(y) - F_{x_i}^2(y)\right)\int K^2(u)du$

Proof of Lemma 3.1

Since $\Omega = (\gamma_t|X_t)$ then $E[\Omega] = 0$ and

$$var[\Omega] = \frac{1}{(nh)^2}var\left(\sum_{t=1}^{n}\gamma_t\right)$$

$$= \frac{1}{(nh)^2}\left\{\sum_{t=1}^{n}var[\gamma_t] + \sum_{t \neq t'}^{n}cov[\gamma_1,\gamma_{t'}]\right\}$$

$$= \frac{1}{(nh)^2}\left\{nE[\gamma_1^2] + 2\sum_{t=2}^{n}(n-t-1)cov[\gamma_1,\gamma_t]\right\}$$

$$= \frac{1}{nh^2}E[\gamma_1^2] + 2\frac{1}{nh^2}\sum_{t=2}^{n}\left(1 - \frac{t-1}{n}\right)cov(\gamma_1,\gamma_t) \tag{3.1.16}$$

where the second term on the right hand side is of negligible magnitude as pointed out and proved in Mwita (2003).

By stationarity,

$$E[\gamma_1^2] = E\left[K^2\left(\frac{x_i - X_t}{h}\right)\left(F_{x_t}(y) - F_{x_t}^2(y)\right)\right]$$

and using Conditions 3.1.4 and the Taylor expansion of $F_{x_t}(y)$ about $F_{x_i}(y)$ and the resulting terms involving the density about $g(x)$, we get

$$E[\gamma_t^2] = (h)V^2(y)g^2(x) + o(h) \qquad (3.1.1.7)$$

Lemma 3.2 *Under Conditions 3.1.3 and 3.1.4*

$$\hat{g}(x) \xrightarrow{p} g(x) \qquad (3.1.1.8)$$

Proof of lemma 3.2

The bias of the density estimator is obtained as

$$E[\hat{g}(x)] - g(x) = \frac{|h|^2}{2}\int u^T \nabla^2 g(x) u K(u) du + o(|h|^3)$$

and the variance is,

$$var[\hat{g}(x)] = \frac{1}{nh} g(x) \int K^2(u) du + o\frac{1}{nh}$$

The mean squared error of $\hat{g}(x_i)$ then becomes of the following order

$$MSE\left(\hat{g}(x_i)\right) = O(\|h\|^4) + O\left(\frac{1}{nh}\right)$$

which goes to zero as n increases. Hence $\hat{g}(x) \xrightarrow{p} g(x)$ ■

Lemma 3.3 *Using the same conditions as in Lemma 3.1 and 3.2, we have*

$$\hat{r}_x(y) \xrightarrow{p} r_x(y) \qquad (3.1.1.9)$$

Proof of Lemma 3.3

To obtain the bias

$$E[\hat{r}_x(y)] - F_x(y)g(x)$$

$$= \frac{1}{2}|h|^2 F_x(y) \int u^T \nabla^2 g(x) u K(u) du$$

$$+ |h|^2 \nabla F_x(y) \int u \nabla g(x)^T u K(u) du$$

$$+ \frac{1}{2} g(x)|h|^2 \int u^T \nabla^2 F_x(y) u K(u) du + o(|h|^3)$$

and using the same arguments as in the proof of Lemma 3.1 we have the variance as

$$var[\hat{r}_x(y)] = \frac{1}{nh} F_x(y) g(x) \int K^2(u) du + o\left(\frac{1}{nh}\right)$$

The mean squared error for $\hat{r}_x(y)$ is of order $O\left(|h|^4 + \frac{1}{nh}\right)$ and by Condition 3.1.2, the mean squared error approaches zero as n increases, hence

$$\hat{r}_x(y) \xrightarrow{p} r_x(y). \qquad \blacksquare$$

By Slutsky's theorem, it can be proved that $\frac{\hat{r}_x(y)}{\hat{g}(x)} = \hat{F}_x(y)$ for $|h| + \frac{1}{nh} \to 0$ as $n \to \infty$. Hence, $\hat{F}_x(y)$ is a consistent estimator of $F_x(y)$ such that $\hat{F}_x(y) \xrightarrow{p} F_x(y)$. The variance and bias for $\hat{F}_x(y)$ is given in the following lemma

Lemma 3.4 *Suppose the conditions in Lemma 3.1 holds, then*

$$E\left(\hat{F}_x(y) - F_x(y)\right) = B_n(y) + o(|h|^3)$$

where

$$B_n(y) = \frac{|h|^2}{g(x)} \left\{\nabla F_x(y)^T \int u \nabla g(x)^T u K(u) du + \frac{1}{2} g(x_i) \int u^T \nabla^2 F_x(y) u K(u) du\right\}$$

and the variance is given by

$$var\left[\hat{F}_x(y)\right] = (n|h|)^{-1}V^2(y) + o((n|h|)^{-1}) \qquad (3.1.2.0)$$

Proof of Lemma 3.4

Because the numerator and the denominator of (3.1.1.4) are stochastic, we can linearize the estimator as

$$\hat{F}_x(y) = F_x(y) + \frac{\hat{r}_x(y) - F_x(y)\hat{g}(x)}{g(x)} + \frac{1}{g(x)}\left(\hat{F}_x(y) - F_x(y)\right)\left(g(x) - \hat{g}(x)\right) \qquad (3.1.2.1)$$

The consistency of $\hat{r}_x(y)$ and $\hat{g}(x)$ as shown in Lemma 3.1 implies that for large $n, h \to 0$, we have $\hat{F}_x(y) - F_x(y) = o_p(1)$ and because $g(x) - \hat{g}(x) = O_p(|h|^2)$, the product of the two quantities is of smaller order in probability. Thus using the proof of Lemma 3.2 and Lemma 3.3 in

$$E\left[\hat{F}_x(y)\right] - F_x(y) = \frac{E[\hat{r}_x(y)] - F_x(y)E[\hat{g}(x)]}{g(x)} + o_p(|h|^2)$$

results to Lemma 3.4

Lemma 3.5 *Assume conditions 2.1.2, 3.1.1 and 3.1.2, then for* $\omega_n \to 0$*, we have*

$$\hat{F}_x(y + \gamma_n) - \hat{F}_x(y) = \omega_n f_x(y) + o_p(\omega_n) + o_p\left(\frac{1}{n|h|}\right)$$

Proof of lemma 3.5

Expressing

$$\hat{F}_x(y + \omega_n) - \hat{F}_x(y) = \frac{1}{n|h|}\frac{\sum_{t=1}^{n} K(x - X_t : h)\left(1_{\{Y_t \le y + \omega_n\}} - 1_{\{Y_t \le y\}}\right)}{\hat{g}(x)} \qquad (3.1.2.2)$$

Using (3.1.1.9) we can simplify (3.1.2.0) using the same arguments of expectation as in Lemma 3.4. Then we expand the resulting $\hat{F}_x(y + \omega_n)$ and other terms involving ω_n about their corresponding function of y we end up with

$$E\left[\hat{F}_x(y+\omega_n)-\hat{F}_x(y)\right]=\omega_n f_x(y)+o(\omega_n\|h\|)$$

Similarly, the variance of (3.1.2.2) becomes

$$var\left[\hat{F}_x(y+\gamma_n)-\hat{F}_x(y)\right]=O\left(\gamma_n\frac{1}{n|h|}\right)$$

The mean squared error goes to zero as $n\to\infty$ hence (3.1.2.2) holds. ■

Lemma 3.6 *Let U be a real random variable with absolutely continuous distribution F_U and θ-quantile, $q_\theta = F_U^{-1}(\theta)$*

a) $P(H_\theta(U)\le\mu) = P\left(q_\theta-\frac{\mu}{1-\theta}\le U\le q_\theta+\frac{\mu}{\theta}\right), for\ all\ 0\le\mu<\infty$

b) Let H_θ be the θ-quantile of $H_\theta(U)$. Then $\Omega=\frac{U-q_\theta}{H_\theta}$ has zero θ-quantile and unit scale

Proof of Lemma 3.6

(a) By definition of $H_\theta(U)=H_\theta(U,q_\theta^U)$:

$$P(H_\theta(U)\le\mu)=P(U>q_\theta,\theta(U-q_\theta)\le\mu)+P(U\le q_\theta,(1-\theta)(q_\theta-U)\le\mu)$$

$$=P\left(q_\theta<U\le q_\theta+\frac{\mu}{\theta}\right)+P\left(q_\theta-\frac{\mu}{1-\theta}\le U\le q_\theta\right)$$

$$=P\left(q_\theta-\frac{\mu}{1-\theta}\le U\le q_\theta+\frac{\mu}{\theta}\right).$$

(b) $P(\Omega\le 0)=(U-q_\theta\le 0)=\theta$, that is the θ-quantile of Ω is 0 and therefore, using (a) we have

$$P(H_\theta(\Omega)\le 1)=P\left(-\frac{1}{\theta}\le\Omega\le\frac{1}{\theta}\right)$$

$$=P\left(q_\theta-\frac{H_\theta}{1-\theta}\le U\le q_\theta+\frac{H_\theta}{\theta}\right)$$

$$=P(H_\theta(U)\le H_\theta)=\theta$$ ■

Theorem 3.1 Suppose Conditions 2.2.1 hold, then as $n \to \infty$,

$$\hat{q}_\theta(x) \xrightarrow{p} q_\theta(x) \tag{3.1.2.3}$$

In addition, if Conditions 3.1.1 are satisfied then

$$\sqrt{nh}[\hat{q}_\theta(x) - q_\theta(x) - B_\theta(x) + o_\theta(h^2)] \xrightarrow{D} N\left(0, \sigma_\theta^2(x)\right) \tag{3.1.2.4}$$

where the bias and variance are given respectively by $B_\theta(x) = -\frac{B(q_\theta(x)|x)}{f(q_\theta(x)|x)}$ and

$$\sigma_\theta^2(x) = \frac{\sigma^2(q_\theta(x)|x)}{f^2(q_\theta(x)|x)} = \frac{v_0\theta(1-\theta)}{f^2(q_\theta(x)|x)g\big(q_\theta(x)\big)}$$

As an application of Theorem 3.1, the asymptotic mean squared error (AMSE) of $\hat{q}_\theta(x)$ is given by

$$AMSE\big(\hat{q}_\theta(x)\big) = \frac{h}{4}\left\{\frac{\mu_2 F^{(2)}(q_\theta(x)|x)}{f(q_\theta(x)|x)}\right\}^2 + \frac{1}{nh}\frac{v_0\theta[1-\theta]}{f^2(q_\theta(x)|x)g\big(q_\theta(x)\big)} \tag{3.1.2.5}$$

The consistent estimate for $\sigma_\theta^2(x)$ is

$$\hat{\sigma}_\theta^2(x) = \frac{v_0\theta[1-\theta]}{\hat{f}^2(q_\theta(x)|x)\hat{g}\big(q_\theta(x)\big)} \tag{3.1.2.6}$$

Proof of Theorem 3.1

First we prove (3.1.2.3) from Lemmas 3.1 and 3.2. We have for all $x \in \Re^d$ and $y, \widehat{F}_x(y) \to F_x(y)$ in probability. Because $F_x(y)$ is a distribution function it follows from Glivenko-Cantelli theorem that for generalized empirical processes based on strong mixing sequences that

$$\sup_{y\in\Re}\left|\hat{F}_x(y) - F_x(y)\right| \to 0 \qquad in\ probability \tag{3.1.2.7}$$

The uniqueness assumption of $q_\theta(x)$ implies that, for any fixed $x \in \Re^d$, there exists a $\psi > 0$ and $\delta(\psi) > 0$ such that

$$\delta = \delta(\psi) = min\{\theta - F_x(q_\theta(x) - \psi), F_x(q_\theta(x) + \psi) - \theta\} > 0$$

This implies that

$$P\{|\hat{q}_\theta(x) - q_\theta(x)| > \psi\} \leq P\left\{\left|F_x\left(\hat{q}_\theta(x)\right) - F_x\left(q_\theta(x)\right)\right| > \delta\right\}$$

$$\leq P\left\{\left|F_x\left(\hat{q}_\theta(x)\right) - \hat{F}_x\left(q_\theta(x)\right)\right| > \delta - \frac{1}{n}\right\}$$

$$\leq P\left\{\sup_y\left|\hat{F}_x(y) - F_x(y)\right| > \delta'\right\} \qquad (3.1.2.8)$$

for arbitrary $\delta' < \delta$ and n large enough. Here, we used $F_x\left(q_\theta(x)\right) = \theta$ and $\theta \leq \hat{F}_x\left(q_\theta(x)\right) \leq \theta + \frac{1}{n}$. (3.1.2.8) tends to zero by (3.1.2.7) hence, (3.1.2.3) holds true. To prove (3.1.2.4), let $b_n = -\frac{B_n\left(q_\theta(x)\right)}{f_x\left(q_\theta(x)\right)}$ and $v = \frac{V\left(q_\theta(x)\right)}{f_x\left(q_\theta(x)\right)}$. For any w

$$q_n(w) = P\left(\frac{\hat{q}_\theta(x_i) - q_\theta(x_i) - b_n}{v} \leq w\right)$$

$$= P(\hat{q}_\theta(q) \leq q_\theta(q) + b_n + vw)$$

As $\hat{F}_x(y)$ is increasing, but not necessarily strictly, we have

$$P\left(F_x\left(\hat{q}_\theta(x)\right) < \hat{F}_x(q_\theta(x) + b_n v + w)\right) \leq q_n(w)$$

$$\leq P\left(F_x\left(\hat{q}_\theta(x)\right) \leq \hat{F}_x(q_\theta(x) + b_n v + w)\right)$$

By the same argument as in (3.1.2.8) we may replace $\hat{F}_x\left(\hat{q}_\theta(x)\right)$ by $F_x\left(q_\theta(x)\right)$ up to an error of $\frac{1}{n}$ at most. Neglecting the $\frac{1}{n}$ term which is asymptotically negligible anyhow, we get

$$q_n(w) \approx P\left(F_x\big(q_\theta(x)\big) \leq \hat{F}_x\big(q_\theta(x) + b_n + vw\big)\right)$$

$$\approx P\left(-f_x\big(q_\theta(x)\big).\,\delta_n \leq \hat{F}_x\big(q_\theta(x)\big) - F_x\big(q_\theta(x)\big)\right) \qquad (3.1.2.9)$$

with $\delta_n = b_n + vw$ where we have used lemma 4.5 and neglected the $o(\delta_n)$ and $o\left(\frac{1}{nh^{(i)}}\right)$. Using the asymptotic normality

$$\sqrt{nh}\left(\left(\hat{F}_X(y) - F_X(y) - B_n(y) + o_p(h^2)\right)\right) \xrightarrow{d} N\big(0, v^2(y)\big) \qquad (3.1.3.0)$$

with $y_\theta = q_\theta(x)$, we get

$$q_n(w) \sim P\left(\sqrt{nh^{(i)}}\,\frac{\hat{F}_x(y_\theta) - F_x(y_\theta) - B_n(y_\theta)}{V(y_\theta)} \geq \frac{-f_x(y_\theta)\delta_n - B_n(y_\theta)}{V(y_\theta)}\sqrt{nh^{(i)}}\right)$$

$$\sim \Phi\left(\left(\sqrt{nh^{(i)}}\right)\frac{f_{x_i}(y_\theta).\,(b_n + vw) + B_n(y_\theta)}{V(y_\theta)}\right)$$

$$= \Phi(w)$$

by our choice of b_n and v. This proves the theorem. ■

These results can be used to construct confidence interval for the estimators as well as other relevant inferences.

3.2 Confidence Intervals for conditional quantiles

Asymptotic distribution of sample quantiles which make statistical inference possible is given by

$$\sqrt{n}f\big(q_\theta(x)\big)\big(\hat{q}_\theta(x) - q_\theta(x)\big) \xrightarrow{d} B(u) \qquad (3.2.0.1)$$

where, $B(u)$, $0 < u < 1$, is Brownian Bridge, zero mean Gaussian process with covariance

$$E[B(s)B(t)] = min(s,t) - st$$

Confidence intervals for a parameter $q_\theta(x)$ can be formed from the asymptotic distribution of $\hat{q}_\theta(x)$ as

$$\sqrt{n}\left(\hat{q}_\theta(x) - q_\theta(x)\right) \xrightarrow{d} B(u)/fq_\theta(x) \qquad (3.2.0.2)$$

This formulae has a severe disadvantage, it requires estimation of $f\left(q_\theta(x)\right)$ from the values of sample quantile function $\tilde{Q}_Y(u)$, similarly for $\tilde{Q}_{Y|X=x}(u)$. One can obtain the confidence interval for $q_\theta(x)$ using facts such as

$$(\sqrt{n}\left(\hat{F}_n\left(q_\theta(x)\right)\right) - u) \xrightarrow{d} B(u) \qquad (3.2.0.3)$$

one can find functions $c_1(u)$ and $c_2(u)$ such that with probability greater than α, for all u,

$$u - \left(c_1(u)/\sqrt{n}\right) < \hat{F}_n\left(Q(u)\right) < u + \left(c_2(u)/\sqrt{n}\right)$$

$$\hat{Q}_n\left(u - \left(c_1(u)/\sqrt{n}\right)\right) < Q(u) < Q_n\left(u + \left(c_2(u)/\sqrt{n}\right)\right) \qquad (3.2.0.4)$$

3.3 Smoothing parameter choice under quantiles

The smoothing parameter selectors in section 2.3 can be adapted to quantile regression estimation. The first modification concerns $log(\hat{\sigma}^2)$ in (2.3.0.3). Since the quantile estimator (3.1.0.2) falls into the class of M-estimators we can proceed and interpret the ρ_θ function as "$-\log likelihood = \rho_\theta$", so that the criterion like AIC is modified by using $\frac{1}{n}\sum_{i=1}^{n}\rho_\theta\left(y_i - \hat{q}_\theta(x_i)\right)$ instead of $\hat{\sigma}$.

The second modification concerns the smoother matrix S_h. Estimator (3.1.0.2) does not lead to a linear estimator $\hat{y}$ = $(S_h)y$ because it is carried out by iteratively reweighted least squares. The smoother matrix(S_h) can be approximated by the implied smoother matrix from the last iteration of the iteratively reweighted least squares fit of the model. With these modifications, the suitable smoothing parameter for the nonparametric quantile regression is choose the bandwidth to the minimiser of

$$2log\left(\frac{1}{n}\sum_{i=1}^{n}\rho_\theta\left(y_i - \hat{q}_\theta(x_i)\right)\right) + \psi(S_h) \qquad (3.3.0.1)$$

where, $\psi(\cdot)$ is one of the penalizing function and S_h is the appropriate smoother matrix.

CHAPTER FOUR

4.0 MODELING EXTREME VALUES

4.1 Extreme value theory

Modeling of extreme events is the central issue in extreme value theory and the main purpose of the theory is to provide asymptotic models with which we can model the tails of a distribution. Energy related utilities, exhibit certain risk characteristics that are different from traditional financial assets like stocks and bonds. Further supply and demand changes of these utilities are translated immediately into price changes of other related complements. Such turbulence in the utility market characterized by the substantial increase in market volatility has generated discussions on the appropriate measures of market risks and margin settings in utility related institutions. Several studies exist that compare the forecasts performances of different risk models among others, Kuester *et.al* (2006), and Manganelli and Engle (2001). But nonparametric quantile regression as a tool of risk estimation is rarely considered. Chen and Tang (2005), investigates nonparametric risk estimation when no regressors are present. Taylor (2008) combines double kernel quantile regression with exponential smoothing in the time domain to study risks. Since data sparseness is more severe in extreme quantiles, we embark upon refining nonparametric quantile regression methods with extreme value theory to model extreme quantiles accurately. We are motivated by the fact that it is prudent for risk managers to focus on the conditional distributions of profit and loss, when extra ordinary or rare events occur. One analytical tool of assessing riskiness of trading activities which is commonly used in practice by institutions and asset managers for minimization of risk is $Value\ at\ Risk$ (VaR).

4.2 Value at Risk

In financial planning, Value at Risk (VaR) can be defined as the maximal loss of a financial position during a given period for a given probability. For a regulatory committee, VaR is the minimal loss which may occur under extra ordinary market circumstances. The main advantage of VaR over other risk measures is its inherent theoretical simplicity. The expected loss given that the loss is at least as large as some given quantile of the loss distribution, herewith VaR is the Expected shortfall ES. According to Artzener *et al.* (1999), Expected shortfall is a coherent risk measure which posses the properties of homogeneighty; increasing the size of a portfolio by a factor should scale its risk measure by some factor, property of monotonicity; a portfolio must have greater risk if it has systematic lower values than another, property of translation variance also referred to as risk free; adding some amount of cash to a portfolio should reduce the risk by the same amount and the property of sub additivity; merging portfolio cannot increase risk.

There are two approaches in VaR calculation.

(i) The unconditional approach which uses the historical returns of the instruments to VaR calculation and

(ii) The conditional approach which uses historical data and explanatory variables to calculate VaR. The explanatory variables may include the macroeconomic variables of an economy and the accounting variables of the companies involved.

Since one bad day in the market makes the probability of the next day somehow worse off, one would expect VaR to increase as the past returns become very negative, as a result VaR depend on the past returns in some way.

Conditional Value at risk herewith $\left(VaR_{(C)}\right)$ can be modeled parametrically and also nonparametrically. Parametric models for $\left(VaR_{(C)}\right)$ can be more efficient if the underlying functions are correctly specified. Engle and Manganelli (2004) applies The Generalized Autoregressive Conditional Heteroscedastic $GARCH$ type parametric model based on regression quantiles to study $\left(VaR_{(C)}\right)$. This volatility based approach to estimating conditional quantiles has two draw-backs. On one hand, the distribution of innovations $\{\varepsilon_t\}$ has to be specified like as standard normal or as a heavy tailed distribution like the t distribution. Secondly volatility based estimates of VaR tacitly assumes that the extreme negative unit rates of change (returns) shows the same kind of random pattern as the majority of typical unit rates of change which mainly determine the volatility estimate.

However a $(VaR_{(C)})$ misspecification may cause serious bias and model constraints and may distort the underlying distribution. To this effect, nonparametric modeling becomes more appealing in the sense that little or no restrictive prior information on functional form is needed and it may further provide a useful insight for further parametric fitting.

4.3 Nonparametric estimation of Value at Risk

Let us assume that the observed data $\{(X_t, r_t); 1 \leq t \leq n\}, X_t \in \Re^d$ are available and they are observed from a stationary time series model. Let r_t be the loss or risk variable which can be the negative logarithm of return in finance or a unit rate of change in econometrics demand modeling. Let X_t be the lag variables of r_t. X_t , may be a vector which may include both the lag variables and the economic (exogenous) variables. We consider the case when X_t is a scalar ($d = 1$). The methodology also applies when $d \geq 1$.

Define the nonparametric estimation of the conditional value at risk $(VaR_{(C)})$ as $q_\theta^c(x)$ the solution of

$$P(r_t \geq q_\theta^c(x) | X_t = x) = Z(q_\theta^c(x)|x) = \theta \qquad (4.3.0.1)$$

We can express $q_\theta^c(x) = Z^{-1}(\theta|x)$, where $Z(r|x)$ is the conditional survival function of r_t given $X_t = x$ that is $Z(r|x) = 1 - F(r|x)$ and $F(r|x)$ is the conditional cumulative distribution function. The nonparametric estimation of the Expected Shortfall ES is defined as

$$\gamma_\theta(x) = E[r_t | r_t \geq q_\theta^c(x), X_t = x]$$

It can be shown that

$$\gamma_\theta(x) = \int_{q_\theta^c(x)}^{\infty} r\, f(r|x)\, dr/\theta \qquad (4.3.0.2)$$

where, $f(r|x)$ is the conditional probability density function of r_t given $X_t = x$. To estimate $\gamma_\theta(x)$ we have

$$\hat{\gamma}_\theta(x) = \int_{\widehat{q_\theta^c}(x)}^{\infty} r\, \hat{f}(r|x)\, dr/\theta \qquad (4.3.0.3)$$

Where, $\widehat{q_\theta^c}(x)$ is a nonparametric estimation of $q_\theta^c(x)$ and $\hat{f}(r|x)$ is a nonparametric estimation of $f(r|x)$, but the bandwidth of $\widehat{q_\theta^c}(x)$ and $\hat{f}(r|x)$ are not necessary the same. The question is how to provide a better estimate of $f(r|x)$ and $q_\theta^c(x)$. Scaillet (2005) used the Nadaraya-Watson type double kernel method to estimate $f(r|x)$ and then estimated $q_\theta^c(x)$ by inverting the estimated conditional survival function, denoted by $\widetilde{q_\theta^c}(x)$ and finally estimated $\gamma_\theta(x)$ by plugging $\tilde{f}(r|x)$ and $\widetilde{q_\theta^c}(x)$ into (4.3.0.2). But it has been shown that the Nadaraya-Watson kernel type procedures have serious drawbacks. The asymptotic bias involves the design density so that they cannot be adaptive thus they require boundary modification. In particular boundary modification effects might cause a serious problem for estimating $q_\theta^c(x)$ since it is only concerned with the tail probability. Despite this Nadaraya-Watson has the properties of positivity and monotonicity as shown in section 2.2.2, which are advantageous if the method of inverting conditional distribution estimator. To overcome these difficulties we apply the Weighted Nadaraya-Watson estimator based on the empirical likelihood principle as discussed in chapter 2 which accommodate all nice properties of monotonicity, continuity, differentiability and lying between zero and one and the attractive asymptotic properties of mathematical efficiency and design adaptation.

We combine the properties of Weighted Nadaraya-Watson estimate and the Double Kernel method of Yu and Jones (1998) and we refer to the estimator as the New Weighted Estimator NWE given as

$$\hat{f}_{NWE}(r|x) = \sum_{t=1}^{n} P_t(x,h) r_t^*(r) \quad (4.3.0.4)$$

where $P_t(x,h)$ is given in section 2.2.3, $r_t^*(r) = K_{h_0}(r - r_t)$, is an initial estimate of $f(r|x)$ smoothing in the r direction. Therefore,

$$\hat{F}_{NWE}(r|x) = \int_{-\infty}^{r} \hat{f}_{NWE}(y|x)dy = \sum_{t=1}^{n} P_t(x,h) G_{h0}(r - r_t) \quad (4.3.0.5)$$

$G_{h_0}(u) = G(u|h_0)$ is the distribution function of $K(\cdot)$ of which for any symmetric density function $K(\cdot)$,

$$E\{K_{h_0}(r - r_t)|X_t = x\}$$

$$= f(r|x) + \frac{h_0^2}{2}\mu_2(K) f''(r|x) + o(h_0^2) \approx f(r|x) \quad as \quad h_0 \to 0 \quad (4.3.0.6)$$

where $\mu_2(K) = \int_{-\infty}^{\infty} u^2 K(u)du$, $f''(r|x) = \frac{\partial^2}{\partial r^2} f(r|x)$ and $\approx$ denotes an approximation ignoring the higher terms.

We show that New Weighted Estimator enjoy the same convergence rates as the Weighted Nadaraya-Watson estimator at both the boundary and the interior points. $\hat{F}_{NWE}(r|x)$, is a monotone cumulative distribution function where,

$$0 \le \hat{F}_{NWE}(r|x) \le 1,\ \hat{F}_{NWE}(-\infty|x) = 0 \text{ and } \hat{F}_{NWE}(\infty|x) = 1.$$

Also $\hat{F}_{NWE}(r|x)$ is continuous and differentiable in r. Plugging in $\widehat{q_\theta^c}(x)$ and $\hat{f}_{NWE}(r|x)$ into (4.3.0.2) we obtain the nonparametric estimator of $\gamma_\theta(x)$ as

$$\hat{\gamma}_\theta(x) = \theta^{-1} \int_{\widehat{q_\theta^c}(x)}^{\infty} r\, \hat{f}_{NWE}(r|x)dr = \theta^{-1} \sum_{t=1}^{n} P_t(x,h) \int_{\widehat{q_\theta^c}(x)}^{\infty} r\, K_{h_0}(r - r_t)dr$$

$$= \theta^{-1} \sum_{t=1}^{n} P_t(x,h) \left[r_t \bar{G}_{h0}\left(\widehat{q_\theta^c}(x) - r_t \right) + h_0 G_{1,h0}\left(\widehat{q_\theta^c}(x) - r_t \right) \right]$$

where

$$\bar{G}(u) = 1 - G(u),\quad G_{1,h_0}(u) = G_1(u|h_0),$$

and

$$G_1(u) = \int_u^{\infty} v\, K(v)dv$$

4.3.1 Distribution theory

Let $\alpha(K) = \int_{-\infty}^{\infty} uK(u)\bar{G}(u)du$ and $\mu_j(P) = \int_{-\infty}^{\infty} u^j P(u)du$, also for any $j \ge 0$, write

$$l_j(u|v) = E\left[r_t^j I(r_t \geq u)|X_t = v\right] = \int_u^\infty r^j f(r|v)dr,$$

$$l_j^{a,b}(u|v) = \frac{\partial^{a,b}}{\partial u^a \partial v^b} l_j(u|v)$$

and

$$l_j^{a,b}(q_\theta^c(x)|x) = l_j^{a,b}(u|v)\Big|_{u = q_\theta^c(x), v = x}.$$

Clearly,

$$l_0(u|v) = Z(u|v) \text{ and } l_j^{2,0}(u|v) = -\left[u^j f'(u|v) + ju^{j-1} f(u|v)\right].$$

We now list the following Assumptions.

Assumption 4.3.1

(A1) *For fixed r and x, $0 \leq F(r|x) < 1$, $g(x) > 0$. The marginal density of X_t is continuous at x and $F(r|x)$ has continous second order derivative with respect to both x and r*

(A2) *The kernels $K(\cdot)$ and $P(\cdot)$ are symmetric bounded and compactly supported Density*

(A3) $h \to 0$, *and* $nh \to \infty$, *and* $h_0 \to 0$ *and* $nh_0 \to \infty$ *as* $n \to \infty$

If $\alpha(t)$ decays geometrically then we have condition (A4).

(A4) *Let $g_{1,t}(\cdot,\cdot)$ be the joint density of X_1 and X_t for $t \geq 2$. Assume that*

$\left|g_{1,t}(u, v) - g(u)g(v)\right| \leq C < \infty$, *for all u and v. And*

(A5) *The process $\{(X_t, r_t)\}$ is stationary $\alpha - mixing$ with the mixing coefficient satisfying $\alpha(t) = O(t^{-\omega-2})$*

(A6) $nh^{\left(\frac{\omega+2}{\omega}\right)} \to \infty$. *This is satisfied by the bandwidth* $h = n^{-\frac{1}{5}}$ *if* $\omega > \frac{1}{2}$

(A7) $h_0 = o(h)$. *This means that the initial step bandwidth should be chosen as small as possible so that the bias from the initial step can be ignored*

Assumption 4.3.2

(B1) *Assume that* $E(|r_t|^{\omega}|X_t = u) \leq C_3 < \infty$ *for some* $\omega > 2$, *in the neighborhood of* x.

(B2) *Assume that* $\left|g_{1,t}(r_1, r_t|x_1, x_2)\right| \leq C_1 < \infty$ *for all* $t \geq 2$, *where* $g_{1,t}(r_1, r_t|x_1, x_2)$ *is the Conditional density of* r_1 *and* r_t *given* $X_1 = x_1$, *and* $X_2 = x_2$. *This implies that the higher the moments involved, the faster the decaying rate of* $\alpha(\cdot)$

(B3) *The mixing coefficient of the* $\alpha - mixing$ *process* (X_t, r_t) *satisfies*

$\sum_{t\geq 1} t^a \alpha^{\left(\frac{\omega-2}{\omega}\right)}(t) < \infty$ *for some* $a > \left(\frac{\omega-2}{\omega}\right)$, *where* ω *is given in assumption (B1)*

(B4) *Assume that there exists a sequence of integers* $z_n > 0$ *such that* $z_n \to \infty$,

$z_n = o\left((nh)^{1/2}\right)$ *and* $\left(\frac{n}{h}\right)^{\frac{1}{2}} \alpha(z_n) \to 0$ *as* $n \to \infty$

(B5) *There exists* $\omega^* > \omega$ *such that* $E\left(|r_t|^{\omega^*}\middle|X_t = u\right) \leq C_4 < \infty$ *in a neighborhood of* x,

$\alpha(t) = O\left(t^{-\theta^*}\right)$, *where* ω *is given in assumption B1,* $\theta^* \geq \omega^* \omega/\{2(\omega^* - \omega)\}$*and* $n^{\omega/4} h^{\omega/\omega^* - 1 - \omega/4} = O(1)$. $\alpha - mixing$ *Imposed in this assumption is weaker than* $\beta - mixing$ *in Hall et.al. (1999).*

4.3.2 Asymptotic properties of The New Weighted Estimator

To investigate the asymptotic behaviour of $\hat{f}_{WNE}(r|x)$ including the asymptotic normality, we have the following lemmas

Lemma 4.1: *Under Assumptions 4.3.1 (A1 – A5), we have*

$$\lambda = -h\lambda_0\{1 + o_p(1)\} \quad \textit{and} \quad p_t(x) = n^{-1}b_t(x)\{1 + o_p(1)\}$$

where $\lambda_0 = \mu_2(W)g^{'}(x)/[2\mu_2(W^2)g(x)]$ *and*

$$b_t(x) = [1 - h\lambda_0(X_t - x)W_h(x - X_t)]^{-1}$$

Further, we have $p_t(ch) = n^{-1}b_t^c(ch)\{1 + o_p(1)\}$ where $b_t^c(x) = [1 + \lambda_c(X_t - x)K_h(x - X_t)]^{-1}$

Lemma 4.2: *Under Assumptions 4.3.1 (A1-A5), we have for any* $j \geq 0$,

$$J_j = n^{-1}\sum_{t=1}^{n} c_t(x)\left(\frac{X_t - x}{h}\right)^j = g(x)\mu_j(W) + O_p(h^2)$$

where $c_t(x) = b_t(x)W_h(x - X_t)$. It follows from Lemmas 4.1 and 4.2 that

$$W_{c,t}(x,h) \approx \frac{b_t(x)W_h(x - X_t)}{\sum_{t=1}^{n} b_t(x)W_h(x - X_t)} \approx n^{-1}g^{-1}(x)b_t(x)W_h(x - X_t) = \frac{c_t(x)}{ng(x)} \quad (4.3.2.1)$$

Theorem 4.1 *Under Assumption 4.3.1 (A1-A6) with* h *in A3 and A6 replaced by* h_0h, *we have*

$$\sqrt{nh_0h}\left[\hat{f}_{WNE}(r|x) - f(r|x) - B_f(r|x)\right] \to N\left\{0, \sigma_f^2(r|x)\right\} \tag{4.3.2.2}$$

where the asymptotic bias is

$$B_f(r|x) = \frac{h^2}{2}\mu_2(W)f^{0,2}(r|x) + \frac{h_0^2}{2}\mu_2(K)f^{2,0}(r|x)$$

and the asymptotic variance is

$$\sigma_f^2(r|x) = \mu_0(K^2)\mu_0(W^2)f\,(r|x)/g(x)$$

Proof of Theorem 4.1

We decompose $\hat{f}_c(r|x) - f(r|x)$ into three parts as follows

$$\hat{f}_c(r|x) - f(r|x) = I_1 + I_2 + I_3 \tag{4.3.2.3}$$

where with $\varepsilon_{t,1} = r_t^*(r) - E(r_t^*(r)|X_t)$

$I_1 = \sum_{t=1}^{n} \varepsilon_{t,1} W_{c,t}(x,h)$, $I_2 = E[(r_t^*(r)|X_t) - f(r|X_t)]W_{c,t}(x,h)$, and

$$I_3 = \sum_{t=1}^{n} \frac{1}{2} f^{0,2}(r|x)W_{c,t}(x,h)(X_t - x)^2 + o_p(h^2)$$

$$= \frac{1}{2} g^{-1}(x) f^{2,0}(r|x) n^{-1} \sum_{t=1}^{n} c_t(x)(X_t - x)^2 + o_p(h^2)$$

$$= \frac{h^2}{2}\mu_2(W)f^{2,0}(y|x) + o_p(h^2)$$

As in the proof of Lemma 4.2, we have

$$I_2 = \frac{h_0^2\mu_2(K)}{2g(x)} n^{-1} \sum_{t=1}^{n} f^{2,0}(r|X_t)c_t(x) + o_p(h_0^2 + h^2)$$

$$= \frac{h_0^2}{2}\mu_2(K)f^{2,0}(r|x) + o_p(h_0^2 + h^2)$$

Therefore

$$I_2 + I_3 = \frac{h^2}{2}\mu_2(W)f^{2,0}(r|x) + \frac{h_0^2}{2}\mu_2(K)f^{2,0}(r|x) + o_p(h_0^2 + h^2)$$

$$= B_f(r|x) + o_p(h^2 + h_0^2)$$

Thus (4.3.2.3) becomes

$$\sqrt{nh_0h}\left[\hat{f}_c(r|x) - f(r|x) - B_f(r|x) + o_p(h^2 + h_0^2)\right] = \sqrt{nh_0h}I_1$$

$$= g^{-1}(x)I_4\left\{1 + o_p(1)\right\} \to N\left\{0, \sigma_f^2(r|x)\right\}$$

where $I_4 = \sqrt{h_0h/n}\sum_{t=1}^{n}\varepsilon_{t,1}c_t(x)$. This together with Lemma 4.2 proves the theorem .

To study the asymptotic behavior of $\widehat{Z}_{NWE}(r|x)$, similar to Theorem 4.1 we have the following asymptotic normality for $\hat{Z}_{NWE}(r|x)$

Theorem 4.2 *Under Assumption 4.3.1 (A1-A6), we have*

$$\sqrt{nh}\left[\hat{Z}_{NWE}(r|x) - Z(r|x) - B_z(r|x)\right] \to N\{0, \sigma_z^2(r|x)\} \quad (4.3.2.4)$$

Where the asymptotic bias is given by

$$B_z(r|x) = \frac{h^2}{2}\mu_2(W)Z^{0,2}(r|x) - \frac{h_0^2}{2}\mu_2(K)f^{1,0}(r|x)$$

and the asymptotic variance is

$$\sigma_z^2(r|x) = \mu_0(W^2)Z(r|x)\,[1 - Z(r|x)]/g(x)$$

In particular, if Assumption 4.3.1*(A7)* holds true, then

$$\sqrt{nh}\left[\hat{Z}_{NWE}(r|x) - Z(r|x) - \frac{h^2}{2}\mu_2(W)Z^{0,2}(r|x)\right] \to N\{0, \sigma_z^2(r|x)\}$$

The $AMSE$ of $\hat{Z}_{NWE}(r|x)$ is

$$AMSE\left(\hat{Z}_{NWE}(r|x)\right) = \frac{\{h^2\mu_2(W)Z^{0,2}(r|x) - h_0^2\mu_2(K)f^{1,0}(r|x)\}^2}{4}$$

$$+ \frac{1}{nh}\frac{\mu_0(W^2)Z(r|x)[1 - Z(r|x)]}{g(x)} \qquad (4.3.2.5)$$

Proof of Theorem 4.2

Similar to (4.3.2.3), we have

$$\hat{Z}_c(r|x) - Z(r|x) = I_5 + I_6 + I_7$$

where $\varepsilon_{t,2} = \bar{G}_{h_0}(r - r_t) - E\left(\bar{G}_{h_0}(r - r_t|X_t)\right)$

$$I_5 = \varepsilon_{t,2}W_{c,t}(x,h).\ I_6 = [\ E\left(\bar{G}_{h_0}(r - r_t|X_t)\right) - Z(r|X_t)]W_{c,t}(x,h)$$

and

$$I_7 = \sum_{t=1}^{n}[Z(r|X_t) - Z(r|x)]\,W_{c,t}(x,h)$$

Similar to the analysis of I_2 by Taylor expansion, Lemmas 4.1 and 4.2, we have

$$I_7 = \sum_{t=1}^{n}\frac{1}{2}S^{0,2}(r|x)\,W_{c,t}(x,h)(X_t - x)^2 + o_p(h^2)$$

$$= \frac{1}{2}Z^{0,2}(r|x)g^{-1}(x)n^{-1}\sum_{t=1}^{n} c_t(x)(X_t - x)^2 + o_p(h^2)$$

$$= \frac{h^2}{2}\mu_2(W)Z^{0,2}(r|x) + o_p(h^2)$$

To evaluate I_6, first, we consider the following

$$E\left(\bar{G}_{h_0}(r - r_t|X_t = x)\right) = \int_{-\infty}^{\infty} K(u)Z(r - h_0u|x)du$$

$$Z(r|x) + \frac{h_0^2}{2}\mu_2(K)Z^{2,0}(r|x) + o(h_0^2)$$

$$= Z(r|x) - \frac{h_0^2}{2}\mu_2(K)f^{1,0}(y|x) + o(h_0^2) \qquad (4.3.2.6)$$

By (4.3.2.6) and following the same arguments as in the proof of Lemma 4.2 and by (4.3.2.5)

$$\sqrt{nh}\left[\hat{Z}_c(r|x) - Z(r|x) - B_S(r|x) + o_p(h^2 + h_0^2)\right] = \sqrt{nh}I_5$$

and

$$\sqrt{nh}I_5 = g^{-1}(x)I_8\{1 + o_p(1)\} \to N\{0, \sigma_Z^2(r|x)\}$$

■

4.3.3 Asymptotic theory of conditional value at risk

By the differentiability of $\hat{Z}_{NWE}\left(\widehat{q_\theta^c}(x)\middle|x\right)$, we use the Taylor expansion and ignore the higher order terms to obtain

$$\hat{Z}_{NWE}\left(\widehat{q_\theta^c}(x)\middle|x\right) = \theta \approx \hat{Z}_{NWE}(q_\theta^c(x)|x) - \hat{f}_{NWE}(q_\theta^c(x)|x)\left(\widehat{q_\theta^c}(x) - q_\theta^c(x)\right)$$

Then by Theorem 4.1

$$\widehat{q_\theta^c}(x) - q_\theta^c(x) \approx \left[\hat{Z}_{NWE}(q_\theta^c(x)|x) - \theta\right]/\hat{f}_{NWE}(q_\theta^c(x)|x)$$

$$\approx \left[\hat{Z}_{NWE}(q_\theta^c(x)|x) - \theta\right]/f_{NWE}(q_\theta^c(x)|x)$$

As an application of Theorem 4.2, we can establish the following theorem for the asymptotic normality of $\widehat{q_\theta^c}(x)$, whose proof is similar to that for Theorem 4.2.

Theorem 4.3: *Under Assumptions 4.3.1 (A1-A6), we have*

$$\sqrt{nh}\left[\widehat{q_\theta^c}(x) - q_\theta^c(x) - B_\epsilon(x)\right] \to N\{0, \sigma_\epsilon^2(x)\}, \qquad (4.3.3.1)$$

Where the asymptotic bias is $B_\epsilon(x) = B_Z(q_\theta^c(x)|x)/f(q_\theta^c(x)|x)$ and the asymptotic variance is

$$\sigma_\epsilon^2(x) = \mu_0(WP^2)\theta(1-\theta)/[g(x)f^2(q_\theta^c(x)|x)].$$

In particular, if Assumption 4.3.1 *(A7)* holds, then

$$\sqrt{nh}\left[\widehat{q_\theta^c}(x) - q_\theta^c(x) - \frac{h^2}{2}\frac{Z^{0,2}(q_\theta^c(x)|x)}{f\left(q_\theta^c(x)\middle|x\right)}\mu_2(P)\right] \to N\{0,\ \sigma_\gamma^2(x)\} \qquad (4.3.3.2)$$

The significance of Theorem 4.3 is that $\widehat{q_\theta^c}(x) - q_\theta^c(x) = O_p\left(h^2 + h_0^2 + (nh)^{-1/2}\right)$ so that $\widehat{q_\theta^c}(x)$ is a consistent estimator of $q_\theta^c(x)$. The term $Z^{0,2}(q_\theta^c(x)|x)$ is the bias term involving the second derivative of the conditional distribution function with respect to x, which utilizes the approximation of the conditional VaR function. Theorems 4.2 and 4.3 implies that if the initial bandwidth h_0 is chosen as small as possible such that $h_0 = o(h)$, thus the final estimates of $Z(r|x)$ and $q_\theta^c(x)$ are not sensitive to the choice of h_0 as long as it satisfies Assumption 4.3.1 *(A7)* making the selection of bandwidth much easier in practice.

By Theorem 4.3 and following Yu and Jones (1998), the AMSE of $\widehat{q_\theta^c}(x)$ is given by

$$AMSE(\widehat{q_\theta^c}(x) = \frac{\{h^2 Z^{0,2}(q_\theta^c(x)|x)\mu_2(P) - h_0^2 f^{1,0}(q_\theta^c(x)|x)\mu_2(K)\}^2}{4f^2\left(q_\theta^c(x)\middle|x\right)}$$

$$+\frac{1}{nh}\frac{\mu_0(P^2)[\theta(1-\theta) + 2h_0 f(q_\theta^c(x)|x)\alpha(K)]}{f^2\left(q_\theta^c(x)\middle|x\right)g(x)} \qquad (4.3.3.3)$$

A comparison of (4.3.3.3) with the Weighted Nadaraya-Watson estimator reveals that (4.3.3.3) has two extra terms which are negligible if assumption *(A7)* is satisfied due to vertical smoothing in the r direction.

By minimizing $AMSE$ in (4.3.3.3) and taking $h_0 = o(h)$, therefore we obtain the optimal bandwidth given by

$$h_{opt,\epsilon}(x) = \left[\frac{\mu_0(P^2)\theta(1-\theta)}{\{\mu_2 Z^{0,2}(q_\theta^c(x)|x)\}^2 g(x)}\right]^{1/5} n^{-1/5} \tag{4.3.3.4}$$

Thus the optimal rate of the AMSE of $\widehat{q_\theta^c}(x)$ is $n^{-4/5}$.

Under Assumption *A7*, the asymptotic result at the boundary point $x = bd$ for $q_\theta^c(x)$ is

$$\sqrt{nh}\left[\left(\widehat{q_\theta^c}(x)(bd)\right) - q_\theta^c(x)(bd) - B_{q,b}\right] \to N\left(0, \sigma_{q,b}^2\right) \tag{4.3.3.5}$$

where the asymptotic bias is

$$B_{q,b} = h^2\beta_2(b)Z^{0,2}(q_\theta^c(x)(0+)|0+)/[2\beta_1(b)f(q_\theta^c(x)(0+)|0+)]$$

and the asymptotic variance is

$$\sigma_{q,b}^2 = \beta_0(0)\theta[1-\theta]/[\beta_1^2(b)f^2(q_\theta^c(x)(0+)|0+)g(0+)]$$

To examine the asymptotic behavior of $\hat{\gamma}_\theta(x)$ we have

Theorem 4.4: *Under Assumptions 4.3.1 (A1 – A4) and (B2 –B5)*

$$\sqrt{nh}\left[\hat{\gamma}_\theta(x) - \gamma_\theta(x) - B_\gamma(x)\right] \to N\left\{0, \sigma_\gamma^2(x)\right\}, \tag{4.3.3.6}$$

where the asymptotic bias is $B_\gamma(x) = B_{\gamma,0}(x) + \frac{h_0^2}{2}\gamma_2(K)\theta^{-1}f(q_\theta^c(x)|x)$ with

$$B_{\gamma,0}(x) = \frac{h^2}{2}\mu_2(P)\theta^{-1}\left[l_1^{0,2}(q_\theta^c(x)|x) - q_\theta^c(x)Z^{0,2}(q_\theta^c(x)|x)\right]$$

and the asymptotic variance is

$$\sigma_\gamma^2(x) = \frac{\gamma_0(P^2)}{\theta g(x)}[\theta^{-1}l_2(q_\theta^c(x)|x) - \theta\gamma_\theta^2(x) + (1-\theta)q_\theta^c(x)\{q_\theta^c(x) - 2\gamma_\theta(x)\}]$$

In particular if Assumption *(A7)* holds true then,

$$\sqrt{nh}\left[\hat{\gamma}_\theta(x) - \gamma_\theta(x) - B_{\gamma,0}(x)\right] \to N\left\{0, \sigma_\gamma^2(x)\right\} \qquad (4.3.3.7)$$

Theorem 4.4 concludes that $\hat{\gamma}_\theta(x) - \gamma_\theta(x) = O_p\left(h^2 + h_0^2 + (nh)^{-1/2}\right)$ so that $\hat{\gamma}_\theta(x)$ is a consistent estimator of $\gamma_\theta(x)$ with a convergence rate$(nh)^{-1/2}$.

The $AMSE$ of $\hat{\gamma}_\theta(x)$ is given by

$$AMSE\left(\hat{\gamma}_\theta(x)\right) = \frac{1}{nh}\sigma_\gamma^2(x) + \left\{B_{\gamma,0}(x) + \frac{h_0^2}{2}\gamma_2(K)\theta^{-1}f(q_\theta^c(x)|x)\right\}^2 \qquad (4.3.3.8)$$

Minimizing $AMSE$ in (4.3.3.8) with respect to h yields the optimal bandwidth given by

$$h_{opt,\gamma}(x) = \left[\frac{\sigma_\gamma(x)}{\gamma_2(P)\theta^{-1}\left\{l_1^{0,2}\left(q_\theta^c(x)\middle|x\right) - q_\theta^c(x)Z^{0,2}\left(q_\theta^c(x)\middle|x\right)\right\}}\right]^{2/5} n^{-1/5} \qquad (4.3.3.9)$$

Therefore the minimal rate of $AMSE$ for $\hat{\gamma}_\theta(x)$ is $n^{-4/5}$.

The asymptotic results for $\hat{\gamma}_\theta(x)$ at the left boundary point $x = bd$, under Assumption *(A7)* is given by

$$\sqrt{nh}\left[\hat{\gamma}_\theta(bd) - \gamma_\theta(bd) - B_{\gamma,c}(x)\right] \to N\left\{0, \sigma_{\gamma,b}^2\right\},$$

where the asymptotic bias is

$$B_{\gamma,b} = h^2\beta_2(b)\theta^{-1}\left[l_1^{0,2}(q_\theta^c(0+)|0+) - q_\theta^c(0+)Z^{0,2}(q_\theta^c(0+)|0+)\right]/[2\beta_1(b)]$$

and the asymptotic variance is

$$\sigma_{\gamma,b}^2 = \frac{\beta_0(0)}{\theta\beta_1^2(b)g(0+)}[\theta^{-1}l_2(q_\theta^c(0+)|0+) - \theta\gamma_\theta^2(0+)$$

$$+ (1-\theta)q_\theta^c(0+)\{q_\theta^c(0+) - 2\gamma_\theta(0+)\}].$$

4.4 Generalized Extreme Value

Instead of focusing on the extremes we focus on exceedances of the measurement over some high threshold and the time at which the exceedances occur. Denote the rate of change of the price of a utility measured in a fixed time interval by r_t. Consider the order statistics of a collection of n rates of changes $\{r_1, \dots, r_n\}$ with $r_{(1)}$ as the minimum and $r_{(n)}$ as the maximum. For a long investment position, loss occurs when the rates of change are negative. Since the theory applies to both maximum and minimum returns, if we are interested in negative rates of change, we have

$$r_{(n)} = \max_{1\le j\le n}(r_j), \text{ and } r_{(1)} = -\max_{1\le j\le n}\{-r_j\} = -r_t$$

The treatment of negative tail is completely analogous to the positive tail only that the scale of the most extreme price movements is different. While the participant large industrial consumers are more concerned about unexpected price increases, others like producers and dealers are more interested in monitoring price drops than large price increases. Assume that the rates of change r_t are serially independent with a common cumulative distribution $F(x)$ and that the range of r_t is $[m, u]$. For log rates, we have $m = -\infty$ and $u = \infty$. In practice the cumulative density function of r_t is unknown, however as n

increases to infinity the cumulative density function becomes degenerate, namely $F_{n,n}(x) \to 0$ if $x < u$ and $F_{n,n}(x) \to 1$ if $x \geq u$ as n goes to infinity. Extreme value theory is concerned with finding two sequences $\{\beta_n\}$ and $\{\alpha_n\}$, where $\alpha_n > 0$, such that the distribution of $r_n \equiv \left(r_{(n)} - \beta_n\right)/\alpha_n$ converges to a degenerate distribution as n goes to infinity. β_n, is the location series and α_n is a series of scaling factors. Under the independent assumption the limiting distribution of the normalized maximum (r_n) is given by

$$F_*(x) = \begin{cases} exp[-(1+\xi x)]^{-\frac{1}{\xi}} & if\ \xi \neq 0 \\ exp[-exp(-x)] & if\ \xi = 0 \end{cases} \qquad (4.4.0.1)$$

Where ξ is the shape parameter that governs the tail behavior of the limiting distribution and $\alpha = \frac{1}{\xi}$ is called the tail index of the distribution. The limiting distribution (4.4.0.1) is the generalized extreme value (GEV) distribution of Jenkinson (1955) for the maximum. It encompasses the three types of limiting distribution of Gnedenko (1943); when $\xi = 0$ we have the $Gumbel$ family which is a thin tailed distribution like the lognormal and the normal. When $\xi > 0$, we have the $Fr\acute{e}chet$ family and when $\xi < 0$, we have the $Weibull$ family. In risk management we are interested with the $Fr\acute{e}chet$ family, with Probability distribution function

$$f_*(x) = \begin{cases} (1+\xi x)^{-\frac{1}{\xi}-1} exp\left[-(1+\xi x)^{-1/\xi}\right] & ; x \neq 0 \\ exp[-x - exp(-x)] & ; x = 0 \end{cases} \qquad (4.4.0.2)$$

Bermam (1964), showed that the same form of the limiting extreme value distribution hold for stationary normal sequences provided that the autocorrelation function of r_t is squared summable that is ($\sum_{i=1}^{n} \rho_i^2 < \infty$), where ρ_i is the $lag\ i$- autocorrelation function of r_t.

The shape parameter ξ can be estimated nonparametrically using the Hill estimator (ξ^h) or the Pickands' estimator (ξ^p) defined as follows. Let k be a positive integer then

$$\xi^h(\kappa) = \frac{1}{\kappa}\sum_{i=1}^{\kappa}\left[ln\left(r_{(N-i+1)}\right) - ln\left(r_{(N-\kappa)}\right)\right] \qquad (4.4.0.3)$$

$$\xi^p(\kappa) = \frac{1}{ln(2)} ln\left(\frac{r_{(N-\kappa+1)} - r_{(N-2\kappa+1)}}{r_{(N-2\kappa+1)} - r_{(N-4\kappa+1)}}\right); \quad \kappa \leq \frac{N}{4} \qquad (4.4.0.4)$$

where the argument (k) is used to emphasize that the estimator depends on k. The choice of k differs between the Hill and the Pickands estimators. The Hill estimator is applicable to the Fré$chet$ distribution only, but when applied it is more efficient than the Pickands. Goldie and Smith (1987) show that $\sqrt{\kappa}[\xi^h(\kappa) - \xi]$ is asymptotically normal with mean zero and variance ξ^2. In practice, one may plot the Hill estimator $\xi^h(\kappa)$ against κ and find a proper κ such that the estimator appears stable. The estimated tail index $\alpha = \frac{1}{\xi^h(\kappa)}$ can then be used to obtain extreme quantiles of the rate of change data.

The Hill estimator for α is given by

$$\hat{\alpha}^h(\kappa) = \frac{1}{\hat{\xi}^h}(\kappa)$$

If F is in the domain of attraction of (4.4.0.1) a GEV distribution, then $\xi^h(\kappa)$ converges in probability to ξ as $k \to \infty$ and $\frac{k}{n} \to 0$ and that $\hat{\xi}^h(k)$ is asymptotically normally distributed with asymptotic variance

$$avar\left(\hat{\alpha}^h(\kappa)\right) = \frac{\alpha^2}{\kappa}$$

Suppose that the loss distribution F is such that $1 - F(x) = x^{-\alpha}L(x)$, with $\alpha = \frac{1}{\xi} > 0$ where $L(x)$ is a slowly varying function. Let $x > r_{(\kappa+1)}$, where $r_{(\kappa+1)}$ is a higher order statistics, then the Hill estimator of $F(x)$ is given by

$$\hat{F}^h(x) = 1 - \frac{\kappa}{n}\left(\frac{x}{r_{(\kappa+1)}}\right)^{-\hat{\alpha}^{h(\kappa)}} \tag{4.4.0.5}$$

Inverting the Hill tail estimator (4.4.0.5) gives the Hill quantile estimator

$$\hat{x}_{\theta,\kappa}^{Hill} = r_{\kappa+1} - r_{\kappa+1}\left(\frac{n}{\kappa}(1-\theta)^{-\hat{\xi}^{Hill}(\kappa)} - 1\right) \tag{4.4.0.6}$$

where $\theta > 1 - \frac{\kappa}{n}$, the Hill quantile estimator (4.4.0.6) is very similar to the MLE GPD quantile estimator.

4.4.1 The Extremal Index

So far our discussions of extremal values are based on the assumption that the data are *i.i.d* random variables. However in reality extremal events tend to occur in clusters because of the serial dependence in the data. We extend the theory and applications of extreme values to cases in which the data form a strictly stationary time series. The basic concept of the extension is that extremal index allows one to characterize the relationship between the dependence structure of the data and their extremal behavior.

Let $r_1, r_2, \ldots$ be a strictly stationary sequence of random variables with marginal distribution function $F(r)$. Consider the case of n observations $\{r_i | i = 1, \ldots, n\}$. Let $r_{(n)}$ be the maximum of the data, that is $r_{(n)} = max\{r_i\}$. We seek the limiting distribution of $(r_{(n)} - \beta_n)/\alpha_n$ for suitably chosen normalizing constants $\alpha_n > 0$ and β_n when $\{r_i\}$ are serially dependent.

Suppose that the serial dependence of the stationary series r_i decays quickly so that r_i and r_{i+l} are essentially independent when l is sufficiently large, in other words, assume that the long range dependence of r_i vanishes quickly. We divide the data into disjoint blocks of size b. Specifically let $g = [n/b]$ be the largest integer less than or equal to n/b. The ith block of the data is then $\{r_j | j = (i-1)*b + 1, \ldots, i*b\}$ where it is understood that the $(g+1)\,th$ block may contain less than b observations. Let $r_{b,i}$, be the maximum of the i^{th} block that is $r_{b,i} = max\{r_j | j = (i-1)*b + 1, \ldots, i*b\}$. The collection of block maxima is $\{r_{b,i} | i = 1, \ldots . g + 1\}$. From the definitions, it is easy to see that

$$r_{(n)} = \max_{i=1,\ldots,g+1} r_{b,i} \qquad (4.4.1.1)$$

that is, the sample maximum is also the maximum of the block maxima. If the block size b is sufficiently large and the block maximum $r_{b,i}$ does not occur near the end of the ith block, then $r_{b,i}$ and $r_{b,i+1}$ are sufficiently far apart and essentially independent under the assumption of weak long range dependence in $\{r_i\}$. Consequently $\{r_{b,i} | i = 1, \dots . g + 1\}$ can be regarded as a sample of identically and independent distributed variables and the limiting distribution of its maximum, which is $r_{(n)}$, should be the extreme value distribution. The proper condition needed for the maximum $r_{(n)}$ of a strictly stationary time series to have the extreme value limiting distribution is obtained in Leadbetter (1974) and is known as the $D(\eta_n)$ condition.

4.4.2 The limiting distribution of a stationary time series

Consider the sample $r_1, r_2, \dots, r_n$. To place limits on the long range dependence of $\{r_i\}$, let η_n be a sequence of thresholds increasing at a rate for which the expected number of exceedances of r_i over η_n remains bounded. Mathematically, this is to say that

$$\lim sup\, n[1 - F(\eta_n)] < \infty$$

where $F(\cdot)$ is the marginal cumulative distribution of r_i. For any positive integers p and q suppose that i_v $(v = 1, \dots, p)$ and j_t $(t = 1, \dots, q)$ are arbitrary integers satisfying

$1 \le i_1 < i_2 < \cdots < i_p < j_1 < \cdots < j_q \le n$, where $j_1 - i_p \ge l_n$, and l_n is a function of the sample size n such that $l_n/n \to 0$ as $n \to \infty$. Let $C_1 = \{i_1, i_2, \dots, i_p\}$ and $C_2 = \{j_1, j_2, \dots, j_q\}$ be two sets of time indices. From the prior condition, elements in C_1 and C_2 are separated by at least l_n time periods.

The condition $D(\eta_n)$ is satisfied if

$$\left| P\left(\max_{i\in C_1\cup C_2} r_i \le \eta_n\right) - P\left(\max_{i\in C_1} r_i \le \eta_n\right) P\left(\max_{i\in C_2} r_i \le \eta_n\right)\right| \le \delta_n . l_n \qquad (4.4.2.1)$$

where $\delta_n . l_n \to 0$ as $n \to \infty$. This condition says that any two events of the form $\{max_{i\in C_1} r_i \le \eta_n\}$ and $\{max_{i\in C_2} r_i \le u\eta_n\}$ can become asymptotically independent as the sample size n increases when the index subsets C_1 and C_2 of $\{1,2,\dots,n\}$ are separated by a distance l_n which satisfies $l_n/n \to 0$ as $n \to \infty$, which is a weak condition.

Lemma 4.4.2:

(i). The $\delta_{n,l}$ appearing in $D(\eta_n)$ may be taken to be non- increasing in l for each fixed n.

(ii) For such $\delta_{n,l}$ taken non increasing in l for each fixed n, the condition $\delta_{n,}l_n \to 0$ as $n \to \infty$, $l_n = o(n)$, may be written as

$$\alpha_{n,[n\lambda]} \to 0 \;\; for\; each \;\; \lambda > 0 \qquad (4.4.2.2)$$

Proof For (i), we simply note that $\delta_{n,l}$ may be replaced by the maximum of the left hand side of (4.4.2.1) over all allowed choices of $i's$ and $j's$ to obtain a possible smaller $\delta_{n,l}$ which is non increasing in l for each n and still satisfies $\delta_{n,l} \to 0$ as $n \to \infty$.

For (ii), it is trivially seen that if $\delta_{n,l} \to 0$ for some $l_n = o(n)$ then (4.4.2.2) holds. The converse may be shown by noting that (4.4.2.2) implies the existence of an increasing sequence of constants m_k such that $\delta_{n,[n/k]} < k^{-1}$ for $n \ge m_k$. If k_n is defined by $k_n = a$ for $m_a \le n \le m_{a+1}$, $a \ge 1$, then $m_{k_n} \le n$ so that $\alpha_{n,[n/k]} < k^{-1} \to 0$, and the sequence $\{l_n\}$ may be taken to be $\{[n/k_n]\}$ ■

Strong mixing implies D, which in turn implies $D(\eta_n)$ for any sequence $\{\eta_n\}$. Also $D(\eta_n)$ is satisfied for appropriately chosen $\{\eta_n\}$ by stationarity normal sequences under weak conditions, whereas strong mixing need not be.

Theorem 4.5 (i) Suppose that $\{r_i|i = 1, \ldots, n\}$ is a strictly stationary time series for which there exist sequences of constants $\alpha_n > 0$ and β_n and a nondegenerate distribution function $F_*(\cdot)$ such that

$$P\left[\frac{r_{(n)} - \beta_n}{\alpha_n} \leq r\right] \underset{d}{\rightarrow} F_*(r) = F_*^{\theta}(r), \quad n \rightarrow \infty \qquad (4.4.2.2)$$

(ii) The min-stable distributions are chosen given in (i) above.

PROOF of Theorem 4.5 (i) Suppose that (4.4.2.2) holds so that writing

$$M_n' = max(-\xi_1, -\xi_2, \ldots, -\xi_n) = -r_{(n)}$$

$$P\{\alpha_n(M_n' + \beta_n) < r\} = 1 - P\{\alpha_n(M_n' + \beta_n) \geq r\}$$

$$= 1 - P\{\alpha_n(m_n - \beta_n) \leq -r\}$$

$$\rightarrow 1 - F(-r) = G(r),$$

say, where convergence occurs at all points r of continuity of G. But for such r and $\varepsilon > 0$ such that G is also continuous at $r + \varepsilon$, since

$$P\{\alpha_n(M_n' + \beta_n) < r\} \leq P\{\alpha_n(M_n' + \beta_n) \leq r\} \leq P\{\alpha_n(M_n' + \beta_n) < r + \varepsilon\}$$

we have, letting $n \rightarrow \infty$ and then $\varepsilon \rightarrow 0$,

$$P\{\alpha_n(M_n' + \beta_n) \leq r\} \rightarrow G(r)$$

so that G is one of the three maximal extreme value distribution functions. Since

$$F(r) = 1 - G(-r)$$

the three forms listed above follow from the three possible forms for G given by the Generalized extreme value Theorem.

(ii) If F is min-stable with distribution function F such that, for each $n = 2,3,\ldots,$ $\left(1 - F(a_n r + b_n)^n = 1 - F(r)\right)$ holds, then the distribution function

$$G(x) = \lim_{\varepsilon \downarrow 0}[1 - F(-r - \varepsilon)] = 1 - F(-r -)$$

satisfies $1 - F(-r) \leq G(x) \leq 1 - F(-r - \varepsilon)$ for all $\varepsilon > 0$ and hence

$$G^n(a_n r - b_n) \leq \left(1 - F(-a_n(r + \varepsilon) + b_n)\right)^n = 1 - F(-r - \varepsilon) \leq G(r + \varepsilon)$$

and

$$G^n(a_n r - b_n) \geq \left(1 - F(-a_n r + b_n)\right)^n = 1 - F(-r) \geq G(r - \varepsilon)$$

for any $\varepsilon > 0$. Since $G(r)$ and $G^n(a_n r - b_n)$ are right continous it follows that $G^n(a_n r - b_n) = G(r)$ so that G is max-stable and, by (4.4.0.1) it is one of the three extreme value distributions for maxima. This proves part (ii). ■

The constant $\theta \in (0,1]$ in (4.4.2.2) is called the extremal index which plays an important role in determining the limiting distribution $F_*(r)$ for the maximum of a strictly stationary time series, $\underset{d}{\rightarrow}$ denotes convergence in distribution. If the condition $D(\eta_n)$ holds with $\eta_n = \alpha_n r + \beta_n$ for each r such that, $F_*(r) > 0$, then $F_*(r)$ is an extreme value distribution function.

For the case $\xi \neq 0$, from the result of (4.4.0.1) $F_*(r)$ is a generalized extreme value distribution and assumes the form

$$F_*(r) = exp\left[-\left(1 + \xi\frac{r-\beta}{\alpha}\right)^{-1/\xi}\right] \quad (4.4.2.3)$$

where $\xi \neq 0$ and $1 + \xi\,(r-\beta)/\alpha > 0$. Based on Theorem 4.5 we have

$$F_*(r) = F_*^{\theta}(r) = exp\left[-\theta\left(1 + \xi\frac{r-\beta}{\alpha}\right)^{-1/\xi}\right]$$

$$= exp\left[-\left(\frac{1}{\theta^{\xi}} + \xi\frac{r-\beta}{\alpha\theta^{\xi}}\right)^{-1/\xi}\right] = exp\left[-\left(\xi\frac{\alpha/\xi + r-\beta}{\alpha\theta^{\xi}}\right)^{-1/\xi}\right]$$

$$= exp\left[-\left(1 + \xi\frac{r-\beta+\alpha/\xi-\alpha\,\theta^{\xi}/\xi}{\alpha\theta^{\xi}}\right)^{-1/\xi}\right]$$

$$= exp\left[-\left(1 + \xi\frac{r-\left[\beta-\frac{\alpha}{\xi}(1-\theta^{\xi})\right]}{\alpha\theta^{\xi}}\right)^{-1/\xi}\right]$$

$$= exp\left[-\left(1 + \xi*\frac{r-\beta*}{\alpha*}\right)^{-1/\xi*}\right] \quad (4.4.2.4)$$

Therefore for a stationary time series $\{r_i\}$ satisfying the $D(\eta_n)$ condition, the limiting distribution of the sample maximum is the generalized extreme value distribution with the shape parameter ξ which is the same as that of the *i.i.d* sequences. On the other hand the location and scale parameters are affected by extremal index θ. Result for the case $\xi = 0$ can be derived via the same approach and we have $\alpha *= \alpha$ and $\beta *= \beta + \alpha ln(\theta)$.

4.4.3 Value at risk for a stationary time series

The relationship $F_*(r)$ of the maximum of a stationary time series can be used to calculate VaR of an institution position when the associated log rates of change form a stationary time series. Specifically from $P\left(r_{(n)} \leq \eta_n\right) \approx [F(r)]^{n\theta}$, the $(1-p)th$ quantile of $F(r)$ is the $(1-p)^{n\theta}$ *th* quantile of the limiting extreme value distribution of $r_{(n)}$.

Consequently, the VaR based on the extreme value theory becomes

$$VaR = \begin{cases} \beta_n - \frac{\alpha_n}{\xi_n}\{1-[-n\theta\, ln(1-p)\,]^{-\xi_n}\} & if \quad \xi_n \neq 0 \\ \beta_n - \alpha_n ln[-n\theta\, ln(1-p)] & if \quad \xi_n = 0 \end{cases} \qquad (4.4.3.1)$$

where n, is the length of the subperiod. From the formulae, we risk underestimating VaR if the extremal index is overlooked.

4.5 Peak over Threshold

Let us denote the daily rate of change by r_t. Let η be a specified high threshold and suppose that the $i-th$ exceedance at day t_i is $(r_{t_i} \geq \eta)$. We shall focus on the data $(t_i, r_{ti}-\eta)$ where $r_{ti}-\eta$ is the exceedance over the threshold η. The occurrence times $\{t_i\}$ provides useful information about the intensity of the occurrence of important “rare events”. A cluster of t_i indicates a period of large declines, the exceeding amount (excesses) provides the actual quantity of interest. Different choices of the threshold η leads to different estimates of the tail index $\frac{1}{\xi}$ and is based on risk tolerance. Since different institutions and investors have different risk tolerance, the choice of η is a statistical problem as well as a financial one. According to Tsay (2010) η = 2.5% may fare well for a stable rate of change series and η may be as high as 10% for a volatile series.

Let the occurrence of the event $\{r_t \leq \eta\}$ follow a point process like a Poisson process. If the intensity parameter λ of the process is time invariant then the process is homogenous. Further, consider the conditional distribution of $r_t \leq x + \eta$ given that $r_t > \eta$

$$Pr(r_t \leq x + \eta | r_t > \eta) = \frac{Pr(\eta \leq r_t \leq x + \eta)}{Pr(r_t > \eta)} \qquad = \frac{Pr(r_t \leq x + \eta) - Pr(r_t \leq \eta)}{1 - Pr(r_t \leq \eta)} \quad (4.5.0.1)$$

Using (4.4.0.1) and the approximation $e^{-y} \approx 1 - y$ we obtain

$$Pr(r_t \leq x + \eta | r_t > \eta) = \frac{F_*(x + \eta) - F_X(\eta)}{1 - F_*(y)}$$

$$= \frac{exp\left\{-\left[1 + \frac{\xi(x + \eta - \beta)}{\alpha}\right]^{-\frac{1}{\xi}}\right\} - exp\left\{-\left[1 + \frac{\xi(\eta - \beta)}{\alpha}\right]^{-\frac{1}{\xi}}\right\}}{1 - exp\left\{-\left[1 + \frac{\xi(\eta - \beta)}{\alpha}\right]^{-\frac{1}{\xi}}\right\}}$$

$$\approx 1 - \left[1 + \frac{\xi x}{\alpha + \xi(\eta - \beta)}\right]^{-\frac{1}{\xi}} \qquad (4.5.0.2)$$

If we substitute $\alpha + \xi(\eta - \beta)$ with $\psi(\eta)$ we have a cumulative distribution function

$$G_{\xi,\psi(\eta)}(x) = \begin{cases} 1 - \left[1 + \frac{\xi x}{\psi(\eta)}\right]^{-\frac{1}{\xi}} & for\ \xi \neq 0 \\ 1 - exp[-x/\psi(\eta)] & for\ \xi = 0 \end{cases} \qquad (4.5.0.3)$$

Which is a generalized Pareto distribution(GPD), when $\psi(\eta) > 0$ and $x \geq 0$.

Large observations which exceed a high threshold can be approximated reasonably well by the Generalized Pareto Distribution. The parameters can be estimated consistently if the threshold exceedances are independent regardless of the true underlying distribution. In general given a high threshold η and a random variable $Y = r_t$, the probability of r_t exceeding η at most by x is given by

$$F_\eta(x) = P[r_t - \eta \leq x | r_t > \eta] = \frac{F(x+\eta) - F(\eta)}{1 - F(\eta)} \quad (4.5.0.4)$$

For a distribution function F, it is possible to find a positive function $\beta(\eta)$ such that

$$\lim_{\to r_{t0}} \sup_{0 \leq x \leq r_{t0}-} \left| F_\eta(x) - G_{\xi,\beta}(x) \right| = 0 \quad (4.5.0.5)$$

with r_{t0} corresponding to the right end point of F. Rearranging (4.5.0.4) and using

$F_\eta(\cdot) \approx G_{\xi,\beta}(\cdot)$, it holds that

$$1 - F(\eta + x) \approx [1 - F(\eta)]\left[1 - G_{\xi,\beta}(x)\right] \quad (4.5.0.6)$$

$1 - G_{\xi,\beta}(x)$ can be obtained by estimating the GPD parameters by maximum likelihood. We estimate $1 - F(\eta) = Z(\eta)$ by the use of the empirical distribution function

$$Z\widehat{(\eta + x)} = \frac{N_\eta}{n}\left(1 + \hat{\xi}\left(\frac{x}{\hat{\beta}}\right)\right)^{-\frac{1}{\hat{\xi}}} \quad (4.5.0.7)$$

Where N_η denotes the number of exceedances over threshold η.

Employing a change of variables $y = \eta + x$ and fixing the distribution value at the probability of interest $F(y) = \theta$, we obtain the quantile estimator $\hat{q}_\theta$ by inverting (4.5.0.7) to get

$$1-\theta = \frac{N_\eta}{n}\left(1+\hat{\xi}\left(\frac{y-\eta}{\hat{\beta}}\right)\right)^{-\frac{1}{\hat{\xi}}}$$

$$\Leftrightarrow \hat{q}_\theta = \eta + \left[\left((1-\theta)\frac{n}{N_\eta}\right)^{-\hat{\xi}} - 1\right] \times \frac{\hat{\beta}}{\hat{\xi}} \qquad (4.5.0.8)$$

where $\hat{\xi}$ and $\hat{\beta}(\eta)$ denotes the Maximum likelihood estimates of ξ and $\beta(\eta)$ respectively. From (4.5.0.1), (4.5.0.2) and the GPD (4.5.0.3) we have

$$\frac{F(y)-F(\eta)}{1-F(\eta)} \approx G_{\eta,\psi(\eta)}(x)$$

If we estimate $F(\eta)$ by the empirical CDF then

$$\hat{F}(\eta) = \frac{n-n_\eta}{n}$$

Consequently by (4.5.0.3)

$$F(y) = F(\eta) + G(x)[1-F(\eta)]$$

$$\approx 1 - \frac{n_\eta}{n}\left[1+\frac{\xi(y-\eta)}{\psi(x)}\right]^{-\frac{1}{\xi}} \qquad (4.5.0.9)$$

For a small upper tail probability p, let $\theta = 1-p$, we estimate the $\theta - th$ quantile Value at Risk of $F(y)$ by

$$VaR_\theta = \eta - \frac{\psi(\eta)}{\hat{\xi}}\left\{1-\left[\frac{n}{N_\eta}(1-\theta)\right]^{-\xi}\right\} \qquad (4.5.1.0)$$

For a generalized Pareto distribution and using the properties of GPD,

$$E(r - VaR_\theta | r > VaR_\theta) = \frac{\psi(\eta) + \xi(VaR_\theta - \eta)}{1 - \xi}$$

provided that $0 < \xi < 1$. Consequently we have the Expected shortfall

$$ES_\theta = \frac{\xi(VaR_\theta)}{1 - \hat{\xi}} + \frac{\psi(\eta) - \xi\eta}{1 - \xi} \qquad (4.5.1.1)$$

4.6 The Poisson Point Process

Applying (4.5.0.2) and considering jointly the exceedances and exceeding times, we view the excess time and excess values as a two dimensional point process (t_i, r_{t_i}). The point process captures the stochastic volatility effects in financial time series. Assume that the baseline time interval is B which is the trading days in a year. Let t be the time interval of the data points, denote the data span by $t = 1,2, \ldots, T$ where T is the total number of data points for a given threshold η. The exceeding times over the threshold are denoted by $\{t_i, i = 1, \ldots, N_\eta\}$ and the observed rate of change at time t_m denoted by r_{tm}. We postulate that the exceeding times and the associated rate of change (t_i, r_{tm}) jointly form a two dimensional Poisson process with an intensity measure. If this process is stationary and there are no clusters asymptotically, the limiting form of the distribution of the process is a homogeneous Poisson with intensity

$$\Lambda[(B_2, B_1) \times (r, \infty)] \frac{B_2 - B_1}{B} Z(r: \xi, \alpha, \beta) \qquad (4.6.0.1)$$

Where (B_2, B_1) is the interval under consideration and

$$Z(r\colon \xi, \alpha, \beta) = \left[1 + \frac{\xi(r-\beta)}{\alpha}\right]_{+}^{-\frac{1}{\xi}}$$

$r > \eta$, $\alpha > 0$, β and ξ are parameters and $[\cdot]_{+} = max(x, 0)$. The extreme value parameters are assumed to change periodically according to a random change point process. This intensity measure says that the occurrence of exceeding the threshold is proportional to the length of the time interval $[B_1, B_2]$ and the probability is governed by a survival function similar to the exponent of the CDF $F_{*}(r)$ in (4.4.0.1). When $\xi = 0$ the intensity measure is taken to be the limit $\xi \to 0$, that is

$$\Lambda[(B_2, B_1) \times (r, \infty)] \frac{B_2 - B_1}{B} exp\left[\frac{-(r-\beta)}{\alpha}\right] \tag{4.6.0.2}$$

(4.6.0.1) can be written as an integral of an intensity function

$$\Lambda[(B_2, B_1) \times (r, \infty)] = \int_{B_2}^{B_1} \int_{r}^{\infty} \lambda(t, z; \xi, \alpha, \beta) dz dt \tag{4.6.0.3}$$

where the intensity function is defined as

$$\lambda(t, z; \xi, \alpha, \beta) = \frac{1}{D} g(z; \xi, \alpha, \beta)$$

and

$$g(z; \xi, \alpha, \beta) = \begin{cases} \frac{1}{\alpha}\left[1 + \frac{\xi(z-\beta)}{\alpha}\right]^{-(1-\xi)/\xi} & if \;\; \xi \neq 0 \\ \frac{1}{\alpha} exp\left[\frac{-z(z-\beta)}{\alpha}\right] & if \; \xi = 0 \end{cases} \tag{4.6.0.4}$$

Using the results of the Poisson Process, we can write down the likelihood function for the observed exceeding times and their corresponding times $\{(t_i, r_i)\}$ over the two dimensional space $[0, T] \times (\eta, \infty)$ as

$$L(\xi, \alpha \; and \; \beta) = \left[\prod_{i=1}^{N_\eta} \frac{1}{D} g\left(r_{t_i}; \xi, \alpha, \beta\right)\right] exp\left[-\frac{T}{B} Z(\eta; \xi, \alpha, \beta)\right] \quad (4.6.0.5)$$

The parameters ξ, α and β can be estimated by maximizing the logarithm of the likelihood function. Since the scale parameter α is non-negative, we use $ln(\alpha)$ in the estimation. As shown in (4.5.0.2), the two dimensional Poisson Process model which employs the intensity measure (4.6.0.1) has the same parameters as those of Extreme Value distribution in (4.4.0.1). Given the upper tail probability θ, with the $(1-\theta)th$ quantile of the rates of change r_t, the Value at Risk based Poisson point process becomes

$$VaR = \begin{cases} \beta - \frac{\alpha}{\xi}\left\{1 - [-Bln(1-\theta)]^{-\xi}\right\} & if \; \xi \neq 0 \\ \beta - \alpha ln[-Bln(1-\theta)] & if \; \xi = 0 \end{cases} \quad (4.6.0.6)$$

4.6.1 Use of explanatory variables

Suppose that $\boldsymbol{x}_t = (x_{1t}, \dots, x_{vt})'$ is a vector of v explanatory variables that are available prior to time t, like the volatility σ^2 of r_t. We postulate that the three parameters ξ, α and β are time varying and are linear functions of the explanatory variables. Specifically when the explanatory variables are available we assume that

$$\xi_t = \gamma_0 + \gamma_1 x_{1t} + \cdots, +\gamma_v x_{vt} \equiv \gamma_0 + \gamma' x_t$$

$$ln\,(\alpha_t) = \delta_0 + \delta_1 x_{1t} + \cdots, +\delta_v x_{vt} \equiv \delta_0 + \delta' x_t$$

$$\beta_t = \vartheta_0 + \vartheta_1 x_{1t} + \cdots, +\vartheta_v x_{vt} \equiv \vartheta_0 + \vartheta' x_t$$

when the three parameters of the extreme value distribution are time varying, we have an inhomogeneous Poisson Process. The intensity measure becomes

$$\Lambda[(B_2, B_1) \times (r, \infty)] = \frac{B_2 - B_1}{B}\left[1 + \frac{\xi_t(r - \beta_t)}{\alpha_t}\right]_+^{-1/\xi} \quad if \;\; r > \eta \qquad (4.6.1.1)$$

The likelihood function of the exceeding times and rates $\{(t_i, r_i)\}$ becomes

$$L = \left[\prod_{i=1}^{N_\eta} \frac{1}{B} g\left(r_{t_i};, \xi_{ti}, \alpha_{ti}, \beta_{ti}\right)\right] exp\left[-\frac{1}{B}\int_0^N Z(\lambda, \xi_t, \alpha_t, \beta_t)dt\right]$$

4.6.2 Validity Assessment

When EVT model is fitted into a sample of extreme outcomes it necessitates the validation of the method in use before it can be used in extreme event forecasting. These diagnostic tests concentrate on evaluating whether the model assumptions are satisfied in practice and the tests are complementary to each other.

Quantile - Quantile plot

This plot is an assessment of the correspondence between the estimated model distribution and the data. It is a plausibility test for modeling excesses. If the points form approximately a straight line this indicates that the model is correctly specified; de Rozario (2002). The plot consists of points

$$\left\{\left(X_{k,n}, H^{-1}_{\xi,\sigma}\left(\frac{n-k+1}{n+1}\right)\right)\middle| k = 1, \dots, n\right\} \tag{4.6.2.1}$$

where $X_{k,n}$ is the $k\ th$ order statistic. In case of a concave curvature in the Q-Q plot the GDP underestimates the empirical distribution tails and the sample data has heavy tail. Convex curvature indicates that sample data has short tail.

Mean Excess Plot

The mean excess plot consists of points

$$\left\{(u, e(u))\middle| X_{n,n} < u < X_{1,n}\right\}$$

where u is the threshold and $e(u)$ is the mean excess function which is defined as

$$e(u) = E[X - u | X > u] \tag{4.6.2.2}$$

The mean excess function describes the expected overshoot of a threshold given that it has been breached. In case of a GPD, (4.6.2.2) has the exact form

$$e(u) = \frac{\sigma + \xi u}{1 - \xi} \tag{4.6.2.3}$$

The empirical estimate of the mean excess function is given by the sample mean excess function

$$e_n(u) = \frac{\sum_{i=1}^{n} (X_i - u)_+}{\sum_{i=1}^{n} I_{\{X_i > u\}}} \tag{4.6.2.4}$$

where $I_{\{X_i > u\}}$ is an indicator function that evaluates to unity if $X_i > u$ and 0 otherwise. $(\cdot)_+$ is as in (4.5.0.3). In the mean excess plot, the mean exceedances in the data are plotted against increasing threshold value. This plot should follow a straight line with slope $\frac{\xi}{(1-\xi)}$ and intercept $\frac{\sigma}{(1-\xi)}$. Corresponding to $\xi = 0$ is a horizontal line.

CHAPTER FIVE

5.0 RESULTS AND DISCUSSION

5.1 Electricity demand distribution

To understand the underlying electricity demand distribution in Kenya, the daily peak hour electricity demand data (in mega watts) from 1^{st} January 2005 to 28^{th} February, 2011 was obtained from Kenya Power and Lighting Company. Our sample comprise of 2250 observations. Figure 1 shows the density estimation of logarithmic daily peak hour electricity demand using the adaptive kernel method of Silverman (1986). Two estimates are presented; for low demand and for high demand. From Figure 1 it is evident that parametric modelling is rejected by this data set unless the data is modelled as a finite mixture of two or more distributions.

Figure 2 shows that even after taking logarithm transformation the trend still persists. This signifies that the trend is non-deterministic and that the data used in this study shows a great deal of seasonality. This seasonality is attributed to the problem of electricity storage.

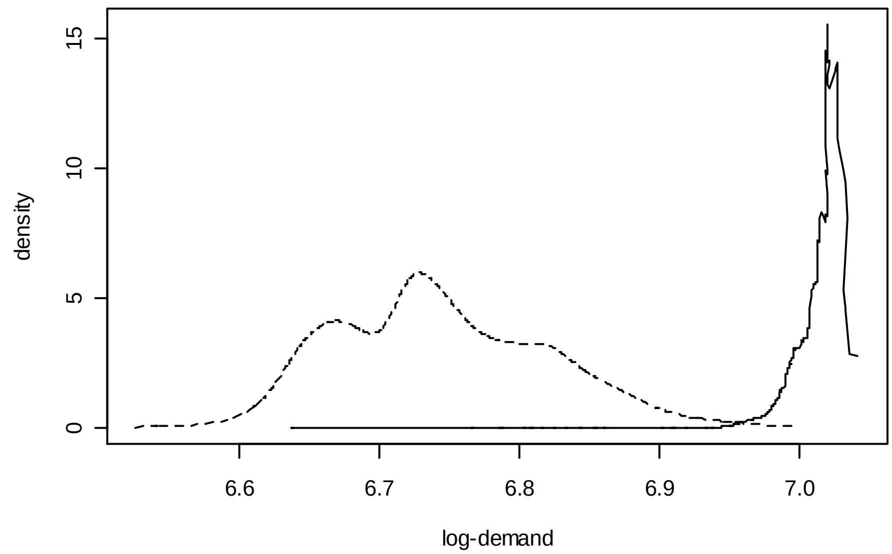

Figure 1: The kernel probability density for low (……….) and high (______) electricity demand at peak hour in megawatts.

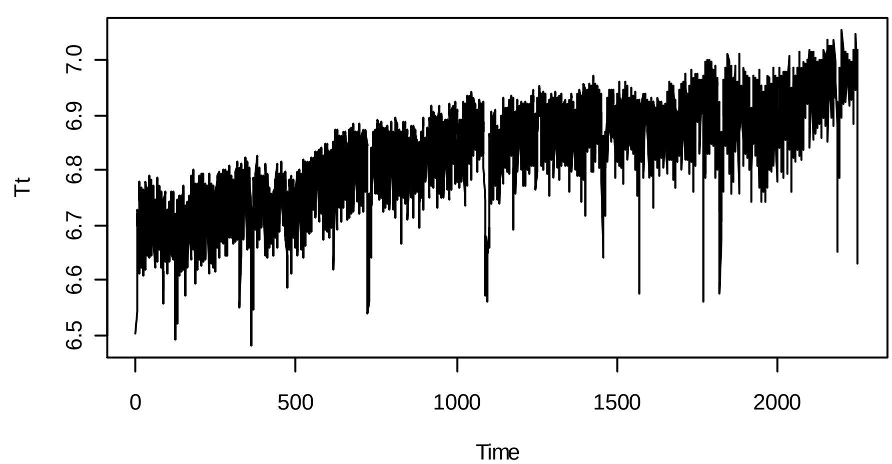

Figure 2: Time series plot for log transformed electricity data.

To introduce stationarity to the series we apply the analysis of Brockwell and Davis (1991) by using the difference operator $(1 - B)(1 - B^s)$. The first part $(1 - B)$ is the ordinary difference component ∇^d and the second part $(1 - B^s)$ is the seasonal difference component ∇_s^D. After differencing, a seasonal difference of order 7 seems appropriate with significant autocorrelation coefficient (ACF) at lags 7, 14 and 21. We obtain a new series $Y_t = (1 - B)(1 - B^7)T_t$, which does not display any apparent deviations from stationarity as shown in Figure 3.

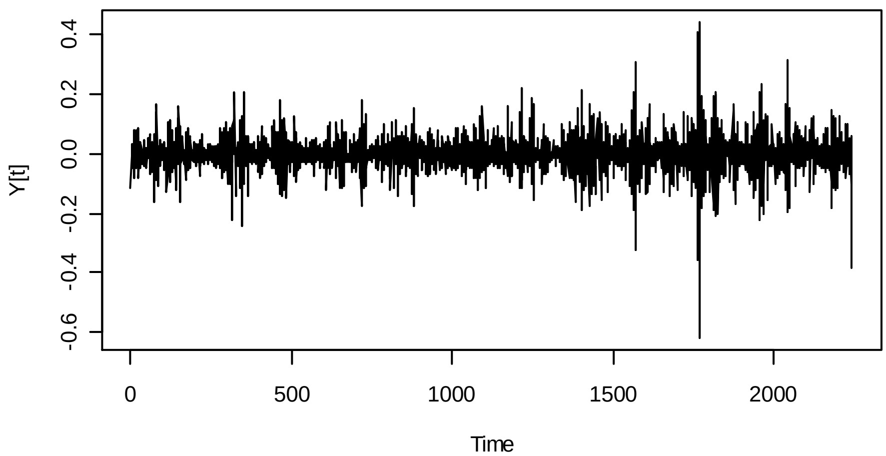

Figure 3: Time series plot of twice differenced log transformed electricity data

Residuals obtained from such detrended models, where economic time series data is regressed on time, can be interpreted as cyclical components in the context of business cycle theory. Such models can be used to estimate trend growth rates in historical context.

We apply the kernel regression estimation and local polynomial fitting methods to estimate the drift and diffusion of the daily rate of change of electricity demand at peak hour. Let Y_t denote the stationary daily rate of change. We model Y_t by assuming that it satisfies the continuous time stochastic differential equation

$$dY_t = \mu(Y_{t-1})dt + \sigma(Y_{t-1})dw_t$$

For simplicity we use $|Y_t|$ as a proxy of volatility where, $\mu(Y_{t-1})$ is the conditional mean of Y_t given Y_{t-1} that is $\mu(Y_{t-1}) = E(Y_t|Y_{t-1})$.

Figure 4 shows a local smooth estimate of $\mu(Y_{t-1})$.The estimate is approximately zero. However to better understand the estimate, Figure 5 shows this estimate on a finer scale. This estimate suggest that when $\mu(Y_{t-1})$ is positive Y_{t-1} is negative and vice versa. Thus the conditional mean is a decreasing function. Figure 6 shows the estimate of the diffusion function of the daily rate of change of electricity demand ($\hat{\sigma}(Y_{t-1})$). The plot shows that the lower the demand the higher the volatility and the higher the demand the higher the volatility. Figure 7 shows this estimate $\hat{\sigma}(Y_{t-1})$ on a finer scale.

To compare the volatility results with parametric model outputs, we find that of all seasonal first differenced parametric models, $ARIMA(1,1,2) \times (1,1,1)_7$ is the most parsimonious model. Likewise, $GARCH(1,1)$, fits well the residuals, with the 0.1, 0.5 and 0.9 quantiles shown in Figure 8, Figure 9 and Figure 10 respectively. These plots are in agreement with the nonparametric results that the more the extreme the demand is the higher the volatility.

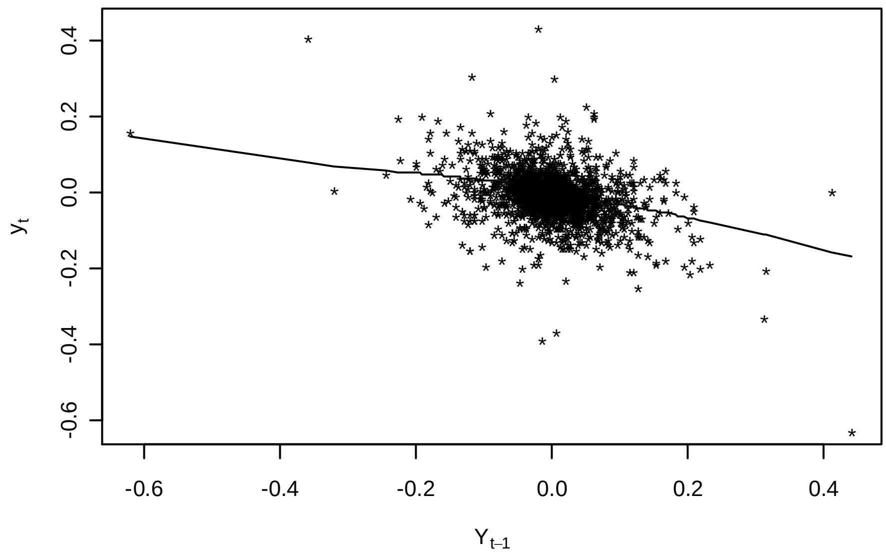

Figure 4 The smooth estimate of the drift function of the daily rate of change of electricity demand

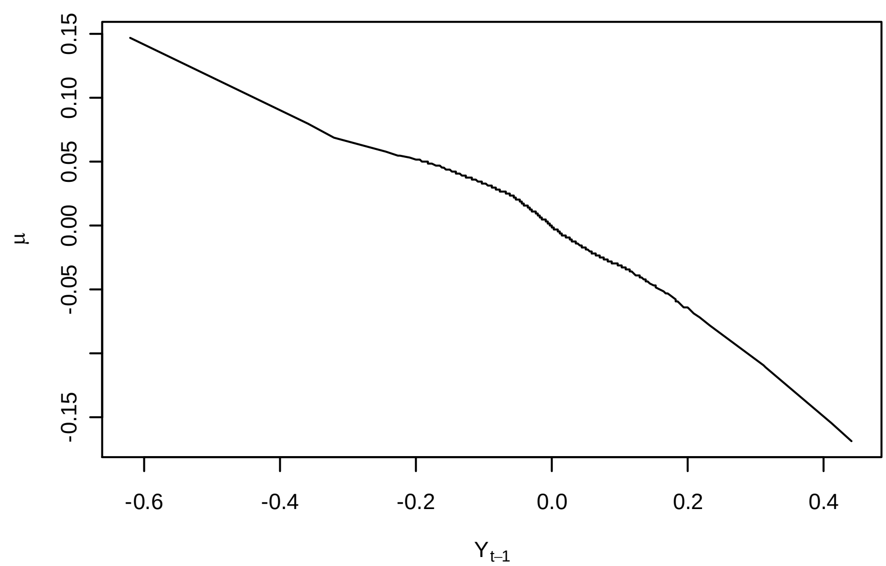

Figure 5 The estimate of the drift function of the daily rate of change of electricity demand $\hat{\mu}(Y_{t-1})$ on a finer scale.

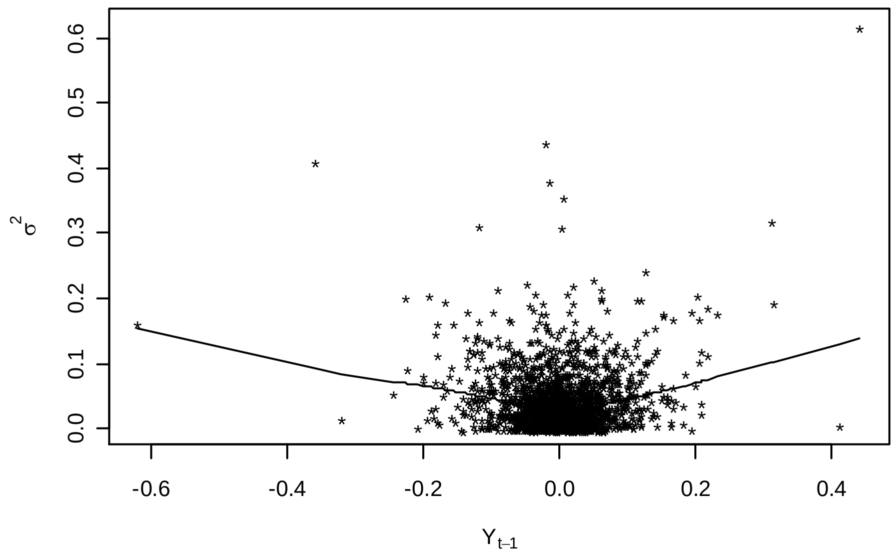

Figure 6 The estimate of volatility for the daily rate of change of electricity demand $\widehat{\boldsymbol{\sigma}}(\mathbf{Y_{t-1}})$.

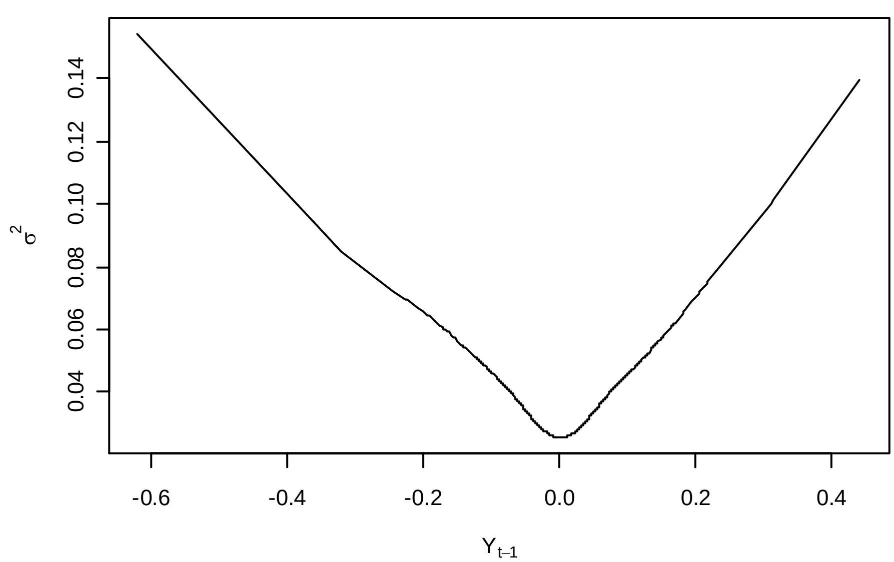

Figure 7 The estimate of volatility for the daily rate of change of electricity demand on a finer scale

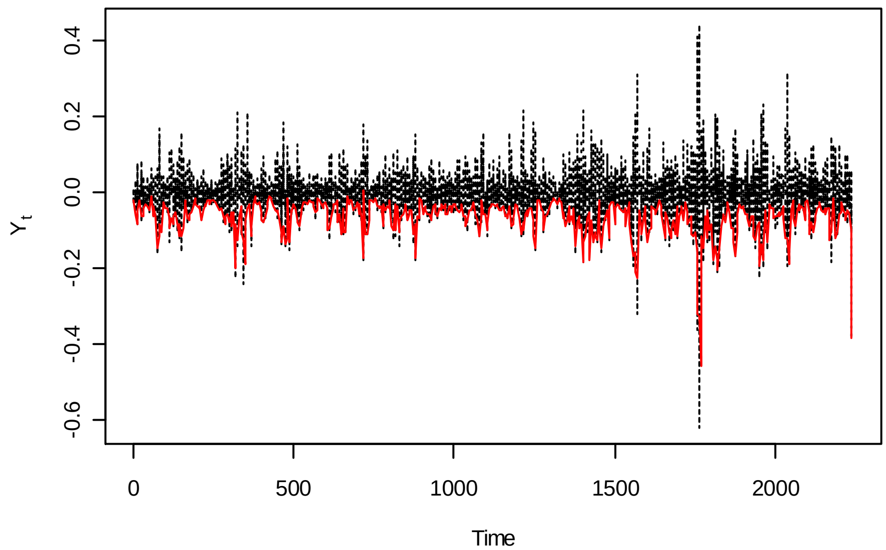

Figure 8 Residuals 0.1 quantile for the stationary series of electricity peak demand under the GARCH model

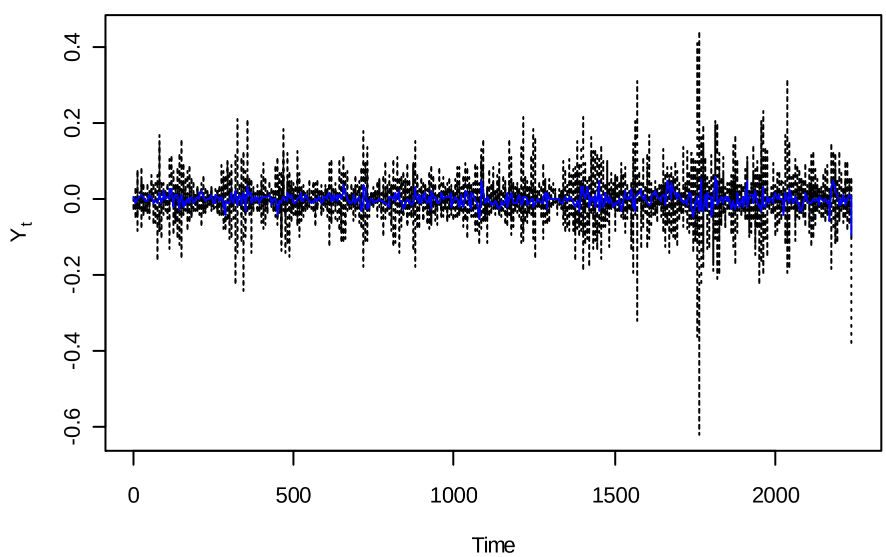

Figure 9 Residuals 0.5 quantile for the stationary series of electricity peak demand under the GARCH model

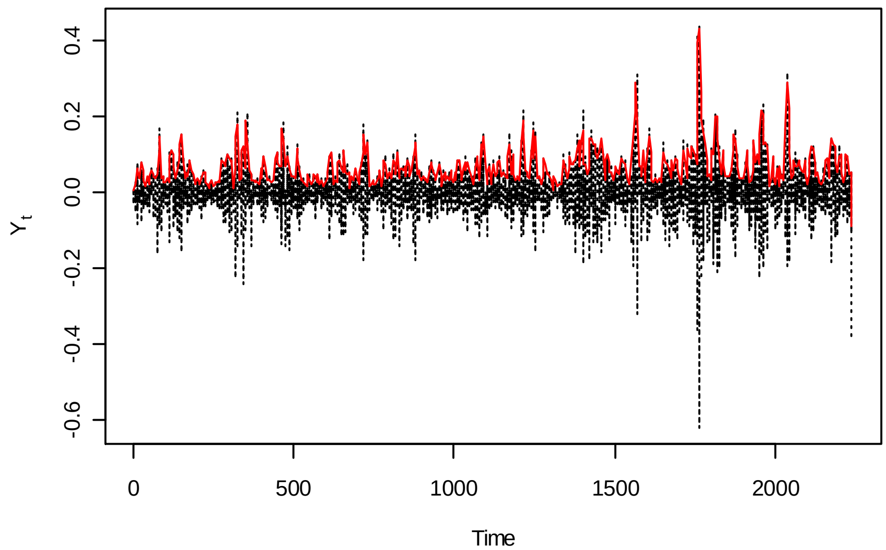

Figure 10 Residuals 0.9 quantile for the stationary series of electricity peak demand under the GARCH model

To try to answer the question if the estimated quantile regression relationships confirm to the location shift transformation that assumes that all the conditional quantiles functions have the same slope parameters, we estimate the quantile fits for electricity logarithm demand data. Figure 11 shows the estimated conditional quantiles at two opposite extremes; quantile (0.9) and quantile (0.1). From Figure 11, we reject the assumption and conclude that electricity demand data has different slopes at different quantiles.

Figure 12 shows the conditional quantile function of electricity demand obtained by inverting the estimated conditional cumulative distribution function using the t kernel estimator.

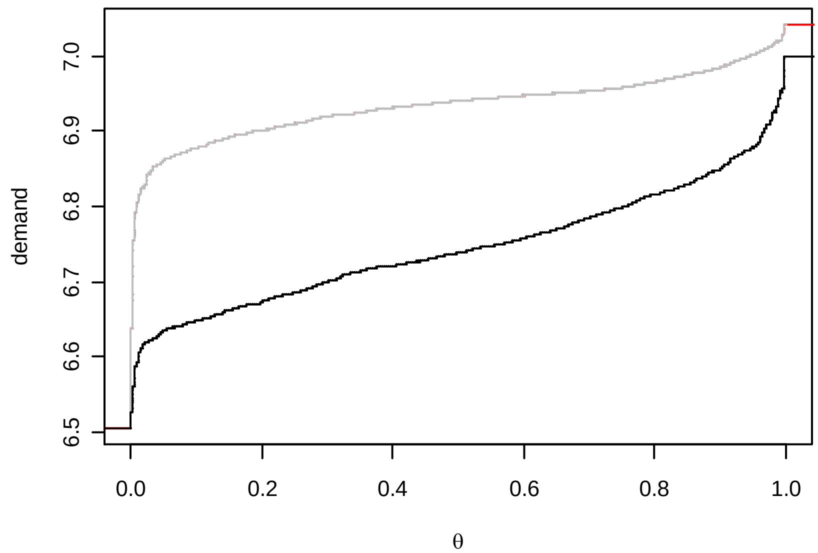

Figure 11 The estimated conditional quantiles of logarithm electricity demand at 0.9 quantile and 0.1 quantile.

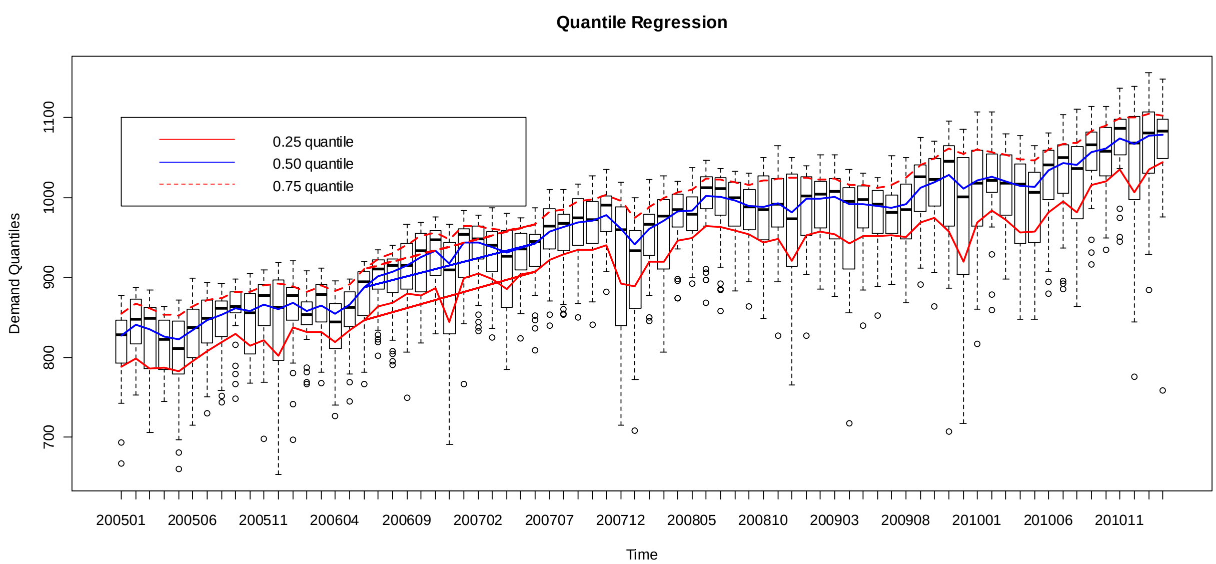

Figure 12 Conditional quantiles for electricity peak hour demand indicating three quantiles, 0.25 quantile, 0.50 quantile and 0.75 quantile.

It is apparent from Figure 12 that electricity demand distribution is heteroscedastic and poses heavy tails. To understand the evolution of electricity demand better, the logarithm daily rate of change in electricity demand is introduced as shown in Figure 13. This figure depicts clearly that there are more extreme positive rates of change than negative rates of change. This can be attributed to the gradual advancement in social economic dynamics in the economy over time. Another important feature displayed in Figure 13 is that aberrant observations tend to emerge in clusters (persistence). We see that apart from year 2005 where the extreme rates of change are evenly distributed throughout, for the other years extremes are clustered. In 2006 clustering is between the months of August and December, in 2007 clustering is between July and October, in 2008 between July and September, in 2009 persistence is between October and November and in 2010 clustering is between the months of May and August. These persistent clustering can possibly be attributed to the market mechanism effects. The intuitive interpretation is that it is difficult to forecast future rates of change unless we understand more the market mechanism.

5.2 Optimal bandwidth for electricity demand data

To determine the optimal nonparametric smoothing parameter (bandwidth) we apply CV, GCV, AICc and RCP. Our results were obtained by writing program code in the R package (Kernsmooth).

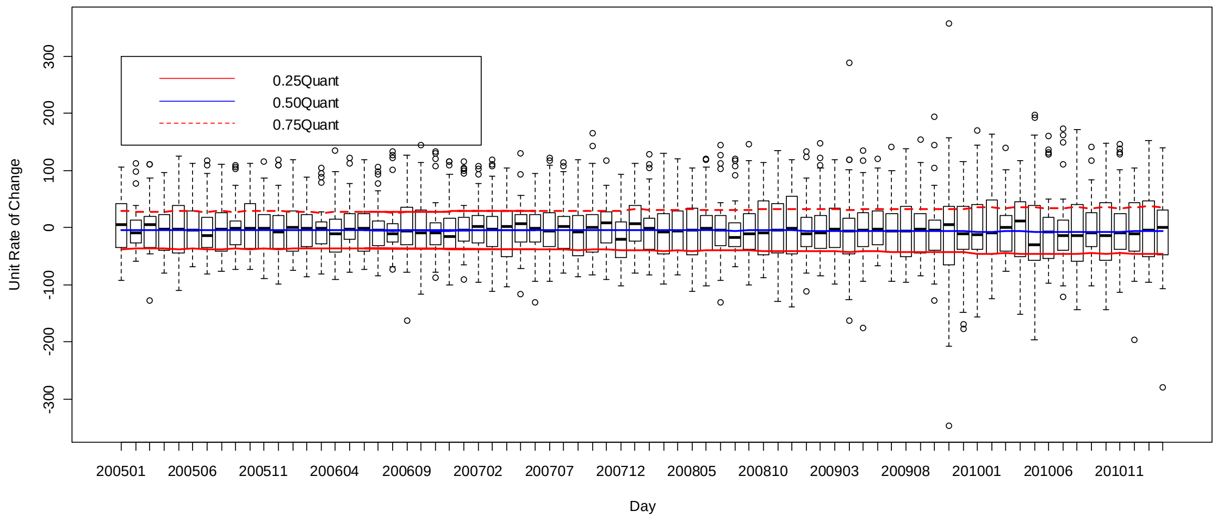

Figure 13 Daily rate of change of electricity demand at peak hour with quantiles 0.25, 0.50 and 0.75

Figure 14 shows the output of bandwidth determination of the logarithm daily rate of change of electricity demand at peak hour under Cross Validation. Figure 15 is the output under GCV while Figure 16 and Figure 17 are the bandwidth determination using AICc and RCP criterion respectively.

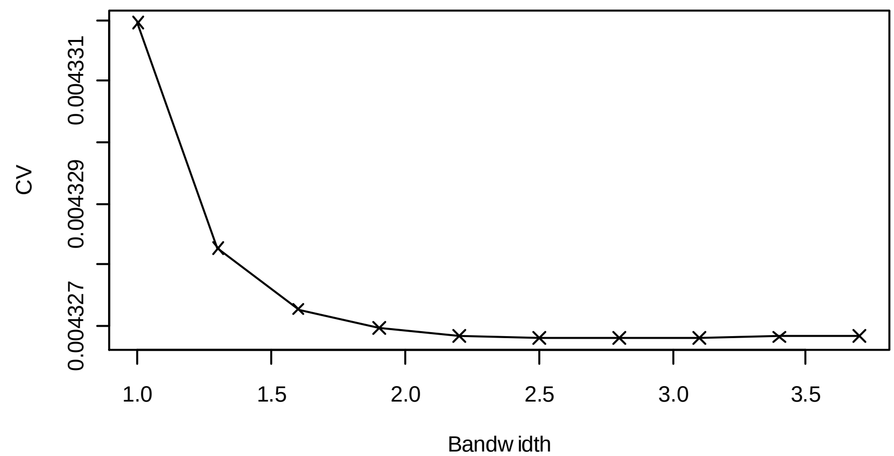

Figure 14 Bandwidth determination of logarithm rate of change of electricity demand using Cross Validation

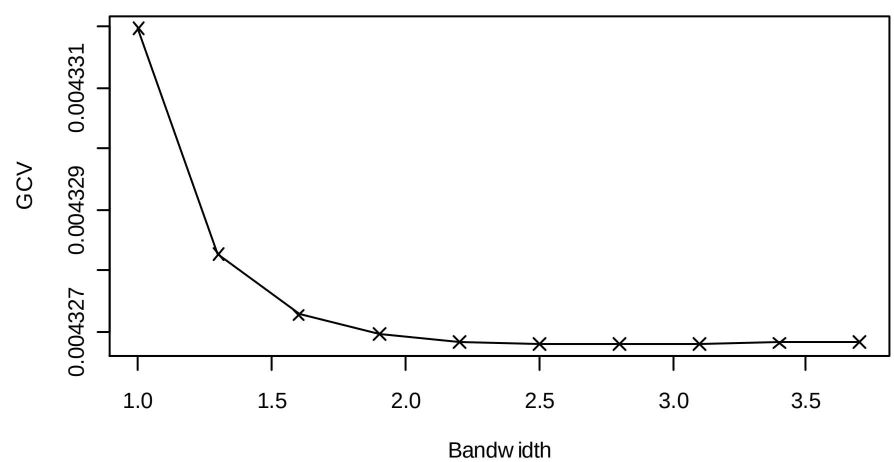

Figure 15*:* Bandwidth determination of Logarithm rate of change of electricity demand using Generalized Cross Validation

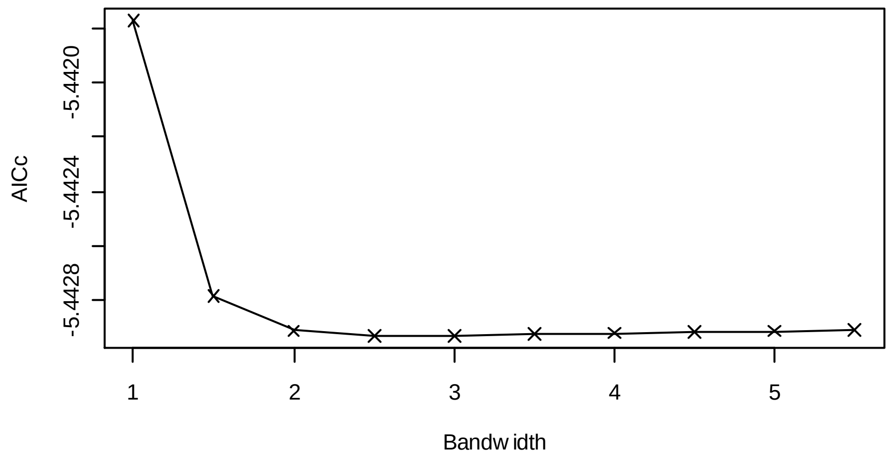

Figure 16: Bandwidth determination of logarithm rate of change of electricity demand using The Improved Akaike Information Criterion

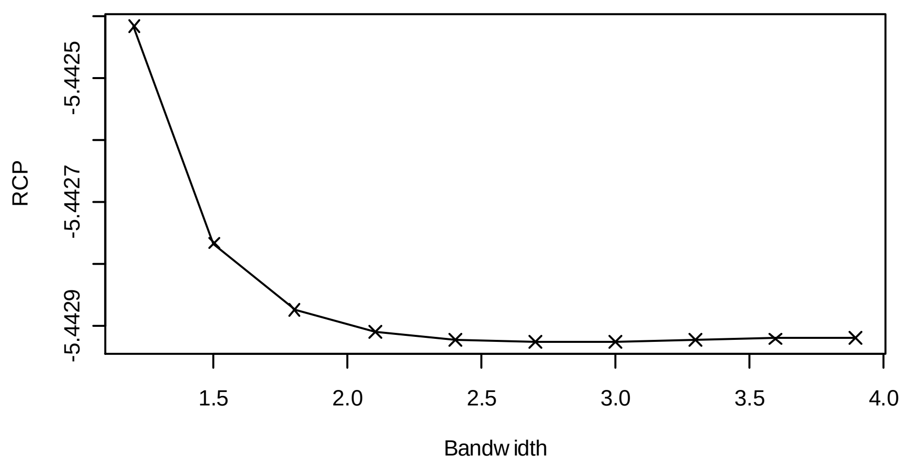

Figure 17 Bandwidth determination of logarithm rate of change of electricity demand using The T-Criterion (RCP)

It is evident from the optimal bandwidth summary Table 3 that AICc overcomes the over fitting tendency and that the cross validation types are asymptotically equivalent. Li and Racine (2004) showed that both CV and GCV performs poorly due to its large sample variance even worse for dependent data.

Table 3 Different optimal bandwidth for electricity demand data.

Criterion	Bandwidth	Minimum value
CV	2.8	0.004326
GCV	2.8	0.004327
AICc	2.5	-5.443
RCP	2.7	-5.443

5.3 Consumers response to change in fuel prices on electricity demand

We refer to the fundamental factors that affect electricity demand and price formation as categorized in Table 3

Table 3 Factors that have impact on spot electricity characteristics (Bunn and Karakatsani, 2003)

Structural Effect	Influencing factors
1. Market Mechanism	Fuel prices, demand polynomial
2. Behavioral variables	Lagged variables, price volatility
3. Time Effect	Daily, weekly, seasonal
4. Efficiency variables	Trading volume, availability of indices
5. Market structure	Margin, concentration of indices
6. Non strategic uncertainties	Demand forecast error

Consider the following model a realization of model (2.7.0.2);

$$y_i = c_0 + g_1(z_{1i}) + g_2(z_{2i}) + \cdots + g_q(z_{qi}) + u_i \qquad i = 1, \dots, n$$

where c_0 is a scalar parameter, z_{ji} is a univariate continuous variable and $g_l(\cdot)$; $l = 1, \dots, q$ are smooth functions. Let y_i be the logarithm change in electricity demand, z_{11} be the logarithm of the previous day electricity demand and z_{22} be the logarithm of the previous day price of fuel (diesel). We wish to find the impact of the explanatory variables to the rate of change of electricity demand.

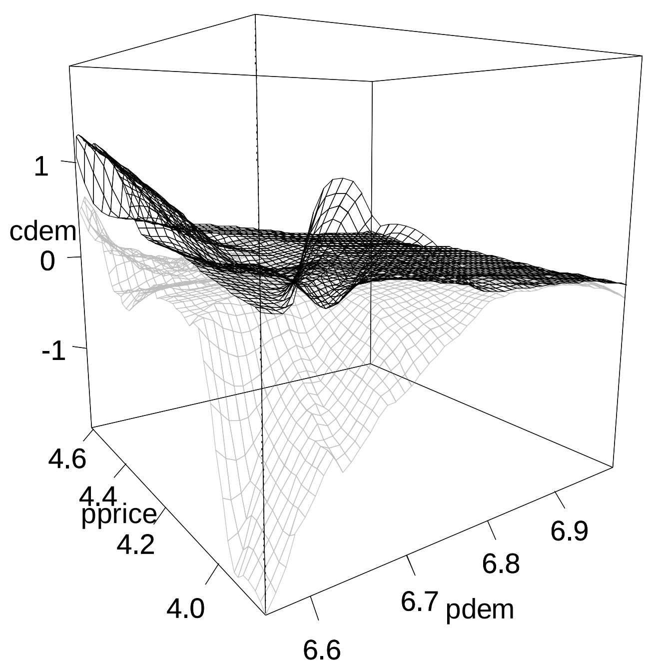

Figure 18 Kernel regression estimate of a two dimensional dependent variable on an explanatory variable: previous day demand and previous day fuel prices on the daily change in electricity demand

Figure 18 shows a kernel regression estimate of daily change in electricity demand ($cdem$) on previous day demand ($pdem$) and the previous day fuel prices ($pprice$). Taking these explanatory variables separately, we have Figure 19 which shows the impact of the previous day demand of electricity on the current daily rate of change of electricity demand. Figure 20 shows the impact of fuel prices on the daily rate of change of electricity demand.

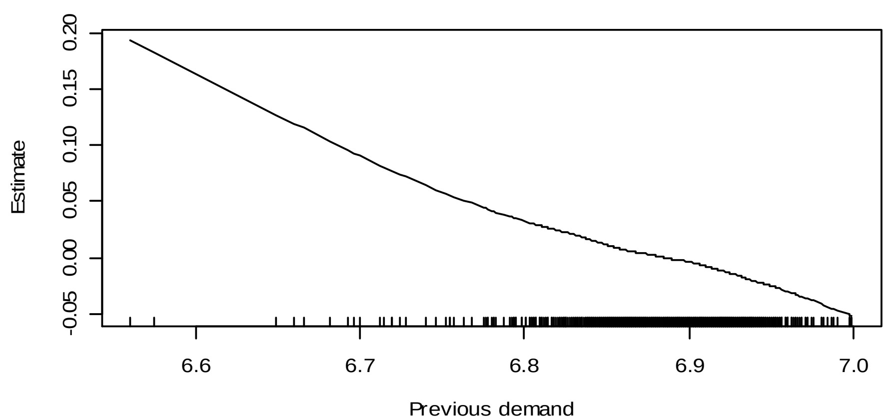

Figure 19 Contribution of the previous day demand of electricity on the daily rate of change of electricity demand

Figure 20 shows that the cross elasticity of demand which measures the responsiveness of the quantity demanded of electricity to changes in price of fuel is indicated to be elastic in the long run. This is because in the long run both the demand for electricity and fuel prices are increasing

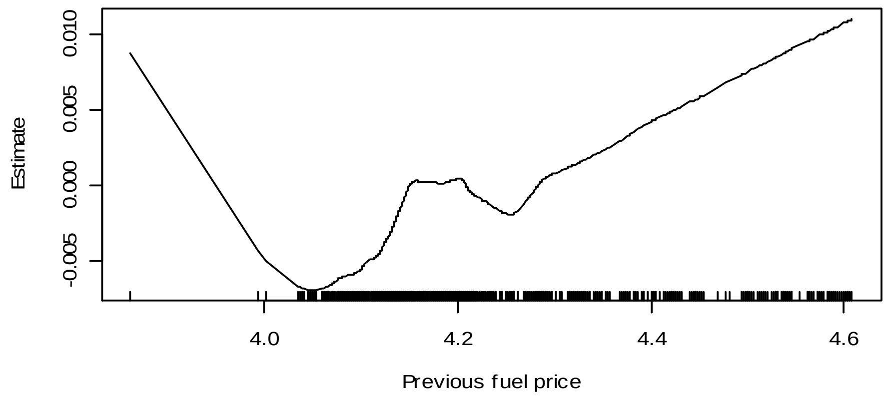

Figure 20 Contribution of fuel prices on the daily rate of change of electricity demand

5.4 Modeling time dependencies

Due to the nature of impact of fuel prices on electricity demand, we focus on the distribution of daily fuel price changes and the associated extreme quantiles by fitting a EVT based model to observed price data. We focus only on the positive tail of the distribution. Since the price changes are so large, we chose logarithmic changes instead of simple net returns $(P_t - P_{t-1})/P_{t-1}$. A problem with using simple returns is that prices are bounded from below and that this makes the return distribution skewed for large positive and negative returns. The data consists of 912 daily prices from July 1, 2006 to June 20, 2009 (Data from Kenya National Oil Corporation).

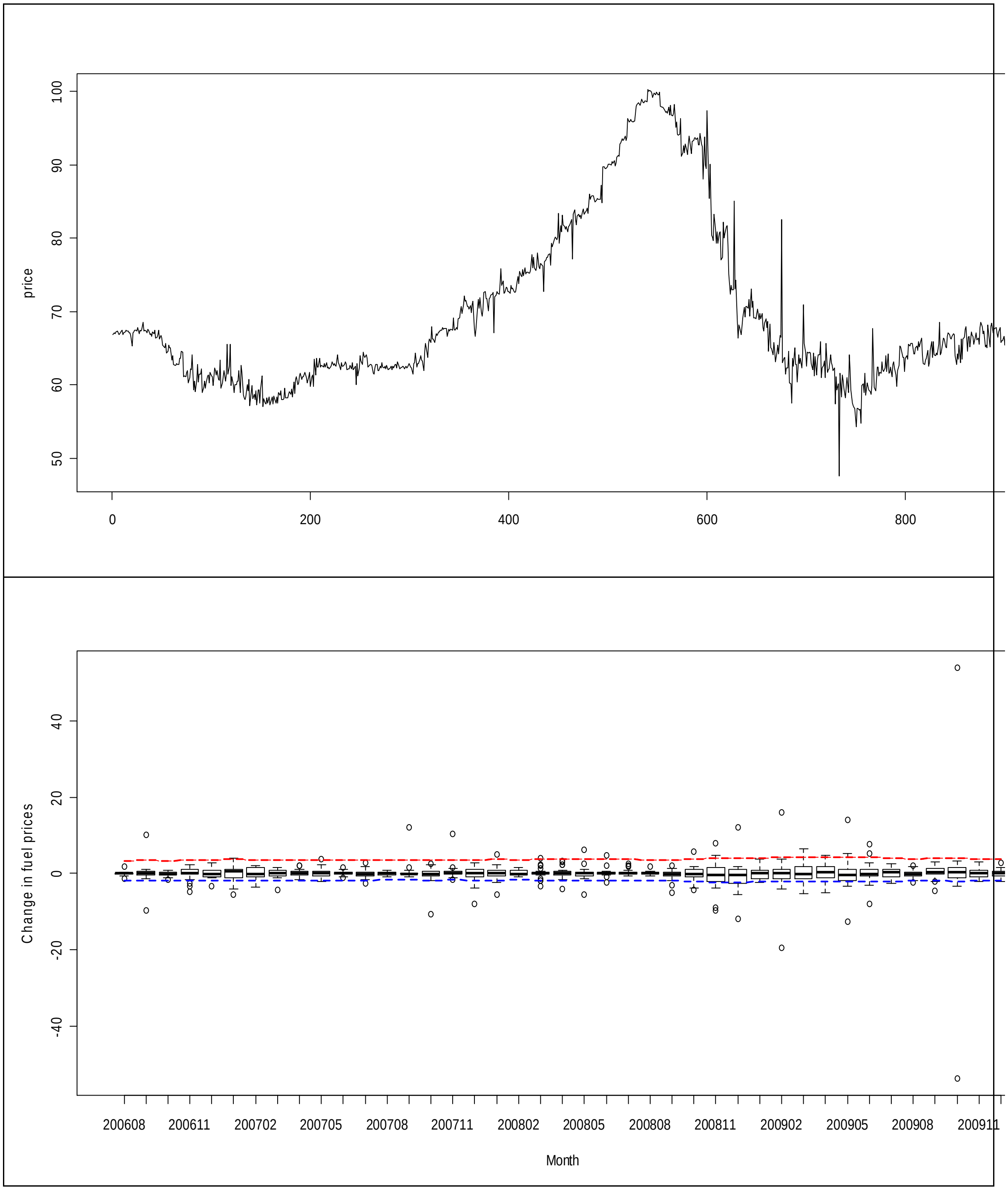

Figure 21 Time series plot of the daily fuel prices from July, 2006 to November 2009, while below is the corresponding daily rate of change in fuel prices.

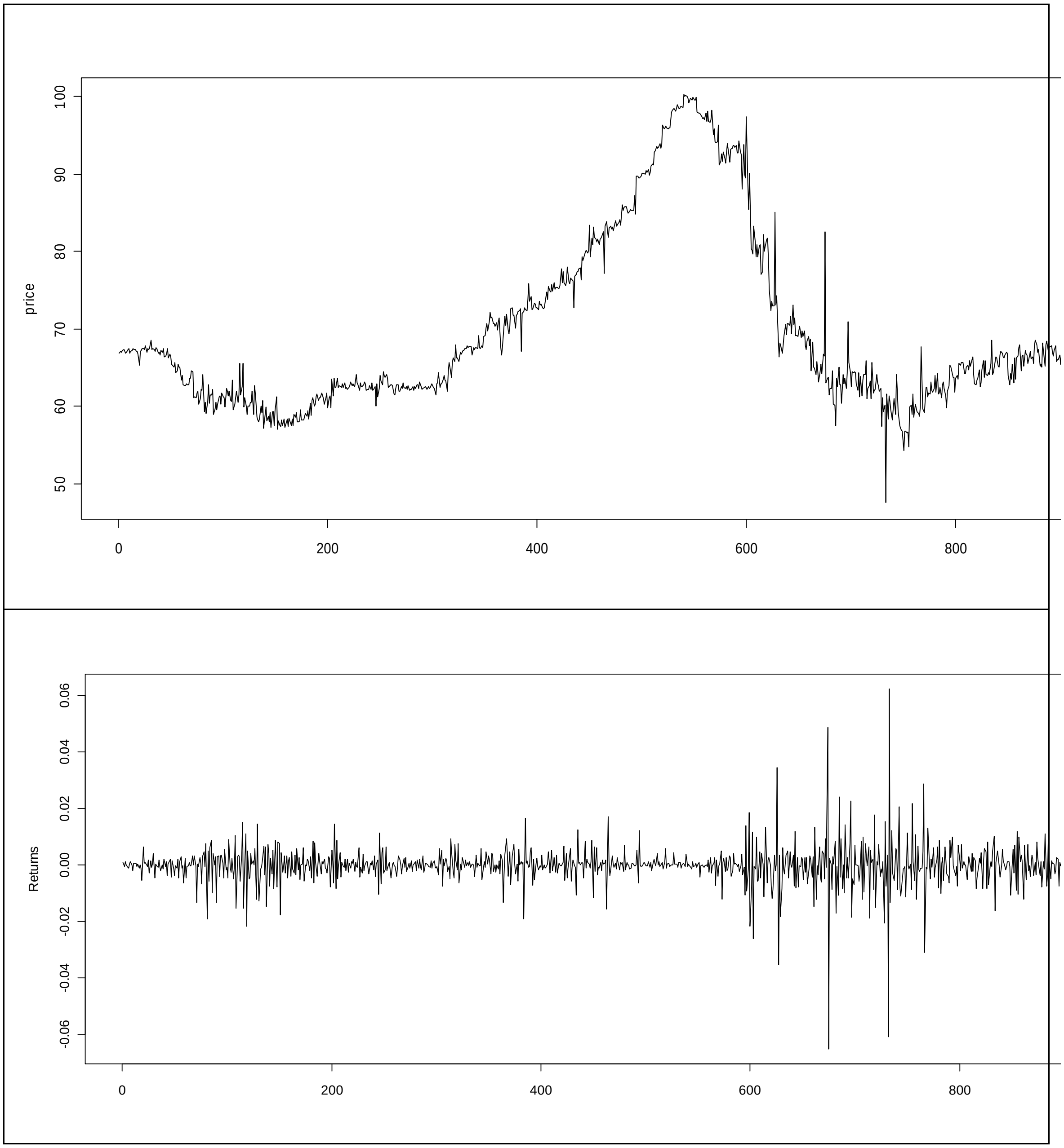

Figure 22 Fuel price changes (%) superimposed on the time series trend of fuel prices

Table 5 Descriptive statistics of daily log fuel price changes

Mean	Standard deviation	Skewness	Kurtosis	Ljung P-value
7.958483e-05	4.573012	-0.2899951	23.3798	2.2e-16

Table 5 reports some statistics on the price changes series. The high volatility is confirmed as evidenced by the high standard deviation and the very high excess kurtosis, also the very small p-value of the Portmanteau test (Ljung) implies that we have enough evidence to reject the null hypothesis that the series are uncorrelated. This together with a visual inspection of Figure 22 indicates a high degree of volatility clustering ($GARCH$) effect.

We extend the classical unconditional extreme value approach by first filtering the data to capture some of the dependencies in the fuel market and there after applying ordinary extreme value techniques. In this way we are able to get better tail estimates in sample, but more importantly we are able to get better predictions for future extreme price changes. In additional the independent and identical assumption behind the EVT based tail quantile estimator is less likely to be violated. In order to prefilter the time series we chose a combined autoregressive moving average ($ARMA$) and the generalized autoregressive conditional heteroscedastic ($GARCH$) model due to the strong and significant volatility clustering in the fuel market. By combining this $ARMA$ model with the simplest $GARCH(1,1)$ model we shall explicitly model the conditional volatility as a function of past conditional volatilities as shown by Bollerslev (1986) and

Engle (1982). We hope to capture the most important dependencies in the rate of change series. Thus we have

$$r_t = a_0 + a_1 r_{t-1} - b_1 e_{t-1} + e_t$$
$$\sigma_t^2 = \varphi_0 + \varphi_1 e_{t-1}^2 + \varphi_2 \sigma_{t-1}^2 \qquad (5.4.1.1)$$

Where σ_t^2 is the conditional variance of e_t. $e_t = \sigma_t \varepsilon_t$, with $\varepsilon_t \sim N(0,1)$ or student t-distributed *i.i.d* innovations. In order to introduce the extreme value theory to the estimation of extreme tail quantiles, the first step is to model the residuals e_t from the normal $(ARMA - GARCH)$ model with the POT model. Since the residual series is much closer to being identical and independent than the original series, it is straight forward to apply EVT. Let the unconditional EVT quantiles of the residual distribution be $\alpha_p = q_\theta$, the second step is to calculate the conditional tail quantiles, $q_{t,\theta}$ of our original rate of change distribution as

$$q_{t,\theta} = a_0 + a_1 r_{t-1} - b_1 e_{t-1} + \sigma_t \alpha_p \qquad (5.4.1.2)$$

where

$$\alpha_p = u + \frac{\alpha}{\xi}\left(\left(\frac{n}{N_u}p\right)^{-\xi} - 1\right) \qquad (5.4.1.3)$$

In Implementing the POT method we fit the GPD to observations in the residual series above a certain high threshold. We rely on a threshold that is approximately 5.5% following the recommendations from the simulation study in McNeil and Frey (2000) and as shown in the mean excess function in Figure 23. Too low threshold value and the asymptotic theory break down. Too high threshold value and one does not have enough data points to estimate the parameters in the excess distribution.

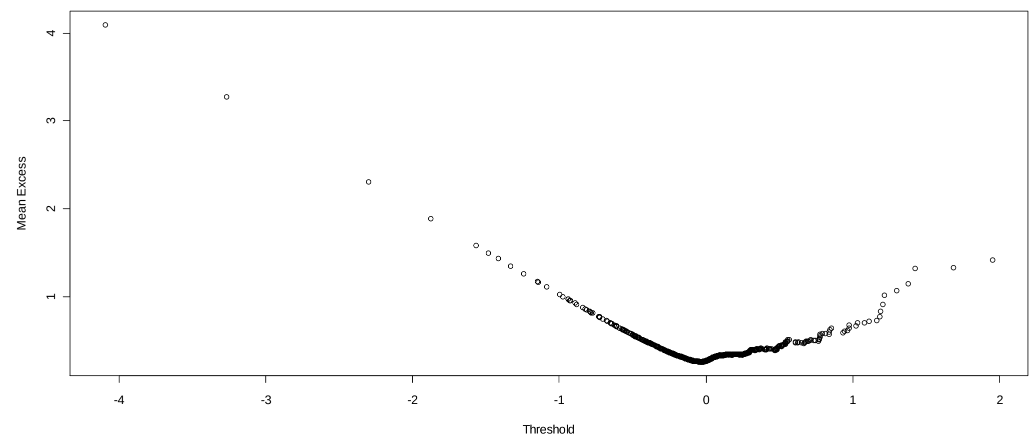

Figure 23 Mean excess function plot of negative fuel price change

In Figure 23, the downward trend shows thin tailed behavior whereas a line with zero slope show an exponential tail. An upward sloping plot indicates heavy tailed behavior. The mean excess plot is linear around $u > 0.01$ a sign of Pareto behaviour with positive tail index.

From table 6, the autocorrelation in the raw rate of change data has actually been removed. We obtain the unconditional α_p by inserting the GPD parameters obtained into (5.4.1.3). Finally the conditional tail quantiles of our original return series are calculated using (5.4.1.2).

Table 6 AR-GARCH parameters, statistics of the standardized residuals as well as GPD parameters.

AR-GARCH parameters	
a_0	***
a_1	-0.0472(0.0576)***
b_1	-0.6168(0.0263)
$\varphi_0 \times 10^{-3}$	1.0576(0.2733)
φ_1	0.14697(0.012)
φ_2	0.8758(0.0090)
Standardized residuals descriptive statistics	
Mean	0.03256153
Standard deviation	1.011791
Skewness	0.0012234
Kurtosis	10.60099
Ljung P-value	0.97
GPD parameters	
ξ	0.04635(0.0393)
α	0.611095(0.0493)
u	**0.055**

Data Source: Kenya National Oil Corporation. *** Not significant parameter estimates. Figures in parentheses are standard errors.

From Table 7, with 5% probability, the daily rate of change of fuel price could be as low as –0.624% and given that the rate of change is less than –0.624% the average rate of change value is –1.404%. With 1% probability, the daily rate of change of fuel price could be as low as –1.3818% and given this rate of change, the average rate of change value is –2.1870%.

Table 7: Risk measures of fuel prices computed via Peak over threshold at 5%, 1% and 0.1% probability

P	quantile	Shortfall
0.95	0.624	1.404
0.99	1.3818	2.1870
0.999	3.2574	4.7766

CHAPTER SIX

6.0 SUMMARY, CONCLUSION AND RECOMMENDATION

6.1 Summary

This study has identified that the distribution of electricity demand in Kenya has continued to evolve as a markedly multimodal, with a stochastic trend. Changes in fuel prices have some impact on electricity demand and in the long run the cross elasticity for all classes of electricity consumers is much larger in the long run than in the short run. From this study, electricity is continuously becoming a necessity rather than a luxury since the cross elasticity in the long run is inelastic. Aberrant observations in electricity demand data tend to occur in clusters due to market mechanism factors which include fuel prices and demand polynomial. It has also been found that cross validation and generalized cross validation are asymptotically equivalent as methods of determining the smoothing parameter of demand data and they do perform poorly due to the large sample variation and even worse for dependent data.

The results on how EVT can be used in the fuel market risk management has also been given and the foundation of EVT has been considered in details. It has been found that with 0.1% probability the daily rate of change of fuel prices would be as low as -3.2574% and given that the average rate of change is less than -3.2574 the rate of change value is -4.7766% . This study serves as a literature review on extreme energy modeling and on market risk management.

Thus the reference in this work provides a collection of relevant information which give additional insight to the theory and application of EVT.

6.2 Conclusion

The energy market is more complex in terms of demand and price formation than other more conventional markets. The complexity arises from the fact that in practice utility like electricity cannot be stored and the demand formation is affected by several other factors such as the previous day demand and fuel price patterns which are not standard across the markets of which its effects are clear. The results are interesting for a number of reasons

1. Identification is nonparametric and may be achievable at some quantiles thus the results offers a possibility of extracting information about the distribution of exogenous impact across different quantiles of the marginal distributions of the observable variables that drive the structural model.
2. The exogenous impact function can be defined and identified in contexts especially in financial markets in which it is attractive to construct models with non existence low order moments.

We may conclude that sequential nonparametric regression can be used in assessing trends in shortfalls on a risky demand rather than relying on emphasis disclosed by researchers. This could be used as an additional analytical tool for investors to know more about demand.

6.3 Recommendation

We recommend to regulators and investors venturing into the energy and energy related industries, to first use this method to consider the risk associated with the venture so that they can hedge appropriately. This study points on the future research to lay emphasis on how the goodness-of-fit of this model can be achieved. This includes among others the multivariate extreme value theory achievable through marginal dependencies using marginal distribution (Copula). Future research should quantify on the set of factors that affect fuel price formation. These factors include economic fundamentals, production constraints, trading inefficiency and market design effects which may be incomplete. Lastly the occasional congestion problems in the power net works need further research on how they can be modeled.

REFERENCES

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P. (2003) Modeling and forecasting realized volatility, *Econometrica*, **71**, 529-626.

Angelica,E. N, Stephen, N.K. and Mwangi, S.K. (2005). Measuring potential output and output gap and macroeconomic policy: The case of Kenya, "Working papers-University of Connecticut, Department of Economics"

Artzener, P, Delbaen, F., Eber, J.M. and Heath, D. (1999). Coherent measure of risk, *Math, Finance* **9**, 203-228

Bandi, F. and Nguyen, T.H. (2000). Fully nonparametric estimators for diffusions, working paper. *Graduate School of Business*. The University of Chicago, Chicago IL

Barone-Adesi, Bourgoin, G. and Giannopoulos, F. (1988). Don't look back, *Risk 11,* (8)

Beirlant, J., Tertius, D.W. and Yuri, G. (2004). Nonparametric estimation of extreme conditional quantiles, *Journal of Statistical Computation*, **74**, (8), 567-580

Berman, S.M (1964). Limiting theorems for the maximum term in stationary sequences, *Annals of Mathematical Statistics*, **35**, 502-516

Bhattacharya, P.K. and GangoPadhyay, A.K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile, *Annals of Statistics*, **18**, 1400-1415

Birol, F. (2001) Energy Price Volatility: Trends and Consequences, Working paper, International Energy Agencies

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy*, **71**, 637-654

Blundell, R.,Browning, M. and Crawford, I. (2003). Nonparametric Engel curves and revealed preference, *Econometrica*, **71 (**1), 205-240

Bollerslev, T (1986). Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics*, **31**, 307-327

Brockwell, P.J. and Davis, R. A. (1991).*Time series: Theory and Methods*, 2nd edition. New York: Springer-Verlag

Bunn, D. W. and Karakatsani, N. (2003) *Forecasting electricity prices*, London Business School

Byström, H.N.E. (2005). Extreme value theory and extremely large electricity price changes, *International Review of Economics and Finance.* **14**, (1), 41-55

Cai, Z. (2002). Regression quantiles for time series, *Econometric Theory*, 169-192.

Chen, S. X and Tang, T. (2005). Nonparametric inference of value at risk for independent financial returns, *Journal of Financial Econometrics*, **3**, 227-225

Cheng, K. F. and Lin, P. E. (1981) Nonparametric estimation of a regression function, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **57**: 233-33.

Clark, R.M. (1977). Nonparametric estimation of smooth regression function, *Journal of the Royal Statistical Society,* **B39**, 107-113

Collomb, G. (1980). Estimating nonparametriqe de probabilities conditionnelles. *C. R. Acad Sci. Paris* .**1** (291), 427-430

Danielsson, J., and de Vries, C.G. (1997). Tail index and quantile estimation with very high frequency data, *Journal of Empirical Finance,* **4**, 241-257

Dawson, A.G. (2000). Tsunami deposits, *Pure and Applied Geophysics*, **157**, 875-897.

de Rozario, R. (2002). Estimating value at risk for the electricity market using a technique from extreme value theory, University of New South Wales, School of Banking and Finance

Dunne J. F. and Ghanbari , M. (2001). Efficient extreme value prediction for nonlinear beam vibrations using measured random response bistories, *Nonlinear Dynamics* **(24)**, 1, 71-101

Embretchts P., C. Klueppelberg and Mikosch, T. (1997). *Modeling extremal events for insurance and finance,* Springer, Berlin

Engle, R. F (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom Inflation, *Econometrica,* **50**, 4,987-1007

Engle, R and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantile, *Journal of Business and Economics Statistics*, **22**, 367-381

Fan, J., Heckman, N. E and Wand, M. P. (1995) Local polynomial kernel regression for generalized linear models and quasi likelihood functions, *Journal of the American Statistical Association,* **90**, 141-150

Fan,J., Hu, T.-C, and Troung, Y.K. (1994). Robust nonparametric function estimation, *Scandinavian Journal of Statistics*, **21**, 433-446

Fan, J., Yao,Q. and Tong, H. (1996). Estimation of conditional densities and sensitivities measures in nonlinear dynamical systems, *Biometrika*, **83**, 189-206

Fan, J. (1993) Local linear regression smoothers and their minimax efficiencies, *Annals of Statistics*, **21**,196-216

Fan, J. and Gijbels, I. (1996) *Local polynomial modeling and its application* Chapman and Hall, London

Gasser, T. and Műller, H. G. (1979). Kernel estimation in regression functions. In Smoothing Techniques for Curve Estimation, *Lecture notes in Mathematics* **742** Springer-Verlag New York

Giot, P., Laurent, S. (2003). Market risk in commodity market: a VaR approach. *Energy Economics*, **25**, 435-457

Glosten, L., Jagannathan, R. and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess returns on stocks *Journal of Finance, **48,*** 1791-1801

Gnedenko, B.V. (1943). Sur la distribution limite du terme d'une Sėrie a le'atoire *Annals of Mathematics*, **44**, 423-453

Goldie, C. M. and Smith, R. L. (1987). Slow variation with remainder: Theory and applications, *Quart. J. Math. Oxford (2)* **38**, 45-71.

Gourieroux, C. (1992). *Modeles ARCH et. Applications Financieres, Economica*, Paris.

Hall, P., Wolff, R.C.L and Yao, Q (1999). Methods of estimating a conditional distribution function, *Journal of the American Statistical Association*, **94**. 154-163

Härdle, W (1989). Asymptotic maximal deviation of M-smoothers, *Journal of Multivariate Analysis*, **29**, 163-17

Harris, R. D. F. (2001). The expectations hypothesis of the term structure and time-varying Risk Premia: a Panel Data Approach, **63,** (2) 149-273

Hausman, J . A and Newey, W. (1995). Nonparametric estimation of exact consumers surplus and dead weight loss, *Econometrica*, **63**, 1445-1476

Hendricks, W. and Koenker, R. (1991) Hierarchical spline models for conditional quantiles and the demand of electricity, *Journal of the American Statistical Association*, **87**, 58-68

Horowitz, J. (2001). Asymptotic normality of a nonparametric Instrumental variables estimator, *International Economic Review* **48** 1329-1349.

International Energy Agency (2010) International Energy Statistics, Short Term Energy Outlook

Jenkinson, A.F. (1955). The frequency distribution of the annual maximum (minimum) values of meteorological elements, Quarterly, *Journal of the Royal Meteorological Society,* **81**, 158-171

Jorion, P (2001). *Value- at- risk, The New Benchmark for Managing Financial Risk*, McGraw-Hill New York

Koenker, R. and Basset, G. S. (1978) Regression quantiles, *Econometrica,* **46**, 33-50

Koenker, R. and Basset, G. S. (1982) Robust test for heteroscedasticity based on regression quantiles. *Econometrica*, **50** (1), 43-61

Koenker, R. (2004). Quantreg: An R package for quantile regression and related methods. ***http://cran.r-project.org.***

Kuester, K. Mittnik, S. and Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, **4**, 53-89

Leadbetter, M.R. (1974). *Extremes and Related Properties of Random Sequences and Processes*, Spriger-Verlag, New York

Li, Q. and Racine, J. S. (2004). Cross validated local linear nonparametric regression. *Statistica Sinica*, **14** (2), 485-512.

Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society* (B) **34**, 1-41.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675

Manganelli, S. and Engle, R. (2001). Value at risk models in finance European Central Bank, Working paper series, 75.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, **24**,165-179

McNeil, A.J. and Frey, R. (2000) Estimation of tailed-related risk measures for heteroscedastic financial time series: an extreme approach, *Journal of Empirical Finance,* **7**, 271-300

McBulty, P.J., Scheick,L.Z.,Roth, D.R., Davis,M.G. and Tortora, M.R.S.,(2000). First failure predictions for EPROMs of the type flown on the MPTB satellite, *IEE Transactions on Nuclear Science*, **47**, 2237-2243

Mhilu, C. F. (2007). Development of regional equations for estimation of low flow regimes of small catchments on Pemba Island, Zanzibar. *Journal of Environmental Hydrology*, **15**, (7), 1-7.

Mwita, P. (2003). Semi parametric estimation of conditional quantiles for time series with applications in finance, A Doctor of Philosophy Thesis Presented to in the *University of Kaiserslautern*. Germany.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probability Application*, 9, 141-142

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge University Press, New York.

Parzen, E. (1962). On estimation of probability density function and mode. *Annals of Mathematical Statistics,* **33**, 1065-1076.

Priestly, M.B. and Chao, M.T. (1972) Nonparametric function fitting. *Journal of the Royal Statistical Society*, **34**, 385-392.

Powell, J. (1986). Censored regression quantiles. *Journal of Econometrics,* **32**, 143-155.

Rice, J. (1984). Bandwidth selection for nonparametric regression. *The Annals of Statistics,* **12**, 1215-1230.

Rosenblatt, M. (1956). A Central Limit Theorem and Strong Mixing Conditions. *Pro.Nat. Acad. Sci.* **4**, 43-47.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, **77**, 828-838

Scaillet, O. (2005). A Kolmogorov-Smirnov type test for positive quadrant dependence. *Canadian Journal of Statistics*, 33, 415-427.

Sebitosi, A. B (2010). Is the South African electricity tariff model conducive to an energy efficiency economy? *Energy for Sustainable Development*, **14**, 315-319.

Sebitosi, A. B. and Okou, R. (2010). Rethinking the power transmission model for Sub-Saharan Africa. *Energy Policy*, **38**, 1448-1454.

Sharpe, W. (1964). Capital asset prices, a theory of market equilibrium under conditions of risk. *Journal of Finance*, **19**, 452-442

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Smith, A. (1973). Bayes estimates in one way and two way models. *Biometrica*, 60, 319-329

Sokolov, Y. A. and Mc Donald, A. (2006). The Nuclear power options for Africa. *ATDF Journal*, **2** (2).

Stone, C.J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, **5**, 595-620.

Taylor, L. D. (1975). The demand for electricity: A survey. *The Bell Journal of Economics*, **6**, (1), 74-110

Taylor, L. D. (1977). The demand for energy: A survey of price and income elasticity in *International Studies of the Demand for Energy*, Ed. William, D. N. North Holland, Amsterdam.

Taylor, J.W. (2008). Using exponentially weighted quantile regression to estimate value at risk and expected shortfall. *Journal of Financial Econometrics*, **6**, 382-406.

Tryon, R.G. and Cruse, T.A. (2000) Probabilistic mseomechanics for high cycle fatigue life prediction. *Journal of Engineering Materials and Technology-Transactions of the AMSE*, **122**, 209-214.

Tsay, R.S. (2010). *Analysis of Financial Time Series*, John Wiley and Sons, New Jersey

Volkonskii, V.A., and Rozanov, Y.A. (1959). Some limit theorems of random functions. *Theory* of *Probability and Its Applications,* 4, 178-197.

Watson, G.S. (1964). Smooth regression analysis. *Sankhya,* **26** (15), 359-372.

Yu, K. and Jones,M. C. (1998). Local Linear Quantile Regression., *Journal of the American Statistical Association*, **93**, 228-237.

Zikovic, S. and Fatur, T. (2007). Measuring oil price volatility as a means of managing commodity risk *Oil Journal*, **58**, 133-145.

APPENDICES

Appendix A.1 Bandwidth determination codes

```
library(KernSmooth)

 kscv1=function(xx,yy,nd,bw,ntrial)

    {kscvgcv=function(bw, xx, yy)

     {nd <- length(xx)

       bwsplus <- bw

       fit.ks <- ksmooth(xx, yy, "normal", bandwidth = bwsplus)

       res <- yy - fit.ks$y

       dhat1 <- function(x2, bw)

     {

     nd2 <- length(xx)

      diag1 <- diag(nd2)

       bwsplus <- bw

     dhat <- rep(0, length = nd2)

     for(jj in 1.:nd2) {

       y2 <- diag1[, jj]

       fit.ks <- ksmooth(x2, y2, "normal", bandwidth = bwsplus)

     dhat[jj] <- fit.ks$y[jj]

     }

     return(dhat)

     }
```

```
        dhat <- dhat1(xx, bw)
        trhat <- sum(dhat)
        sse <- sum(res^2)
 sigma<-log(sse/nd)
rice<-sigma-log(1-((2*trhat)/nd))
  AICc<-sigma+(((1+trhat)/nd)/(1-((trhat+2)/nd)))
  return(AICc)
      }
     cvgcv<-lapply(as.list(bw),kscvgcv,xx=xx,yy=yy)
     cvgcv<-unlist(cvgcv)
     return(cvgcv)
     }
```

Appendix A.2 Partial plots codes

```
bw.all=npregbw(cdem~pdem+pprice,regtype="ll",bwmethod="cv.aic")
 model.np<-npreg(bws=bw.all)
plot(model.np,plot.errors.method="bootstrap",plot.errors.boot.num=100,plot.errors.type
="quantiles",plot.errors.style="band",common.scale=FALSE)
```

Appendix A.3 Local polynomial regression quantiles codes

```
"lprg" <- function(x,y,h,m,theta=.50)
     {
        xx <- seq(min(x),max(x),length=m)
        fv <- numeric(length(xx))
        dv <- numeric(length(xx))
```

```
        for(i in 1:length(xx))
        {
        z <- x - xx[i]
        wx <- dnorm(z/h) # kernel
        r <- rq(y ~ z,weights=wx,tau=theta,ci=FALSE)
        fv[i] <- r$coef[1]
        dv[i] <- r$coef[2]
        }
        list(xx=xx,fv=fv,dv=dv)
       }
 Xx <- 1:2241
 fit <- lprg(Xx, Yt,h=2.5,m=500,theta=0.1)
win.graph(width=4.85,height=3.5,pointsize=8)
 plot(1:length(Yt),Yt,type="l",xlab="Time",ylab=expression(Y[t]),lty=3)
 lines(fit$xx,fit$fv,col=2,lty=3)
 plot(1:length(Yt),Yt,type="l",xlab="Time",ylab=expression(Y[t]),lty=3)
 lines(fit$xx,fit$fv,col=2)
 fit1 <- lprg(Xx, Yt,h=2.5,m=500,theta=0.5)
 plot(1:length(Yt),Yt,type="l",xlab="Time",ylab=expression(Y[t]),lty=3)
 lines(fit1$xx,fit1$fv,col=3)
 plot(1:length(Yt),Yt,type="l",xlab="Time",ylab=expression(Y[t]),lty=3)
 lines(fit1$xx,fit1$fv,col=4)
 fit2 <- lprg(Xx, Yt,h=2.5,m=500,theta=0.9)
 plot(1:length(Yt),Yt,type="l",xlab="Time",ylab=expression(Y[t]),lty=3)
```

```
lines(fit2$xx,fit2$fv,col=2)
```

Appendix A.4 PROOF

Proof of (2.4.4.2)

$$m(X_t) = m(x)f_n(x) + m''(x)\underbrace{\frac{1}{n}\sum_{t=1}^{n}(X_t - x)K_h(X_t - x)}_{J_1(x)}$$

$$+ \frac{1}{2}\underbrace{\frac{1}{n}\sum_{t=1}^{n} m''(x_t)(X_t - x)^2 K_h(X_t - x)}_{J_2(x)}$$

then,

$$E[J_1(x)] = E[(X_t - x)K_h(X_t - x)]$$

$$= \int (u - x)K_h(u - x)f(u)du$$

$$= h\int uK(u)f(x + hu)du$$

$$= h^2 f'(x)\mu_2(K) + o(h^2)$$

$$nh\,Var\big(J_1(x)\big) = O(1)$$

$$E[J_2(x)] = E[m''(x_t)(X_t - x)^2 K_h(X_t - x)]$$

$$= h^2\int m''(x + \theta hu)u^2 K(u)f(x + hu)du$$

$$= h^2 m''(x)\mu_2(K)f(x) + o(h^2)$$

and $Var\big(J_2(x)\big) = O(1/nh)$,

therefore $E\{J_2(x)\} = h^2 m''(x)\mu_2(K)f(x) + o_p(h^2)$

hence

$$I_1 = m(x) + m'(x)J_1(x)/f_n(x) + \frac{1}{2}J_2(x)/f_n(x)$$

$$= m(x) + \frac{h^2}{2}\mu_2(K)[m'(x) + 2m'(x)f'(x)/f(x)] + o_p(h^2)$$

By the fact that $f_n(x) = f(x) + o_p(1)$,

$$B_{nw}(x) = \frac{h^2}{2}\mu_2(K)[m''(x) + 2m'(x)f'(x)/f(x)]$$

■