ESTIMATION OF VARIANCE IN HETEROSCEDASTIC DATA

Titus K. Kibua

Statistics and Actuarial Science Department, Kenyatta University, Nairobi, Kenya
Email - kibua.titus@ku.ac.ke

Abstract: *Data which exhibit none constant variance is considered. Smoothing procedures are applied to estimate these none constant variances. In these smoothing methods the problem is to establish how much to smooth. The choice of the smoother and the choice of the bandwidth are explored. Kernel and Spline smoothers are compared using simulated data as well as real data. Although the two seem to work very closely, Kernel smoother comes out to be slightly better.*

Key Words: *Smoothing, Kernel, Spline, Heteroscedastic, Bandwidth, Variance.*

1. INTRODUCTION:

Let us have the observations $x_1, x_2, ..., x_n$ .The mean is given by $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $i = 1,2,...,n$.

The deviations of each observation away from the mean is given by $x_i - \bar{x}$, $i = 1,2,...,n$.

Now, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ has no meaning. To do away with this zero, we square, sum and average the deviations. This quantity is the variance. It is a constant. The problem is to find the variance between any one observation and the next. Thus we need to estimate the variance between x_1 and x_2 , x_2 and x_3 , …, x_{n-1} and x_n . Variance in this case is no longer a constant but a variable. We need to estimate this variable. In this work, analyzing regression data when the classical assumption of constant variance is violated is considered. If the variance function contains unknown (β) parameters, these must be estimated, perhaps by using classical methods such as: Least squares, Weighted Least squares, Maximum Likelihood, and Ridge regression among others. Least squares estimation has been used over the years. When the form of the distribution of the errors is known the method of maximum likelihood can be applied. Weisberg (1980) referred to this method as weighted least squares. The usual regression model makes four basic assumptions and the analysis is often based on a fifth one. The first four assumptions are that the model is correctly specified on average and that the errors are normal, identical and independently distributed (have the same distribution) and have constant variability. That is $E(y_i) =$ expected value of $y_i = f(x_i, \beta)$. Thus $y_i - f(x_i, \beta) = \varepsilon_i$. Variance $(y_i) = \text{var}(\varepsilon_i) = \sigma_i^2$. The errors ε_i have the same distribution for each given value of x . Given x_i the errors $\varepsilon_i = y_i - f(x_i, \beta)$ are independently distributed. However if any of these assumptions are violated the estimates obtained under the classical or usual assumption are not good. Therefore we hope to obtain better estimates of $\beta = (\beta_1, \beta_2, ..., \beta_p)$ when the estimation of the variance is incorporated.

Therefore, we need to investigate and incorporate the information about the variance estimates of the errors which are needed for better understanding of the variability of the data. In heteroscedastic regression models the variance is not constant. Often as in the case with the mean, the heteroscedasity is believed to be in functional form which is referred to as variance function. We try to understand the structure of the variances as a function of the predictors such as time, height, age and so on. Two procedures of estimating the variance function includes the parametric and nonparametric methods. The parametric variance function estimation may be defined as a type of regression problem in which we see variance as a function of estimable quantities. Thus, the heteroscedasticity is modeled as a function of the regression and other structural parameters. This function is completely known, specified up to these unknown parameters. Estimation of these parameters is what entails parametric methods.

However, for many practical problems the degree to which components of the statistical model can be specified in a parametric form varies drastically. When the model is miss – specified, the resulting model fit can be biased and the possibility for making wrong inferences exist. On the other hand, when part of the model is parametric fitting it is useful to do so in order to have a more analytical and traceable model that can be able to use traditional

inference techniques. Traditionally, the regression parameter β has been estimated using linear regression which has not given good estimates. To get better estimate of β there is need to incorporate the use of variance estimation. For the last twenty years, nonparametric method for estimating the variance function has been considered.

If n data points $\{x_i, y_i\}$, $i = 1,2,...,n$ have been collected, the regression relationship can be modeled as $y_i = \beta(x_i) + \varepsilon_i$, $i = 1,2,...,n$ with unknown regression function coefficients β and observation errors ε_i. A look at the scatter plot of x_i versus y_i does not always establish an interpretable relationship. The eye is sometimes distracted by extreme points. In such a situation one is interested in estimating the mean of a certain function. The aim of a regression analysis is to produce a reasonable approximation of β to the unknown response function. Reducing the observational errors, it allows interpretation to concentrate on important details of the mean dependence of Y on X.

This task of approximating the mean function can be done essentially in two ways. The quite often used parametric approach is to assume that the mean curve has some pre specified functional form, for example a line with unknown slope and intercept. As an alternative one could try to estimate β non parametrically without reference to a specified form. The term nonparametric thus refers to the flexible functional form of the regression curve. The question of which approach should be taken in data analysis was a key issue in the bitter fight between Pearson and Fisher in the twenties. Fisher pointed out that the nonparametric approach gave generally poor efficiency whereas Pearson was more concerned about the specification question.

2. BASIC IDEA OF SMOOTHING:

If β is believed to be smooth, then the observation at x_i, near x should contain information about the value of β at x. Thus it should be possible to use something like local average of data near x to construct an estimator of $m(x)$. Smoothing of a data set $\{x_i, y_i\}$, $i = 1,2,...,n$ involves the approximation of the mean response curve β in the regression relationship. The function of interest could be the regression curve itself, certain derivatives of it or functions of derivatives such as extrema or inflection points.

In the trivial case in which $\beta(x)$ is a constant, estimation of β reduces to the point of location, since an average over the response variables y yields an estimate of β. In practical studies though it is unlikely that the regression curve is constant, the assumed curve is modeled as a smooth continuous function of a particular structure which is nearly constant in small neighborhoods around x. It is not easy to judge from looking even at a two – dimensional scatter plot whether a regression curve is locally constant. A reasonable approximation to the regression curve $\beta(x)$ will therefore be any representative point close to the centre of this band of response variables. A quite natural choice is the mean of the response variables, near a point x. The local average should be constructed in such a way that it is defined only from the observations in a small neighborhood around x, since y – observations from points far away from x will have, in general very different mean values. The local averaging procedure can be viewed as the basic idea of smoothing. More formally this procedure is defined as $\hat{m}(x) = n^{-1}\sum_{i=1}^{n} w_{n_i}(x) y_i$, where $\{w_{n_i}(x)\}$, $i = 1,2,...,n$ denotes a subsequence of weights which may depend on the whole vector $\{x_i\}$, $i = 1,2,...,n$.

Every smoothing method described in this work is, at least asymptotic, of the form $\hat{m}(x) = n^{-1}\sum_{i=1}^{n} w_{n_i}(x) y_i$. Quite often the regression estimator $\hat{m}(x)$ is just referred to as a smoother and the outcome of smoothing procedure is simply called the smooth. Special attention has to be paid to the fact that smoothers, by definition average over observations with different mean values. The amount of averaging is controlled by the weight sequence $\{w_{n_i}(x)\}$, $i = 1,2,...,n$ which is tuned by a smoothing parameter. This smoothing parameter regulates the size of the neighborhood around x meaning a local average over too large a neighborhood would cast away the good with the bad. In this situation an extremely over smooth curve would be produced, resulting in a biased estimate $\hat{m}(x)$. On the other hand defining the smoothing parameter so that it corresponds to a very small neighborhood would not shift the chaff from the wheat .Only a small number of observations would contribute none negligibly to the estimate $\hat{m}(x)$ at x making it very rough and wiggly. In this case the variability of $\hat{m}(x)$ would be inflated. Finding the choice of smoothing parameter that balances the trade – off between over smoothing and under smoothing is called the smoothing parameter selection problem.

3. CHOOSING THE SMOOTHER:

Some of the smoothing techniques include Kernel, Spline, Locally weighted regression, Recursive Regressogram, Convolution, Median, Split linear fit and K-Nearest Neighbor among others. One of the most active research areas in Statistics in the last 20 years has been the search for a method to find the "optimal" bandwidth for a smoother. There are now a great number of methods to do this. Unfortunately none of them is fully satisfactory. Here a comparative study of the two mostly used and easy to implement smoothers is presented. The Kernel and the cubic spline smoothers .The comparison is preformed on a simulated data set. Looking at the kernel smoothing, a variety of kernel functions have been studied. Kernel smoothing technique is one of the simplest estimation which is straight forward to implement without further mathematical knowledge and it is understandable. The only problem is that the decision about the right amount of smoothing is crucial. The challenge in smoothing is to choose the best bandwidth that balances the desire to reduce the variance of the estimator (which needs lots of data points) yet capture significant small – scale features in the underlying distribution (which needs a narrow bandwidth). Every smoothing method has to be tuned by some smoothing parameter which balances the degree of fidelity to the data against the smoothness of the estimated curve. A choice of the smoothing parameter has to be made in practice and controls the performance of the estimators. One thing that has to be noted here is that the user of a nonparametric smoothing technique should be aware that the final decision about an estimated regression curve is partly subjective since even asymptotically optimal smoothers contain a considerable amount of noise that leaves space for subjective judgment. It is therefore of great importance to make such a decision in interaction with the data, which means that ideally one should have a computer resource with some sort of interactive graphical display. The main interest here is to show the best of these functions more specifically when we are considering kernel to be normal – the Gaussian density is given by $K(x)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}x^2\right],\ |x|<1$. However, both practical and theoretical considerations limit the choice of the kernels. Given the bivariate data $(x_i, y_i), i=1,2,...,n$ the smoothed value $\hat{m}(x)$ produced by a kernel function $K(x)$ can be given as

$$\hat{m}(x)=\frac{n^{-1}\sum_{I=1}^{N}K\left(\frac{x-x_i}{h}\right)Y}{n^{-1}\sum_{i=1}^{N}\left(\frac{x-x_i}{h}\right)},\ 0<x, x_i<1, i=1,2,...,n$$ Nadaraya [1964] and Watson [1964]

A common measure of "fidelity to the data" for a curve g is the residuals sum of squares $\sum_{i=1}^{n}(y_i-g(x_i))^2$ if g is allowed to be any curve unrestricted in functional form. Then this distant measure can be reduced to zero by any g that interpolates the data. Such a curve would not be acceptable on the ground that it is unique and that it is too wiggly for a structure oriented interpolation. The spline smoothing approach avoids this implausible interpolation of the data by quantifying the competition between the aims to produce a curve without too much rapid local variation. There are several ways to quantify local variation. One could define a measure of roughness based, for instance, on the first, second and so forth derivatives. In order to explicate the main ideas the integrated squared deviation is most convenient that is the roughness penalty $\int(g''(x))^2dx$ is used here to quantify local variation. Using these measures the weighted sum can be defined as

$s_\lambda(g)=\sum_{i=1}^{n}(y_i-g(x_i))^2+\lambda\int(g''(x))^2dx$ .Where λ denotes a smoothing parameter. The smoothing parameter λ represents the rate of exchange between residual errors and roughness of the curve g. The problem of minimizing $s_\lambda(.)$ over the class of all twice differentiable functions on the interval $[a, b]=[x_{(1)}, x_{(n)}]$ has a unique solution $\hat{m}_\lambda(x)$ which is defined as the cubic spline. The observations considered are in a small neighborhood of x since y observations far away from x will have in general very different mean values.

4. CHOOSING THE SMOOTHING PARAMETER:

The problem of deciding how much to smooth is of great importance in non parametric regression. Focus here is to find a good way of choosing the smoothing parameter for various smoothing methods. What conditions do we require for a bandwidth selection rule to be "good"?. First of all it should have theoretical desirable properties.

Secondly it has to be applicable in practice. Regarding the first condition there have been a number of criteria proposed that measure in one way or another how the estimate is to the true curve.

Before embarking on technical solutions of the problem it is worth noting that a selection of the smoothing parameter is always related to a certain interpretation of the smooth. If the purpose of smoothing is to increase the "signal to noise ratio" for presentation or to suggest a simple (parametric) model, then a slightly "over smoothed" curve with a subjectively chosen smoothing parameter might be desirable. On the other hand, when the interest is purely in estimating the regression curve itself with an emphasis on local structures then a slightly "under smoothed" curve may be appropriate. However, a good automatically selected parameter is always a useful starting (view) point. An advantage of automatic selection of the bandwidth for kernel smoothers is that comparison between laboratories can be made on the basis of a standardized method. Another advantage of the same lies in the application of additive models for investigation of high-dimensional regression data.

In obtaining the smooth curve the selection of the kernel function is not enough but rather the consideration of the bandwidth h. Several procedures of obtaining the bandwidth have been studied. Here the choice for the best bandwidth is to come up with several plots and select the bandwidth which outperforms the rest. Using large simulated data Kernel and cubic spline techniques are compared and the best smoothing parameter is therefore found by looking at the curves plotted. In the data sets, variation of the smoothing parameters is carried out until the best is found. The best smoothing parameter that is obtained here can be used straight forward to smooth any given data of any size instead of using trial and error methods.

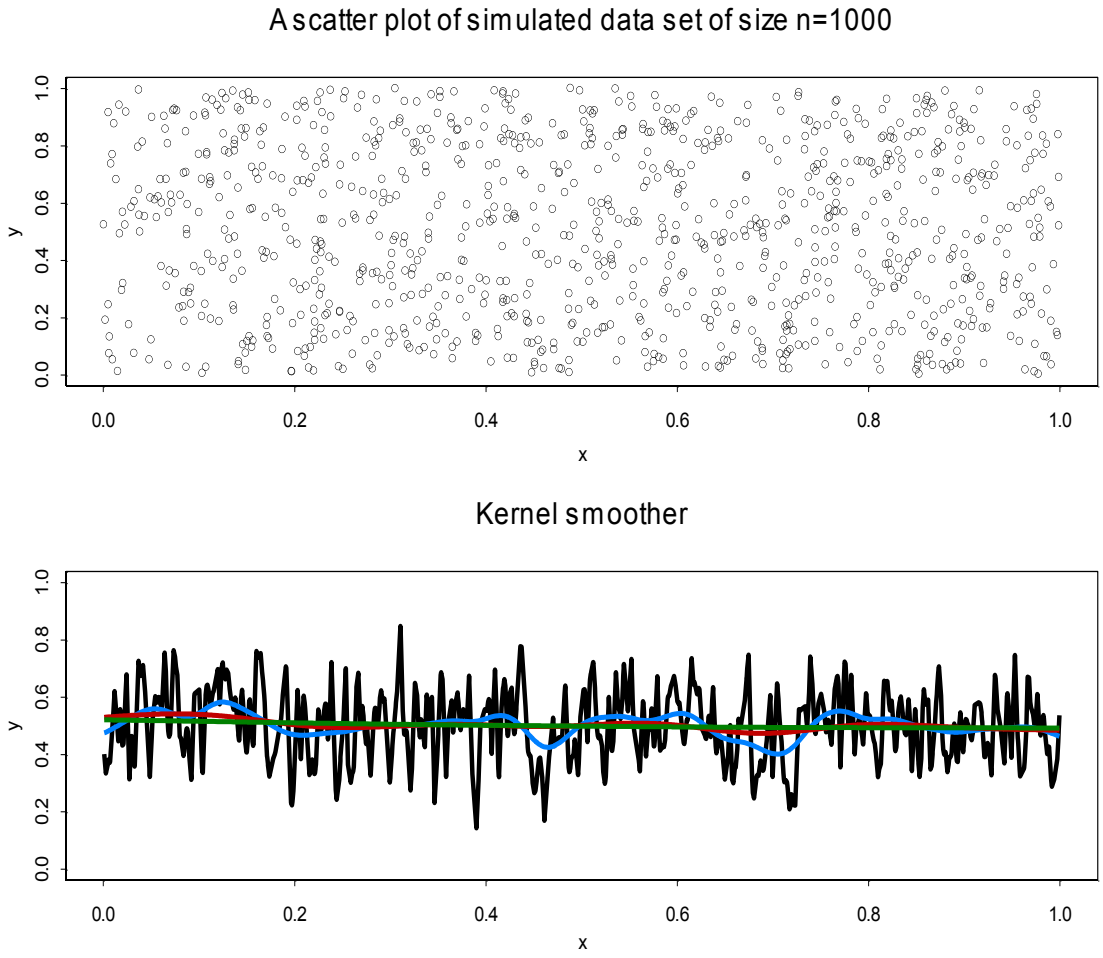

Figure 1

Figure 1: shows a simulated data of size $n = 1000$ data points with a kernel smoothing technique with smoothing parameters: $h = 0.004,\ 0.175,\ 0.3$ and 0.7. The blue curve with smoothing parameter $h = 0.175$ seems to have the optimal bandwidth. It outperforms the rest as the best smoothing parameter. Smoothing parameters less than 0.175 would under smooth the data and smoothing parameters greater than 0.175 would over smooth the data.

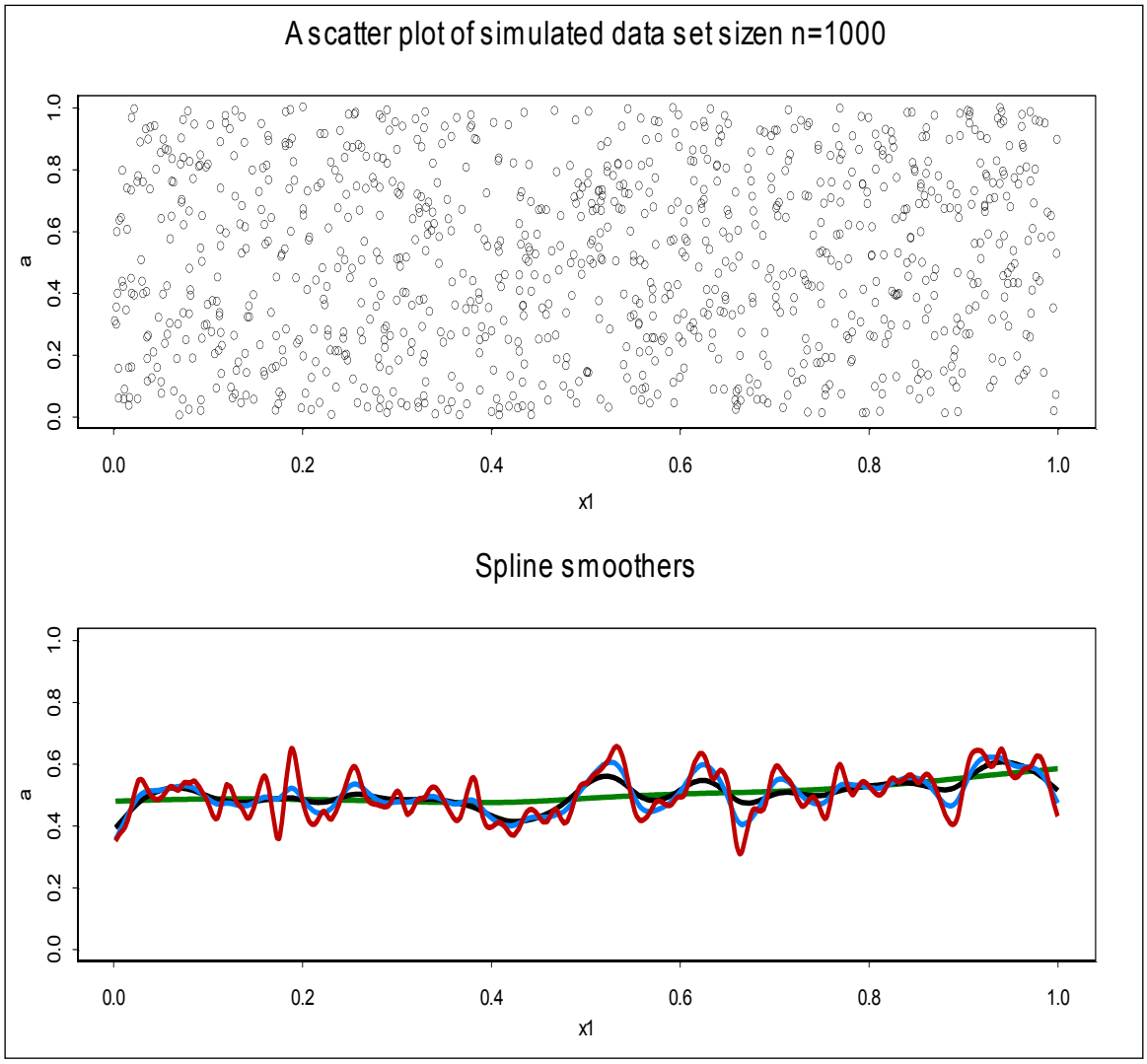

Figure 2

Figure 2: shows a simulated data of size $n = 1000$ data points with a spline smoothing technique with smoothing parameters: $\lambda = 5,\ 20,\ 35$ and 70 .The blue curve with smoothing parameter (bandwidth) $\lambda = 35$ seems to have the optimal bandwidth. It looks fairly smooth and hence it outperforms the rest as the best smoothing parameter.

5. EMPIRICAL STUDY (REAL DATA):

Here, it is assumed that the variance function is completely unspecified.
Let this variance function be denoted by $V(x_i, \beta)$ or $V\{c_i\}$. In estimating the variance function, residuals are used. The residuals are defined as $y_i - f(x_i, \hat{\beta})$.
Then the expectation of the squared residuals gives the estimate of the variance function given by $E(r_i^2) = E[y_i - f(x_i, \beta)]^2 \cong V(x_i, \beta)$. It is also possible to have the model in the design alone which is defined as $Var(y_i) = \sigma_i = V(c_i)$. Where $V(.)$ is unknown and $\{c_i\}$ is a set of identically and independently distributed random variables independent of $\{e_i\}$. In practical studies, to achieve the smooth conditions large data sets are use. The two smoothing techniques are applied on data obtained from the Nairobi stock exchange. Share volume in successive months over a period of 5 years 4 months of the Kenya commercial bank is studied. A regression model is constructed that relates to time (x_i) and (y_i) the share volume. The variables are: Time, Share volume in Kenyan shillings, Residuals and Residuals squared.

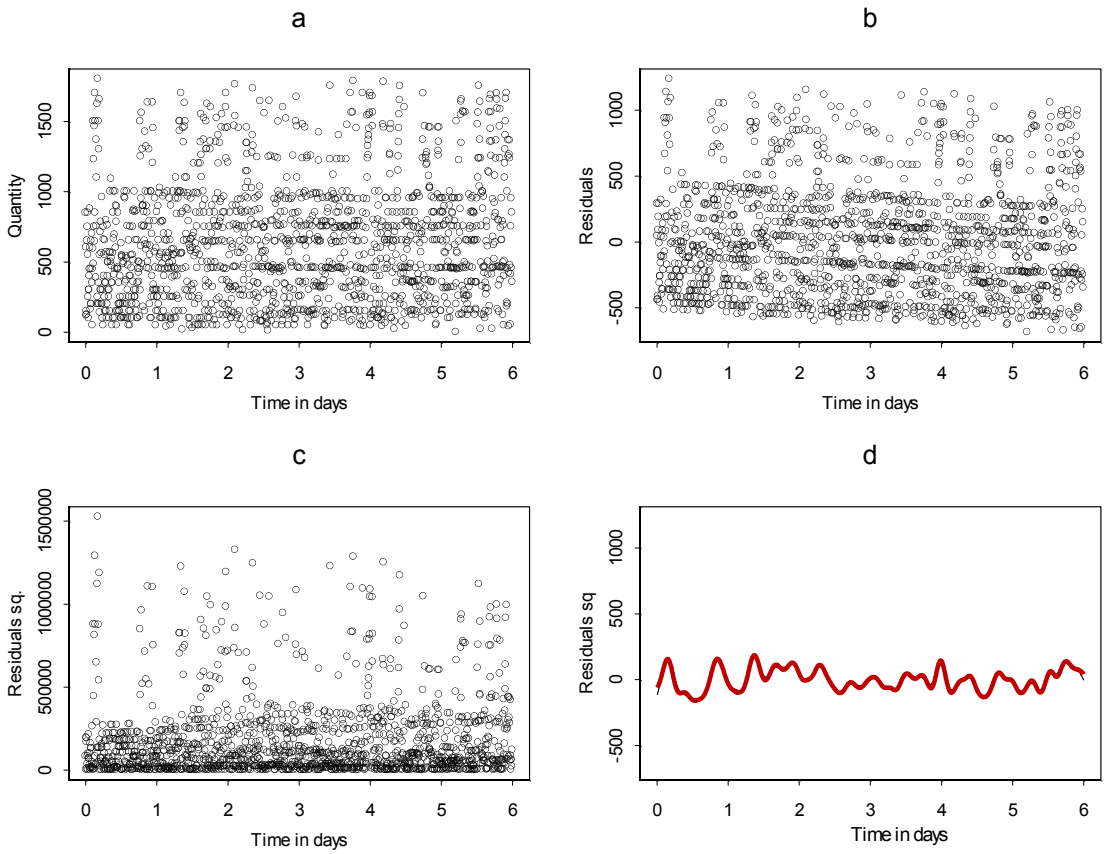

Figure 3

Figure 3: a) shows a scatter plot of share volume against time. b) Shows residuals squared against time. c) Shows residuals squared against time.
d) Shows both kernel and spline smooth of squared residuals against time.

6. CONCLUSION:

It is seen that the Kernel smoother produces the best estimate since its variance is less than that of the Spline smoother. From figure 3(d), it is seen that the spline is more on top around the middle and this clearly shows that its variance is more. There is a great difference along the boundaries (that is the beginning and the end) and the recommendation here is that more research work need to be done to find out why there is such much discrepancy.

REFERENCES:

1. Amemiya, T. (1977). A note on a heteroscedastic model. *Journal of Econometrics* 6, 365 – 370.
2. Buckley, M. J., Eagleson, G. K. and Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* 75, 189 – 200.
3. Cao, R., Cuevas, A. and Manteiga, W. G. (1994). A comparative study of sevsral smoothing methods in density estimation. *Computational Statistics and Data Analysis* 17, 153 – 176.
4. Carroll, R. J. and Rupert, D. (1988). *Transformation and weighting in regression*. New York: Chapman & Hall.
5. Carroll, R. J. and Ruppert, D. (1987). Variance function estimation. *Journal. of American Statistical Association* . 82, 1079-1091.
6. Carroll, R. J. and Ruppert, D. (1982b). Robust estimation in heteroscedastic linear models. *Annals of Statistics* 10, 429 – 441.
7. Cook, D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* 70, 1 – 10.
8. Diblai, A. and Bowman, A. (1997). Testing for a constant variance in a linear model. *Statistics & Probability Letters* 33, 95 – 103.
9. Gasser, T.,Muller, H. G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society* B 47, 238 – 252.
10. Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. Journal of the Royal Statistical Society, series B, 51, 3 – 14.

11. Hardle, W [1993]. *Applied Nonparametric Regression*. Cambridge, U.K, Cambridge University press.
12. Kibua, T. K. (2006). Bandwidth selection in smoothing functions. *East African Journal of Statistics* 1(2), 161 – 174.
13. Kibua, T. K. (2006). Some results on nonparametric estimation of variance function. *East African Journal of Physical Sciences* 7(1/2), 51 – 64.
14. Kibua, T. K. (2005). Some asymptotic theory for variance function smoothing. *East African Journal of Statistics* 1(1), 9 – 22.
15. Lian, H., Liang, H and Carroll, R. J. (2014). Variance function partially linear single – index models. *Journal of the royal statistical society*, series B, 77, 171 – 194.
16. Muller, H. G. and Stadtmuller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Annals of Statistics* 15, 610 – 625.
17. Muller, H. G. and Stadtmuller, U. (1987). Variable bandwidth kernel estimators of regression curves. *Annals of Statistics* 15, 182 – 201.
18. Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications* 9, 141 – 142.
19. Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society* B 47, 1 – 50.
20. Silverman, B. W. (1986). *Density Estimation for Statistical and Data Analysis*, London: Chapman and Hall.
21. Watson, G. S. (1964). Smooth regression analysis. *Sankhya* A 26, 359 – 372.
22. Weisberg, S. (1985). *Applied Linear Regression*. Wiley, New York.

WEB REFERENCE:

- https://documentslide.com/documents/ebook-pdf-statistics-applied-nonparametric-regression1.html