BOOTSTRAP UNIFORM CONFIDENCE BANDS FOR A LOCAL LINEAR NONPARAMETRIC ESTIMATOR AND APPLICATIONS TO FINANCIAL RISK MANAGEMENT

***O. K. Ananda*[1]*, P. N. Mwita*[2]**
Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Naiorbi, Kenya

Abstract
This paper considers the problem of bootstrapping a local linear estimator in conditional quantile estimation of a financial time series assuming independent and identically distributed errors. A nonparametric regression bootstrap generating process is estimated, then bootstrap confidence bands fitted to the quantile estimates. Under appropriate assumptions, the local linear bootstrap estimator is known to be consistent.

Key words: Quantile estimation, bootstrap, local linear, consistency

1.0 Introduction

Consider a partitioned stationary α-mixing time series (X_{t+1}, ξ_t) where the real valued response variable $X_{t+1} \in \Re$ is F_{t+1} measurable and the variate $\xi_t \in \Re^P$ which is F_{t-} measurable. For some $0 < \theta < 1$, we are interested in estimating the conditional θ quantile of X_{t+1} given the filtration F_t, and assuming that it is completely determined by ξ_t , we can consider the model:

$$X_t = f(X_{t-1}) + s(X_{t-1})\xi_t \qquad \text{t} = 1,2,3,... \qquad 1.1$$

Here the innovations ξ_t are assumed to be iid random variables with $E(\xi_t) = 0$ and $E(\xi_t^2) = 1$. Under these assumptions it holds that:

$$\begin{aligned} E(X_t / X_{t-1} = x) &= f(x) + E[s(X_{t-1})\xi_t / X_{t-1})] \\ &= f(x) + s(x)E(\xi_t) \\ &= f(x) \end{aligned} \qquad 1.2$$

A similar calculation gives $s^2(x) = \text{var}(X_t / X_{t-1} = x)$. The unknown functions f and s describe the conditional mean and the conditional volatility of the process, which we want to estimate. We have developed a nonparametric regression methodology that will help estimate the values of s and f as shown above.

Nadaraya (1964) and Watson (1964) discussed a kernel smoothing for the nonparametric estimator function of f and s as in the model equation (1.1) above. Assuming finite moments of up to order 4, Franke & Wenzel (1992) and Kreutzberger (1993) proposed an autoregression bootstrap re-sampling scheme that approximates the laws of a kernel estimator for f and s. Franke et. al (2002) considered two estimators for the estimation of conditional variance and gave consistency of the residual based and Wild bootstrap procedures for f and s.

Although the model errors in model equation (1.1) are homoscedastic, we'll followed similar lines as in Franke et. al. (2002) for bootstrapping f and s. We also developed uniform confidence bands to give an idea about the global variability of the estimate from the model equation (1.1) since it's clear that curve of fit contains the lack of fit test as an immediate application. Hardle and Song (2010) used strong approximation of empirical process and extreme value theory to construct the uniform band over the estimator; however the poor convergence of the extremes of a sequence of n independent normal random variables is also investigated by Fisher and Tippet (1928). The slow convergence of kernel estimators as identified by Pritsker (1998) is the main reasons for the poor finite sample performance. In their write up, the casted doubt in the applicability of first order asymptotic theory of nonparametric methods in finance, since persistent serial dependence is a stylized fact for interest rates and many other high frequency financial data. Another fact is that a kernel estimate produces biased estimates near the boundaries of the data as discussed by Hardle (1990) and Fan and Gijbels (1996). Boundary bias can generate spurious nonlinear drift, giving misleading conclusions of the dynamics of X_{t+1}. Recently, Hong & Li (2002) have developed a nonparametric test for the model using the transition density, which can capture the full dynamics of X_{t+1}. It has been suggested that to avoid the boundary bias then kernel smoothing can be applied by methods of local polynomial Fan and Gijbels (1996) or a weighted Nadaraya Watson kernel estimator Cai (2001)

2.0 The Local Linear Estimator

The local estimation of $\mu(x)$ means estimating $\mu(.)$ separately for each $(m \times 1)$ vector $x = (x_1, x_2, \ldots, x_m)'$ of interest. Note that x is scalar if $m = 1$ and x_t is scalar. The starting point for deriving the local linear estimator is the fact that, although $\mu(x)$ is not observable, it appears in a first-order Taylor expansion of $\mu(x)$ taken at x

$$\mu(x_t) = \mu(x) + \frac{\partial \mu(x)}{\partial x'}(x_t - x) + R(x_t, x) \qquad 1.5$$

Where $R(x_t, x)$ denotes the remainder term. Inserting this expansion into the model equation (1.1) the gives

$$y_t = \mu(x) I + \frac{\partial \mu(x)}{\partial x'}(x_t - x) + R(x_t, x) + \xi_t \qquad 1.6$$

where the ξ_t denotes the stochastic error term.

The right hand side contains two known terms, the constant one multiplied by the unknown $\mu(x)$ and the known term $(x_t - x)$ multiplied by a vector of unknown first partial derivatives $\frac{\partial \mu(x)}{\partial x'}$ i.e. were there no remainder term $R(x_t, x)$, one would have a simple OLS regression problem in which the estimated parameters correspond to the estimated function value $\hat{\mu}(x)$ at x and the estimated vector corresponds to the estimated function value $\overline{\mu}(x)$ at x and the estimated vector $\frac{\hat{\partial} \mu(x)}{\partial x'}$ of partial derivatives also evaluated at x. However, whenever the conditional mean function is non-linear, the remainder term $R(x_t, x)$ maybe nonzero at x. Using the standard OLS estimation would then result into biased estimates for which the size of bias depends on all remainder terms $R(x_t, x)$, $t = 1,2,3,\ldots,T$. One possibility to reduce the bias is to use only those observations x_t that are in some sense close to x. More generally, one down-weighs those observations that are not in a local neighborhood of x. If more data is available, it is possible to decrease the size of the local neighborhood, where the estimation variance and bias can decrease, i.e. the approximation error of the model can decline with sample size. Thus the underlying idea of nonparametric estimation.

The weighing is controlled by a so called kernel function $K(u)$ where the following we can assume the function is symmetrical, compact, non-negative univariate probability density so that $\int K(u)du = 1$. To adjust the size of the neighborhood one introduces a bandwidth h such that for a scalar x the kernel function becomes $\frac{1}{h}K(\frac{x_t - x}{h})$. So the larger the value of h the larger the neighborhood around x, where the sample observations receive a larger weight and the larger may be the estimation bias. Because a larger h implies function estimates will look smoother, the bandwidth h is called the smoothing parameter. Since the observations in the local neighborhood of x are the most important, this estimation approach is called local estimation. If $m > 1$ and $x = (x_1, x_2, \ldots, x_m)'$ is a vector, one uses a product kernel

$$K_h(x_t - x) = \prod_{i=1}^{m} \frac{1}{h^m} K\left(\frac{x_{ti} - x_i}{h}\right) \qquad 1.7$$

Here the component x_{ti} denotes the ith component of x_t. Instead of using a scalar bandwidth that imposes the same degree of smoothing in all directions, it is also possible to use a vector bandwidth that determines the amount of smoothing in each direction separately. The kernel variance can be given to as $\sigma_k^2 = \|K\|_2^{2m} = \int u^2 K(u)du$ and the kernel constant $\int K(u)^2 du$ both influence the asymptotic behavior of the local linear estimator.

Owing to the introduction of the kernel function, one has to solve a weighted least-squares problem

$$[\hat{c}_1, \hat{c}_2, ..., \hat{c}_m] = \arg\min_{\hat{c}_1, \hat{c}_2, ..., \hat{c}_m} \sum_{t=i_m+1}^{T} \left\{ y_t - c - \sum_{i=1}^{m} c_i (x_{ti} - x_i) \right\}^2 K_h (x_t - x) \qquad 1.8$$

which delivers the local linear function estimate $\hat{\mu}(x,h) = \hat{c}$ at the point x.

The bandwidth h is also included as an argument to indicate the dependence method for bandwidth choice, which is based on statistical procedures. A well grounded bandwidth is also necessary for valid asymptotic properties of these estimators.

3.0 Asymptotic Properties of the Local Linear Estimator

Let $f(x)$ denote the density of the lag vector at the point x and $tr(A)$ - the trace of matrix A. The asymptotic normal distribution of local linear estimator is given by:

$$\sqrt{Th^m} \left\{ \hat{\mu}(x,h) - \mu(x) - b(x)h^2 \right\} \xrightarrow{d} N(0, v(x)) \qquad 1.9$$

Where the asymptotic bias $b(x)$ and asymptotic variance $v(x)$ which can be given as:

$$b(x) = \frac{\sigma_K^2}{2} tr\left(\frac{\partial^2 \mu(x)}{\partial x \partial x'} \right) \qquad 2.0$$

$$v(x) = \frac{\sigma^2(x) \|K\|_2^{2m}}{f(x)} \qquad 2.1$$

This then becomes clear that, for the asymptotic normal distribution to exist, one has to require that $Th^m \longrightarrow \infty$ and $h \longrightarrow 0$ as $T \longrightarrow \infty$. Otherwise, the asymptotic distribution would collapse to a point or the bias would grow infinitely large. Inspecting the asymptotic bias term (1.9) more closely reveals that the second order partial derivatives of $\mu(x)$ have to exist. In fact for (2.1) to hold this has to be the case in a neighborhood of x hence one has to assume $\mu(.)$ is twice continuously differentiable on the support of $f(x)$. Because both the density $f(x)$ and the conditional variance $\sigma^2(x)$ enter the asymptotic variance (2.1), one also has to assume that both are continuous and the latter is positive on the support of $f(x)$. Initially the asymptotic distribution (1.9) was derived under the assumption that $\{y_t, x_t\}$ is a sample of i.i.d observations. Then x_t does not contain lags of y_t and there is no stochastic dependence between observations at different times. In the current situation, where x_t is a vector of lagged $y_t's$, a stochastic dependence clearly exists. Hadle & Yang (1998) showed that the asymptotic behavior of the local linear estimator (1.8) is the same as that encountered in the case of i.i.d variables if the stochastic dependence is sufficiently weak. At this point, it is sufficient to state that a stationary $ARMA(p,q)$ process satisfies the required conditions if its driving error process is not completely ill-behaved. For empirical work, it is most important to transform a given time series to be stationary. Thus prior to local linear estimation, one has to remove unit roots.

Some consequences of the implications of the asymptotic normal distribution can be given as:

$$\hat{\mu}(x,h) \approx N\left((\mu(x) + b(x)h^2, \frac{1}{Th^m} v(x) \right) \qquad 2.2$$

This nicely shows the asymptotic bias-variance trade-off. If h gets larger, the bias increases but the variance diminishes and vice versa. This asymptotic trade off will be used to obtain an asymptotically optimal bandwidth. By inspecting the formulae below that its rate of decline is $T^{-\frac{1}{(m+4)}}$, thus if we denote a positive constant by β, any bandwidth for which $h = \beta T^{-\frac{1}{(m+4)}}$, hold has the optimal rate to guarantee a balanced decline of bias and variance. Inserting, $h = \beta T^{-\frac{1}{(m+4)}}$ into (1.9) delivers the rate of convergence of the local linear estimator with respect to the number of observations T that is:

$$T^{\frac{2}{(m+4)}} \{\hat{\mu}(x,h) - \mu(x)\} \longrightarrow N\left(b(x)\beta^2, \frac{1}{\beta^m} v(x) \right) \qquad 2.3$$

It becomes apparent that the rate of convergence of the local linear estimator depends on the number m of lags and becomes quite slow if there are many lags, often called the curse of dimensionality of nonparametric estimators. Note that the rate of convergence is slower than for parametric estimators even if $m = 1$. This is the price one pays in nonparametric estimation for allowing the model complexity to increase with number of observations and thus to let the bias reduce with sample size. Such an increase in model complexity is in general not possible if one wants to obtain the parametric $\sqrt{T}$ rate.

By inspection, (2.0) one can see that the estimation bias also depends on the second partial derivative of the conditional mean function as well as on the kernel variance σ_K^2. The asymptotic variance (2.1) increases with conditional variance $\sigma^2(x)$ and decreases with the density $f(x)$. The intuition for the latter is that the larger the density, the more observations are on average close to the point x and thus available for local estimation, which in turn reduces the estimation variance.

4.0 Bandwidth and Lag Selection

The method for nonparametric bandwidth and lag selection described here is based on Tscherning & Yang (2000). For a lag selection it is necessary to specify a set of possible lag vectors a priori by choosing the maximal lag M. Denote the full lag vector containing all the lags up to M by $y_{t,M} = (x_{t-1}, x_{t-2}, ..., x_{t-M})'$. The lag selection task is now to eliminate from the full lag vector $y_{t,M}$ all lags that are redundant. This depends on choosing a relevant optimality criterion. A widely used criterion is the mean integrated squared error of prediction commonly known as final prediction error (FPE). We state it by using a weight function $w(.)$ needed for obtaining consistency of the lag selection procedure. One has to choose a weight function $w(.)$ that is continuous and non negative and for which $f(x_M) > 0$ for x_M in the support of $w(.)$. The simplest example is the indicator function.

$$C = \int \left(tr\left[\frac{\partial^2 \mu(x)}{\partial x \partial x'} \right] \right)^2 w(y_M) f(y_M) dy_M = E\left(\left(tr\left\{ \frac{\partial^2 \mu(x)}{\partial x \partial x'} \right\} \right)^2 w(y_{t,M}) \right) \qquad 2.4$$

And $b(h) = \|K\|_2^{2m} (T - i_m)^{-1} h^{-m}$ and $c(h) = \sigma_K^4 h^4 / 4$,

And $b(h)$ and $c(h)$ depend on the bandwidth and kernel constants.

The integrated variance of estimation and the integrant squared bias of estimation go to 0 for increasing sample size if $h \longrightarrow 0$ and $Th^m \longrightarrow \infty$ as $T \longrightarrow \infty$ holds.

5.0 Bandwidth Estimation

For minimizing the AFPE with respect to h, i.e. by solving the variance bias tradeoff between $b(h)B$ and $c(h)C$, one obtains the asymptotically optimal bandwidth:

$$h_{opt} = \left\{ \frac{m\|K\|_2^{2m} B}{(T - i_m)\sigma_K^4 C} \right\}^{\frac{1}{(m+4)}} \quad 2.5$$

In order for the optimal bandwidth to be finite, one has to assume that C defined is positive and finite. This requirement implies that, in the case of local linear estimation, an asymptotically optimal bandwidth h_{opt} for linear processes that is finite doesn't exist. This is because a first order approximation bias doe not exist, and thus a larger bandwidth has no cost i.e. clearly one should take a bandwidth as large as possible. It should be noted that h_{opt} is asymptotically optimal on the range where the weight function $w(.)$ is positive. For this reason it is also called the global asymptotically optimal bandwidth. Starting from the mean squared error of prediction:

$$\int \left[\int (\tilde{x} - \hat{\mu}(x,h))^2 f\left(\frac{\tilde{x}}{y}\right) d\tilde{x} \right] f(x_1,..., x_T) dx_1,..., dx_T \quad 2.6$$

which is computed at a given x, one would obtain a local asymptotically optimal bandwidth, which, by construction may vary with x If h_{opt} is estimated by consistent estimators for the unknown constants B and C, the resulting bandwidth estimate is known as plug-in bandwidth $\hat{h}_{opt}$. One way to estimate the expected value B consistently is given by averaging the weighted squared errors from the local linear estimates:

$$\hat{B}(h_B) = \frac{\frac{1}{T - i_m} \sum_{t=i_m+1}^{T} \{x_t - \hat{\mu}(x_t, h_B)\}^2 w(x_t, M)}{\hat{f}(y_t, h_B)} \quad 2.7$$

Where, $\hat{f}(.)$ is the Gaussian kernel estimator of the density $f(y)$.

Estimating h_B, one has to use Silverman's rule of thumb bandwidth Silverman (1986)

$$\hat{h}_B = \hat{\sigma}\left(\frac{4}{m+2}\right)^{\frac{1}{(m+4)}} T^{-\frac{1}{(m+4)}} \quad 2.8$$

With $\hat{\sigma} = \left(\prod_{i=1}^{m} \sqrt{\mathrm{var}(x_{ti}}\right)^{\frac{1}{m}}$ which denotes the geometric mean of the standard deviation of the regressors.

An estimator of C is give by

$$\hat{C}(h_C) = \frac{1}{T - i_m} \sum_{t=i_m+1}^{T} \left[\sum_{j=1}^{m} \hat{\mu}^{(jj)}(x_t, h_C) \right]^2 w(x_t, M) \quad 2.9$$

Where $\hat{\mu}^{(jj)}(.)$ denotes the second order direct derivative of the function $\mu(.)$ with respect to x_{ti}. For estimating higher order derivatives, one can use local polynomial estimation of appropriate order. In estimating second order direct derivatives it is sufficient to use the direct local quadratic estimator:

$$[\hat{c}_0, \hat{c}_{11}, ..., \hat{c}_{1m}, \hat{c}_{21}, ..., \hat{c}_{2m}] = \arg\min_{\hat{c}_0, \hat{c}_{11}, ..., \hat{c}_{1m}, \hat{c}_{21}, ..., \hat{c}_{2m}}$$

$$\sum_{t=i_m+1}^{T} \{x_t - c_0 - c_{11}(y_{t1} - y_1) - ... - c_{1m}(y_{tm} - y_m) - c_{21}(y_{tm} - y_m)^2 - ... - c_{2m}(y_{tm} - y_m)^2\}^2 K_h(y_t - y) \quad 3.0$$

The estimation of the direct second derivatives are then given by $\hat{\mu}^{(jj)}(x_t, h) = 2\hat{c}_{2j}, \ \mathrm{j} = 1, ..., \mathrm{m}$. Excluding all cross term does not affect the convergence rate while keeping the increase in the parameters $c_0, c_{1j}, c_{2j}, \quad \mathrm{j} = 1, .., \mathrm{m}$ linear in the number of lags m. This approach is a partial cubic estimator proposed by Yang & Tschernig (1999), who also showed that the rule-of-thumb bandwidth:

$$\hat{h}_C = 2\hat{\sigma}\left(\frac{4}{m+4}\right)^{\frac{1}{(m+6)}} T^{-\frac{1}{(m+6)}} \quad 3.1$$

Has the optimal rate. The plug-in bandwidth

$$\hat{h}_{opt} = \left\{\frac{m\|K\|_2^{2m} \hat{B}(\hat{h}_B)}{(T - i_m)\hat{C}(\hat{h}_C)\sigma_K^4}\right\} \quad 3.2$$

6.0 Results and Discussions

Clearly the returns of the financial time series process can be seen to be normally distributed as shown in the Figure 1 below.

For such a process, then fitting a polynomial by a plug –in approach, the function that appears in the diagram gives an asymptotically optimal bandwidth best approximated with a local linear polynomial.

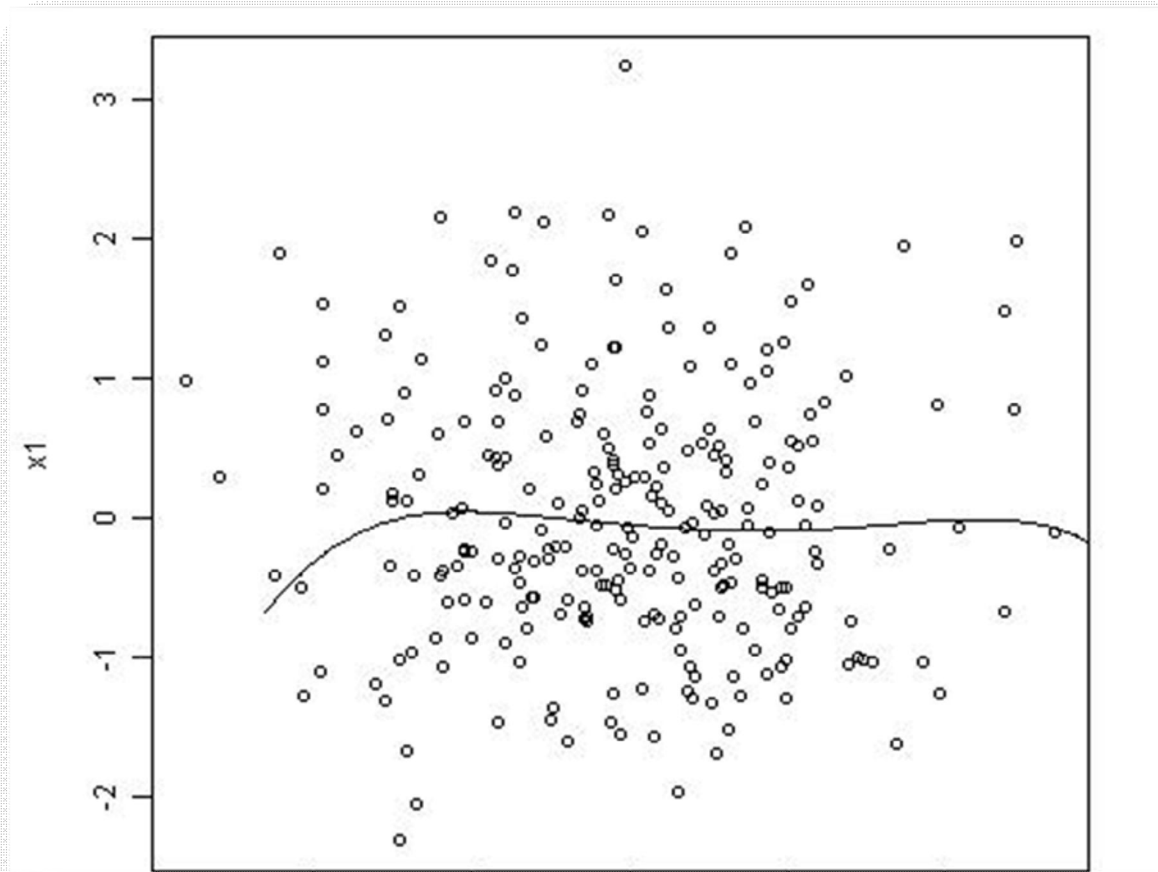

Figure 1: Returns of a financial time series

Fitting our returns data with a linear model with 2 coefficients (AR [1] process) for 250 training points produced the coefficients 0.9817942 that represented the y intercept and a gradient of 0.9727030. Hence the hypothesized model could be represented as $y = 0.9817942 + 0.9727030x$ where x represents the returns and y as the actual financial time series. The estimates for the standard errors for the coefficient above respectively gave 0.06867458 and 0.06262530 as the best variance bounds.

Again, fitting our returns data with a nonparametric regression local linear estimator to 250 training points to 2 variable(s) for 2 nonparametric regressor(s) using optimal bandwidth as given below.

Table 1: A table of optimal bandwidth and coefficients of a local linear polynomial on 250 points

		y
Bandwidth(s):	0.1896613	0.3412962
	x	
Bandwidth(s):	0.001552832	0.7407379
	0.999803300	0.2324291
	f	s
Coefficient(s):	0.9476291	0.958792
Residual standard error: 0.977513		
R-squared: 0.7270775		

Using the estimators fitted in the previous discussion, then the diagrams as shown below illustrate the power of local linear compared to the linear ordinary regression in trying to uncover the data generating mechanism for the returns data. As we can see, the vector of the bandwidth enables us to smooth the local linear estimator in all directions not necessarily equal but essentially uniformly at various values of the function. The final prediction error can be seen to be equals to 0.977513 and a coefficient of multiple determination of 0.7270775 that implies 73% of the returns are taken care of by the local linear estimator as the optimal lag vector. The remaining 27% lags can be described to be redundant. The coefficients in the local linear estimator are 0.9476291 and 0.958792 respectively which compared to an OLS as given above are slightly smaller

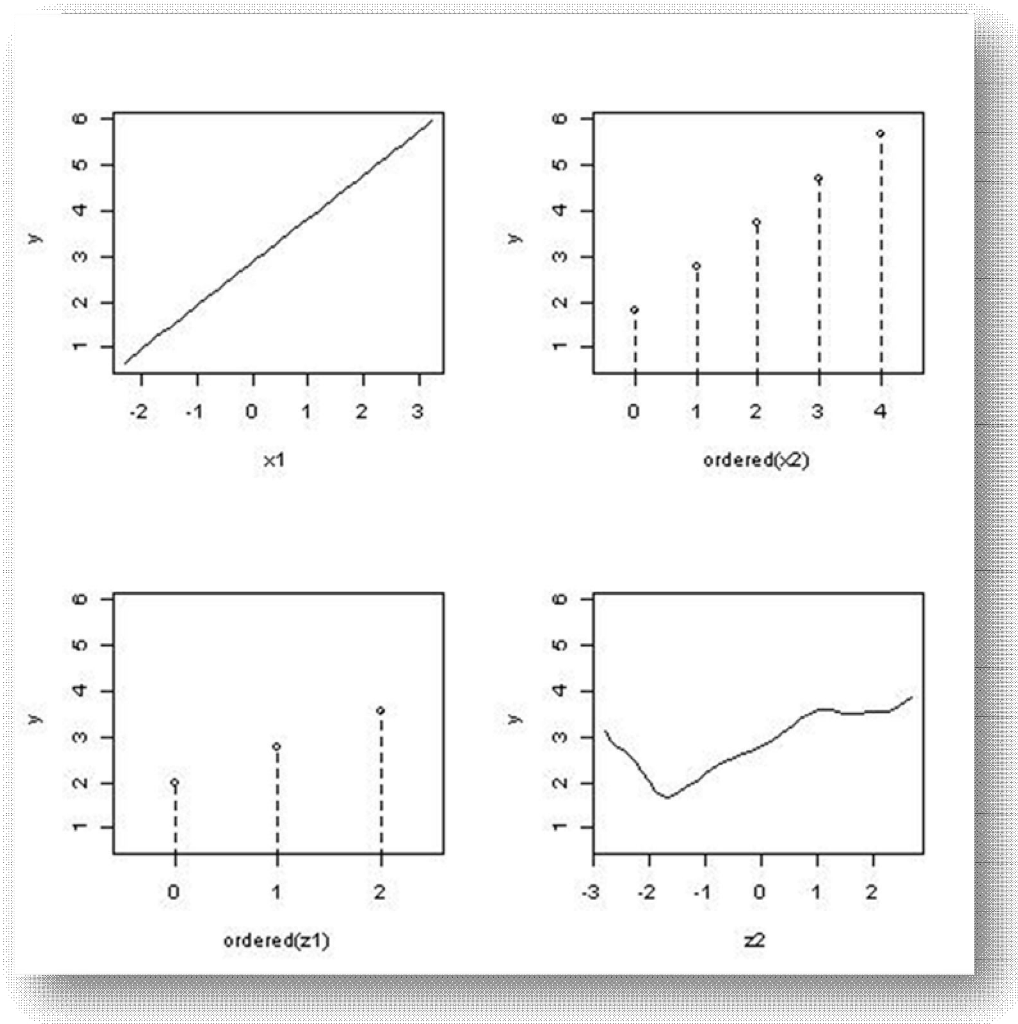

Figure 2: The plots of the OLS fitted and local linear polynomial

To plot regression surfaces with variability bounds constructed from bootstrapped standard errors, then the diagram below shows the uniform bands.

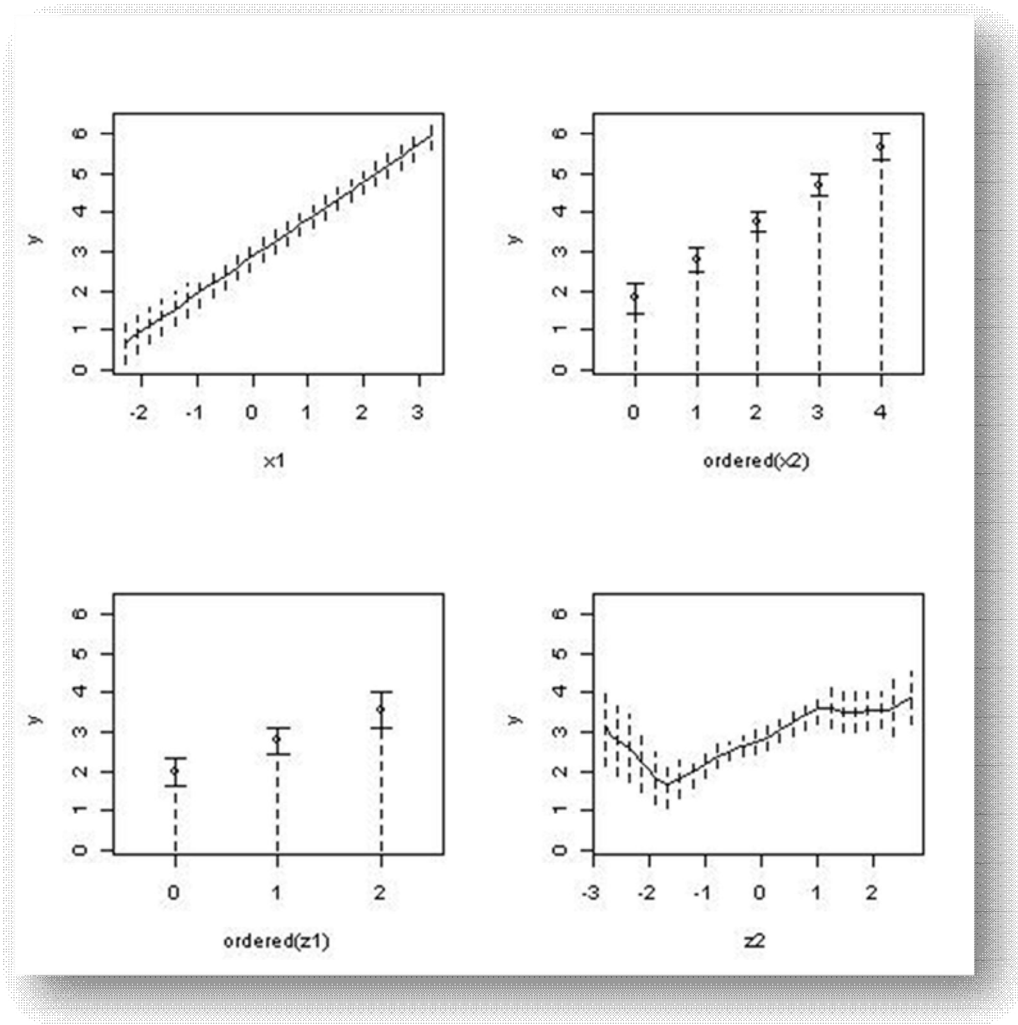

Figure 3: Bootstrapped estimators for both OLS and the local linear

7.0 Conclusions

In this paper we have shown that the local linear estimator out-performs the ordinary linear regression in mimicking the financial time series data generating mechanism. With an optimal bandwidth and lags, we are able to demonstrate that we can estimate the values of the quantile estimates for conditional mean and variance that are used to put a bound on the risk levels applied in financial risk management.

Also we have demonstrated the bootstrap confidence for both estimators in both the ordered and the unordered sets will give a uniform bound on the expected extends of the values of the estimates

References

Ait-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate. *Review of Financial Studies* ***9***, pp 385-426.

Baillie, R. T. and Bollerslev, T. (1992). Prediction in dynamics models with time dependent conditional variances. *Journal of Econometrics,* **52** , pp 91-113.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities, 71. *Journal of Poolitical Economy*, pp 637-653.

Claeskens, G. van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics,* **31**(6), pp 1852-1884.

Fama, E. (1970). Efficient Capital markets: A review of theory and empirical work. *Journal of Finance,* **25**, pp 383-417.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. London. Chapman and Hall.

Gospodinov, N. (2002). Median unbiased forecasts for highly persistent autogressive process. *Journal of Econometric*.

Hardle, W. (1990). Applied Nonparametric Regression. New York: Cambridge University Press.

Hardle, W. and Bowman, A. W. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *Journal of American Statistical Assiociation,* **83**(401*)*, PP 102-110.

Hardle, W. and Marron, J. S.. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Annals of Statistics,* **19**(2), pp 778-796.

Kim, J. H. (2001). Boostrap-after-boostrap prediction intervals for autoregressive models. *Journal of Business & Economics,* **19**, pp 117-128.

Li, Q. and Racine, J. (2007). Nonparametric Econometrics: Theory and Applications,. Princeton University Press.

Mandelbrot, B. (1963). The variation of certain speculative prices. *J.,Bus* , pp 394–419.

McMurry, T. and Politis, D. N. (2004). Nonparametric regression with infinite order flat-top kernels. *Journal of Nonparametric Statistics,* **16**(3-4) , pp 549-562.

Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science,* **4**, pp 141-183.

Miguel, J. A. and Olave, P. (1999). Forecast intervals in ARCH models: Boostrap verses parametric methods,. *Applied Economics Letters,* **6**, pp 323-27.

Neumann, M. H. Polzehl, J. (1998). Simulataneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics,* **9**, pp 307-333.

Pagan, A. R. and Ullah, A. (1999). Nonparametric Econometrics,. New York: Cambridge University Press.

Pascual, L., Romo, J. and Ruiz, E. (2000). Forecasting returns and volatilities in GARCH processes using boostrap. Working paper 00-68(31), Universidad.

Pascual, L., Romo, J. and Ruiz, E. (1998). Boostrap predictive inference for ARIMA processes. *Universidad Carlos III de Madrid, Spain* , Working paper 98-86.

Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance* **52**, pp 1973-2002.

Taylor, S. (2005). Asset Price Dynamics, Volatility, And Prediction. Princeton, NJ.: Princeton University Press.

Tsay, R S. (2005). Analysis of Financial Time Series, 2th Edition. New York.: John Wiley and Sons.

Wang, and Wahba, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison to bayesian confidence intervals. *Journal of Statistical Computation and Simulation,* **50**, pp 263-279.