How to Objectively Rate Investment Experts in Absence of Full Disclosure? An Approach based on a Near Perfect Discrimination Model

Patrick Wessa[1]

Abstract

The result of this investigation is an operational model that can be used to accurately identify real stock market time series. In other words, if we are presented with a collection of blinded time series (real-life time series and simulated Random-Walks) then the proposed model will allow us to discriminate between both categories. In addition, it is shown that the type II error of this model quickly converges to zero as the time series length increases. The most remarkable feature of this model is its simplicity: a (bias-reduced) logistic regression with a single exogenous variable (the kurtosis p-value) based on the Quasi Random-Walk model that relates returns of equity and the entire market in times of large market returns. This model can be used as an objective rating benchmark for the models that are used by hedge funds to identify the stocks that should be used in a market neutral arbitrage strategy of long and short positions. In addition, it allows independent auditors to objectively evaluate the added value of statistical and technical analysis techniques that are often used in investment decisions. A rating mechanism that is based on the proposed benchmark, provides valuable information about the investment strategy even in absence of full disclosure.

1   Introduction

My main argument in this paper is that real stock market time series exhibit fundamental, testable differences when compared to Random-Walk series that are - by definition - Fama-efficient. The discrimination between both types of time series is based on a simple logistic regression model with the p-value of the small sample Kurtosis as exogenous variable. The discrimination quality of the model is extraordinary which makes the model an effective benchmark (and challenge) for any discrimination model that is used by hedge funds in a market neutral investment approach.

To provide empirical evidence for this statement I follow a three step procedure. First, a theoretical introduction and formal definition of an alternative model for the efficient market hypothesis (the so-called Quasi Random-Walk model) is provided. Second, a comprehensive dataset is collected, and complemented with simulated Random-Walk time

[1] Integrated Faculty of Business and Economics (Lessius, dept. of Business Studies), KULeuven Association; patrick@wessa.net

series that represent the efficient market hypothesis. The third step involves the development of a very simple (bias reduced) logistic regression model that discriminates between both types of time series. It is shown that the discrimination quality is near perfect (in terms of type I and type II errors) when the time series under investigation has more than 500 observations.

All computations that are presented in this paper were performed with the R language (R Core Team, 2007).

2   The model

2.1   Introduction

The application of the Quasi Random-Walk (QRW) model in finance was proposed by Airoldi (2001) based on previous research and the following observation of Cizeau et al. (2001): "Empirical evidence shows that for large market returns almost 90% of the equities have the same sign as that of the market...".

Airoldi formulates his model for N equities S_i for i = 1, 2,..., N that exhibit movements dSi = $\pm s$ following a Quasi Random-Walk with "hopping probabilities" P_{dSi} that may depend on previous market returns Airoldi (2001). Since this dependence is associated with periods of turmoil we can introduce two separate states of the market:

$$h= 1 : Pas, (MC"**)) = \pm \qquad (M \,^\wedge\,) \qquad\qquad \, _,,$$
$$h = 0 : P_{0Si} \qquad\qquad = I \qquad\qquad\qquad\qquad _{K \, , \, \}}$$

with $M\text{«-**)}$ = £ E f=! $^{,,,,(t-A1)}$, 19 (M) I < 1.

Obviously, if h = 0 then P_{dSi} behaves like an ordinary Random-Walk and does not depend on market movements. In this case, the probability of an increase (or a decrease) is exactly | and does not depend on anything else. This implies that there is no way to predict the outcome of future returns (h = 0 corresponds to the efficient market). As a consequence, the time series of equity returns is not autocorrelated when h = 0. This automatically leads to the conclusion (based on the central limit theorem) that equity returns must be normally distributed (as long as h = 0).

On the other hand, P_{dSi} in (2.1) depends on the previous market movements if h = 1. As pointed out before, in times of large market returns there is a high probability that the sign of equity returns is the same as that of the market. Therefore an additional term that describes the dependence on past market returns is included in the equation. As a consequence, the time series of equity returns (for periods where h = 1) is autocorrelated and fat-tailed instead of normally distributed.

The QRW model does not explain when, how, or why large market returns occur. It also does not describe the exact specification of the function g (.) that relates large market returns to the returns of equities. This however is not a problem for the development of a discrimination model as will be explained in the following sections.

2.2 Bias reduced logistic regression model

I expand on Airoldi's idea, and introduce a simple logistic relationship $f = e^{l+\&X}$ where $P\{h = 1) = jjj$ and where X represents a "discriminating statistic" that is used to model the states of the market. Parameter estimates 7,[7] are obtained with the Bias Reduced Logistic Regression method introduced by Firth (1992) which yields adequate variance estimates in cases of near perfect discrimination - unlike ordinary logistic regression estimation (Firth (1993), and Heinze et al. (2002)).

The key to discriminate between $h = 1$ and h = 0 is provided by the observation that the sign of large market returns can be associated with the sign of equity returns (Cizeau et al., 2001). Moreover, it has been well established in empirical research that large market returns are more likely to occur than what we could expect in efficient markets. Hence, if market returns exhibit fat tails then we should expect excess kurtosis (as compared to the normal distribution).

In addition, the QRW model predicts that autocorrelation measures are bad predictors for a discrimination model. The reason is simply due to the fact that the observations of a QRW is a mixture of two different states where only state h =1 displays any meaningful autocorrelation pattern. Any attempt to measure this pattern requires the investigator to filter the time series to exclude the randomness in state h = 0.

In this study we employ the following (sample) measure of kurtosis (Borghers et al., 2006a) for the equities $S_i = (xi, x_2, ...x_n\}$ with i = 1, 2,..., N;

$$K = l \quad \frac{(n-l)n}{(n-2)(n-3)(n-4)j^\wedge} \quad s \quad J \quad J \quad \frac{3(ra-2)^2}{(n-3)(n-4)} \quad K \quad)$$

with S $\quad$ _ $_{Ti}$f and vj = V Ina;,- and r, = $\quad$ y
$\sum_{j=2}$ $\quad\quad\quad\quad$ $_{j=2}$

The kurtosis measure K_i in (2.2) can be used for large and small samples. The standard error of K_i is

$$SK = \backslash I \quad \frac{4 Un - 1)^2 - 1) s^2}{(n-4)(n+4)} \quad\quad (2,3)$$

with $\quad s = \frac{6(nn1)n}{(n-3)n(n+2)}$ $\quad$. The test statistic is z $=_{SK}^\wedge <- N(0,1)$.

Based on the research of Van der Vorst (2005), I use the p-value of K_i as the only discriminating statistic X in the logistic regression. This means that (given a population kurtosis of K_i) the type I error $P(K_i = 0 \mid H_0 : K_i = 0)$ is transformed into a probability of the state of the market P (h =1). In order to be able to make clear-cut decisions a threshold-based rule must be introduced to identify Quasi Random-Walk Stock Market Returns.

The threshold u is chosen such that the type I error is P (P (h =1) > u | h = 0) < 0.05. Obviously, the type II error is hoped to be small enough to allow for practical applications.

2.3   Applications

It is well known that a large proportion of hedge funds employs a so-called "market neutral" investment strategy. This particular strategy is often employed because it theoretically promises to yield positive returns that are independent of the market trend. Basically, this is achieved by holding simultaneous long and short positions in different subsets of a (pre-specified) universe of stocks or other assets. In theory, such strategies allows the hedge fund to generate positive investment returns that are independent of the overall market trend.

The underlying decision process of this type of hedge fund is based on some sort of discrimination model or methodology. For obvious reasons, this will be kept secret from the outside world, making the hedge fund resemble a black box that can only be rated indirectly, through the analysis of historical performance.

Non disclosure of the actual discrimination model or methodology is problematic for the hedge funds and their clients. The hedge fund managers are unable to adequately demonstrate the qualities of their investment approach and the clients have no information about the risk they are actually taking by using the hedge fund investment vehicle. Therefore there is an obvious need for an objective rating methodology that can be used by clients to assess and compare hedge fund investment strategies without the need to disclose any secret information about the underlying discrimination model.

How can the proposed model be used to rate hedge funds that employ a "market neutral" investment strategy?

At first, it is important to identify the fundamental problem that the hedge fund manager is facing in the market neutral strategy. To put things simply, the manager faces the following list of decisions:

- define a universe of tradable items (for instance all large-cap stock in the U.S.A.)

- determine the investment horizon (often this ranges from 2-30 days)

- create a discrimination model that classifies the stocks in three piles (long, short, and neutral)

The fundamental (most difficult) problem is to appropriately identify the stocks that have to be attributed to the short, long, and neutral category. The hedge fund's portfolio consists of a combination of simultaneous long positions (from the long pile stocks), and short positions (from the short pile stocks) during the investment horizon. The hedge fund manager hopes to achieve an arbitrage profit that is caused by rising prices of long pile stocks and decreasing prices of short pile stocks.

Unfortunately, the hedge fund manager is unable to disclose information about the discrimination model or methodology that is used because it would otherwise lose its competitive advantage. At the same time, the manager is not allowed (due to legal restrictions) to advertise the quality of the investment strategy.

In second instance, one could look at the decision problem of the manager of an "ordinary" fund. Here, the investment strategy is subject to many restrictions that are imposed by law and senior management. Many funds will manage a basket of assets that contains cash, bonds, and a portfolio of stocks that is similar to a market index. From time to time, these managers will decide to change the composition of the basket or stock

portfolio. These changes are typically small and "cautious". Therefore the stock-picking problem is not as acute as in the case with hedge funds. Nevertheless, many financial institutions and institutional investors often base their decisions on the analysis of experts such as: technical analysts, econometricians, and fundamental analysts. Even though there is a legal requirement to be transparent to the outside world, the details of such analysis and resulting decisions are not readily disclosed.

In both cases there is no full disclosure. This is problematic because clients have no way to assess the risk that is associated with their investments. In addition, there are seems to be a growing concern about the black-box nature of the hedge fund industry which explains why some are now openly calling for more regulation.

If the proposed model in (2.1) turns out to be true and if the suggested exogenous factor proves to be effective in discriminating between the states of the market *(h = 1* and h = 0) then it is possible to create fast algorithms that make a preselection of equities (from the universe of all equities under consideration) that show a high logistic regression probability that h = 1. This is of particular importance for hedge funds, employing the market-neutral investment strategy.

Furthermore, any model that selects equity from the universe and assigns them to either a long or short position portfolio, must have a statistical discrimination quality that is at least as good as the quality of the proposed model. In other words, the power of the proposed logistic regression is a benchmark for any equity selection algorithm when feeded with simulated and true stock market time series. This allows us to create a rating mechanism that allows us to compare ratings from different (hedge) funds even if they operate in different markets or circumstances. The logic behind this is that the rating of the fund depends on its ability to discriminate between Random-Walk and real Stock Market time series relative to the discrimination quality of the benchmark model. If the model's performance is not too sensitive with respect to place or time then the relative rating promises to be fair. In any case the fund's performance should be at least as good as the performance of the proposed benchmark.

3   Dataset

I collected 66 index time series from Yahoo! Finance (2007) about various important markets and made them available in an on-line archive (see Figure 1 and Statistical Computations at FreeStatistics.org (2007)). Each time series consists of daily closing prices ranging from the first trading day in 1995 (earliest date) until the last trading day of December 2006. The lengths of the time series vary due to obvious reasons.

The data collection represents a variety of important markets including:

- several important stock exchanges of the U.S.A.

- U.S.A. bonds, notes, and treasury bills

- gold and silver
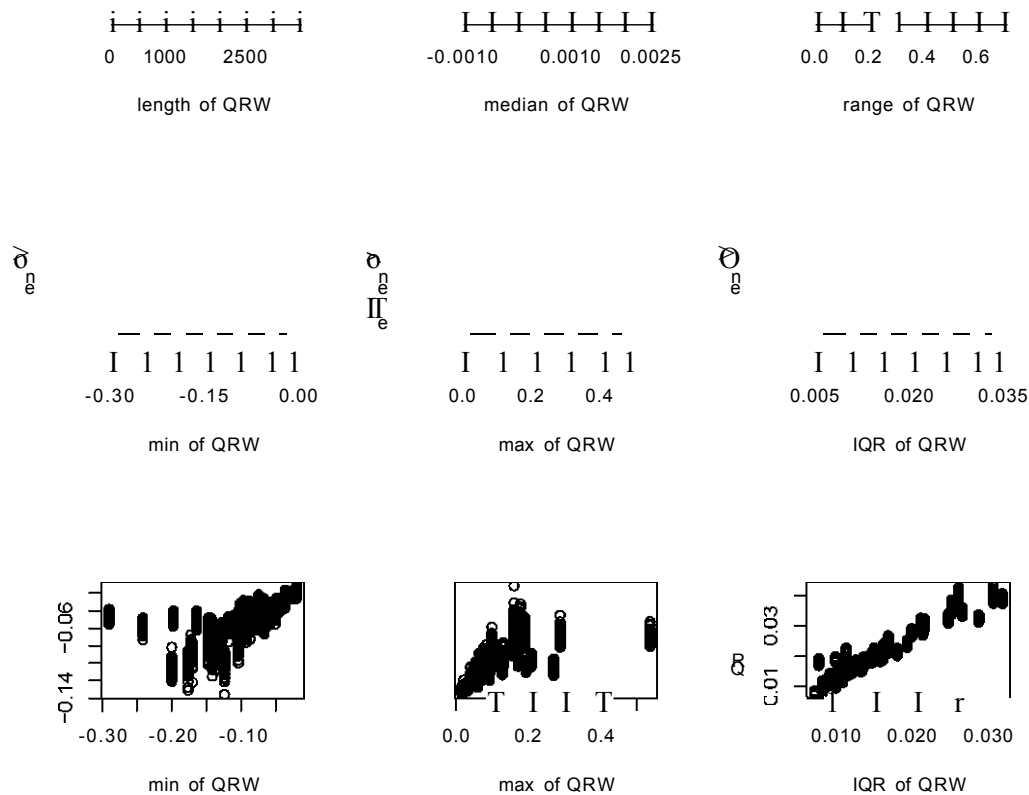
- several well-known stock exchanges in Europe and Asia

Figure 1: Descriptive statistics - dataset
http://www.freestatistics.org/blog/date/2007/0ct/18/26qptxvtt3arzpz1192711431.htm

Figure 1 shows the histograms about various statistics of the log returns of all ob-served time series (denoted QRW). It can be observed that many series contain more than 2500 trading days, and only a small minority of series have less than 1000 observations (variable "length"). The descriptive statistics about extreme values (c.q. range, mini-mum, maximum, and interquartile range) have highly skewed distributions. In addition, the variation about these statistics is substantial, indicating that the sample of index series exhibits a variety in terms of extreme returns.

For every time series I simulated 20 Random-Walks (denoted RW) that are - by def-inition - known to satisfy the criteria of weak form-efficiency. Each of the 20 simulated series has the same mean, and standard deviation as the original time series. More pre-cisely, the formal definition of the symmetric Random-Walk is $(1 - B)lnY_t = e_t$ with $BY_t = Y_{t-i}$.

This definition contains a logarithmic transformation because it allows the simulated stock market prices to have the property that local volatility increases with the local mean. This property is realistic as can be easily verified by the use of the Standard Deviation-

Mean Plot (Borghers et al., 2006b). In addition, it should be noted - as pointed out by Larry Weldon - that the definition of the RW is symmetric. This means that the probability of an increase or decrease of stock market prices is exactly 50%. This comment makes sense because a weak drift (c.q. a constant term) might be added to the definition to account for the long term trend on the stock markets. In this investigation the constant term is assumed to be zero because of the following reasons:

- the drift component in a simulated RW of a maximum of 500 daily observations is negligible

- if the drift component is tested in real stock market time series (c.q. QRW) it is almost never significant

- if a drift component would be included in the simulation it would result in a skewed distribution and non-zero autocorrelation of the log returns - a property that would be easily detected by a discrimination model

Figure 1 shows that the variation of the deviation of extremes (minimum or maximum) between the simulated and original time series converges as the extreme is closer to zero. The interquartile range (IQR) of simulated and observed time series however are highly correlated, even if the variation of the observed time series is large. In other words, the RW and QRW series are similar when only the mid 50% of observations are considered. The differences between both types of series are clearly located in the tails of the log return distribution.

4 Analysis

4.1 Introduction

In applied research it is common to use market time series of daily returns that are relatively short (between 6 months and 2 years). The underlying assumption is that only the recent history contains any information of interest. For this reason, we analyze time series that are obtained by introducing a minimum n_{min} and a maximum number of observations n_{max}. The methodological approach outlined below however, is generic and can be reproduced for any set of integer values with $n_{min} < n_{max}$.

The maximum number of subsets M_{max} is defined by the following relationship $M_{max} = floor\left(\frac{max(length(S_1), length(S_2),...,length(S_M))}{n_{min}}\right)$ where $floor(x)$ returns the nearest integer value that is equal to or smaller than x, and $length(S_i)$ returns the number of observations of the time series.

The actual subsets (for every equity i and actual length j) are defined by the following relationship $M^{\wedge} = floor \;\wedge\hat{a}\hat{i}hi^{\wedge}l^{\wedge}j$ for $i = 1,2,..., N$ and $j = n_{min}, n_{min} +1,..., n_{max}$.

The p-values p_{j_q} of the previously defined kurtosis measures K_{ijq} were computed for the log returns r_{ijq} of every actual subseries $S_{ijq} = \{x_{j(q-1)+1}, X_{j(q-1)+2}, ... X_{j(q-i)+j}\}$ for $i = 1, 2,..., N$ and $j = n_{min}, n_{min} + 1,..., n_{max}$ and $q = 1, 2,..., M_j$. Due to the large number of subseries the computational effort is quite substantial. It is therefore not feasible to use many more than 20 Random-Walks per observed time series. Several
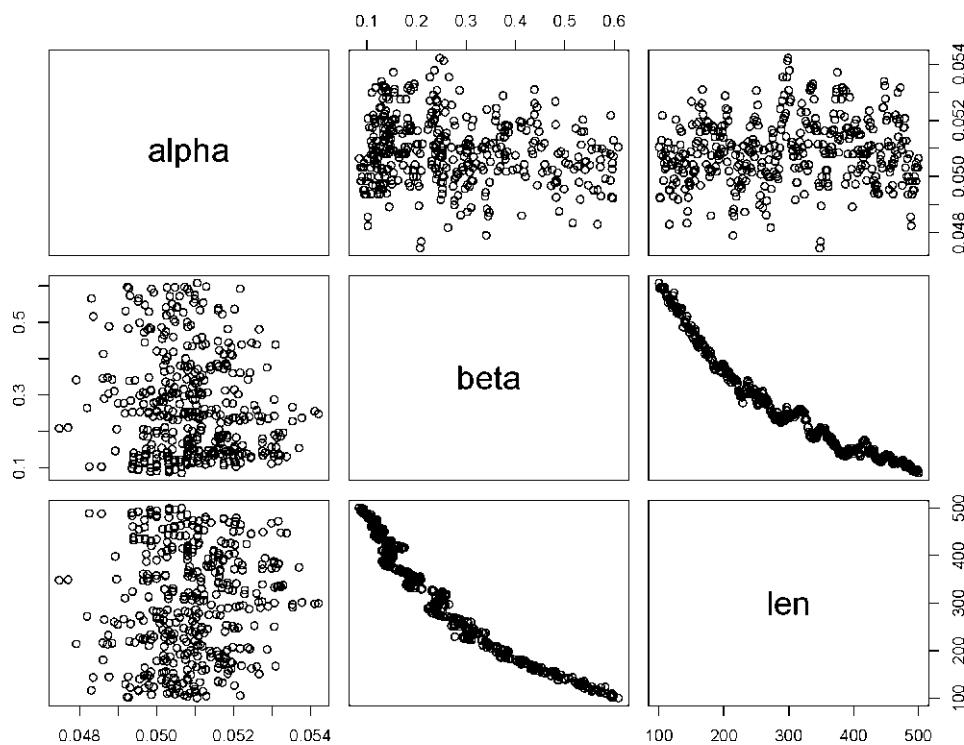
Figure 2: Type II errors depend on the time series length given a fixed type I error of 5%.

experiments applied to a subset of the data with 50 Random-Walks instead of 20 yielded results that were very similar to those portrayed below.

In order to be able to create a model that discriminates between (efficient) Random-Walks and true financial time series (Quasi Random-Walk) the kurtosis p-values p'_{ijq} were computed based on the subseries of 20 simulated log returns r with the same mean and standard error as the original time series S_i.

For every length $j = n_{min}, n_{min} + 1, ..., n_{max}$ a bias reduced logistic regression was computed where the binary status variable h = 0 or $h = 1$ is explained by the corresponding p-values of the kurtosis p'_{ijq} and p_{ijq}. For obvious reasons, the degrees of freedom of every regression depends on the number of subsets that can be created of each length j. All bias reduced logistic regressions however, have highly significant [5] parameters (t-stat between -15.9 and -27.5).

4.2   Type I & II errors

From the results presented below it can be concluded that the discrimination quality of the logistic regressions strongly depends on the length of the time series j = $n_{min}, n_{min} + 1,..., n_{max}$ with $n_{min} = 100$ and $n_{max} = 500$. The relationship between the type I error (alpha), type II error (beta), and the time series length (len) is shown in Figure 2.

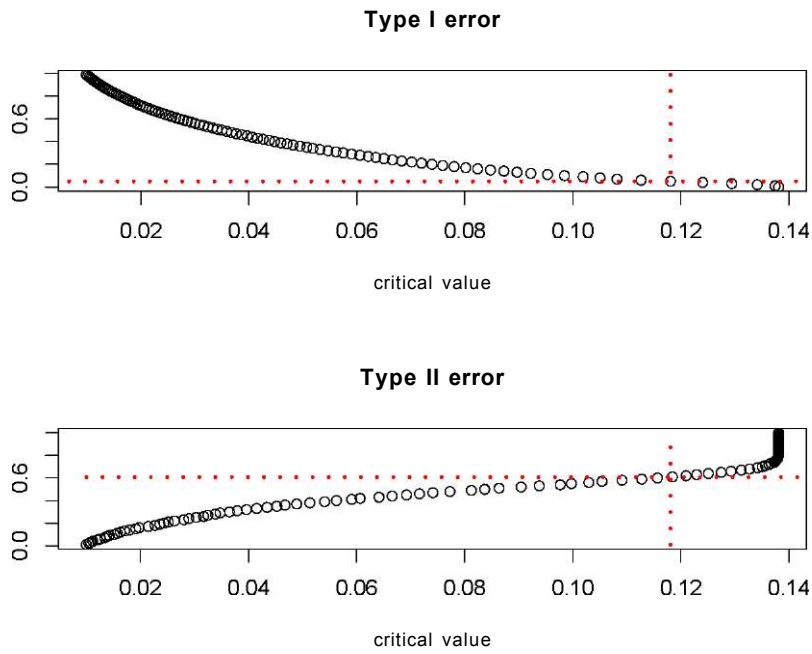The type I error is 5% up to a rounding-off error: the exact values range from 4.5%-

Type I error

Type II error

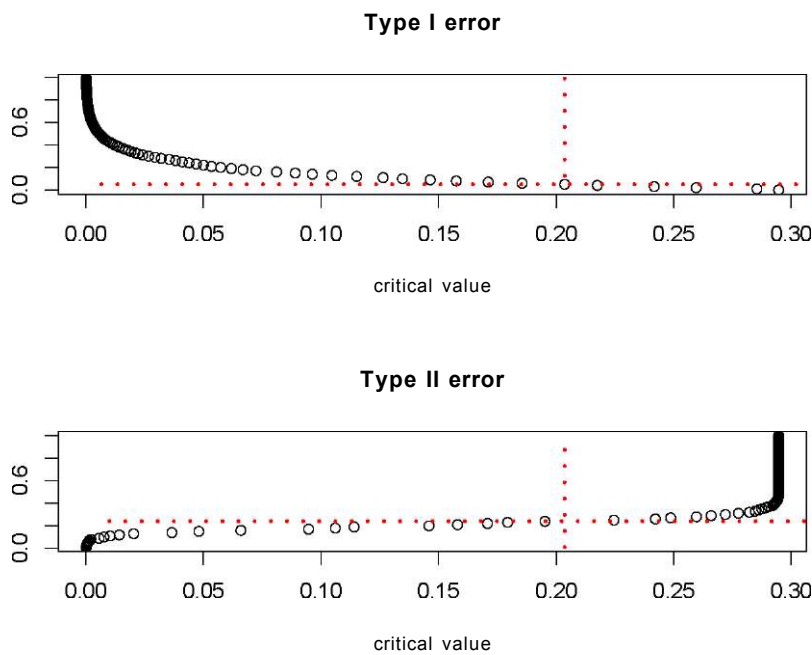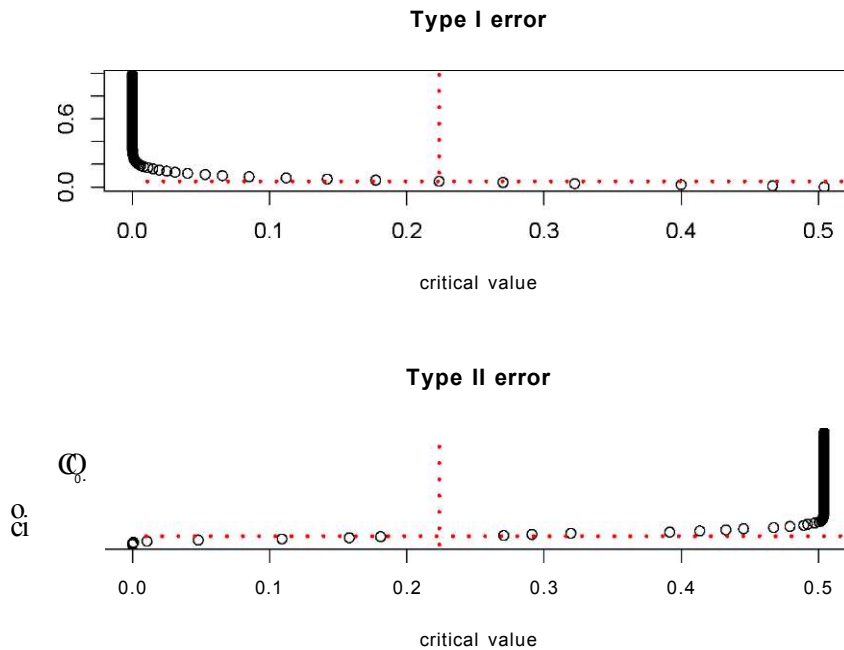**Figure 3:** Type II error of kurtosis-based discrimination model in relationship with type I error (time series length = 100).

Type I error

Type II error

Figure 4: Type II error of kurtosis-based discrimination model in relationship with type I error (time series length = 300).

Type I error

critical value

Type II error

critical value

Figure 5: Type II error of kurtosis-based discrimination model in relationship with type I error (time series length = 500).

5.5%. The pattern that emerges from the sequence of type I errors is caused by the varying lengths of the original data. The same applies to the wiggly pattern in the sequence of type II errors. More importantly, there is a clear (non-linear) negative relationship between the length of the time series ("length") and the type II error ("beta"). Also note that there is a very sharp reduction in type II error (from a maximum of 60.8% to 8.44%) with increasing length. The shortest time series in this investigation have 100 observations. For obvious reasons we should not expect the kurtosis to be significantly different from zero in every single instance. If a type I error of 5% is maintained, the statistical power is expected to be rather low. Figure 3 shows the relationship between the chosen type I error (alpha), its corresponding critical value (c.q. threshold u), and the type II error. The power of the statistical test is only 39.2% when a 5% type I error is selected - this is inadequate for practical applications.

With increasing number of observations, the power of the test is improved. For example, a time series length of $j = 200$ yields a power of 61.6%. The result for $j = 300$ is shown in Figure 4 - the power is 75.9%. The trade-off between type I & II errors is much smaller than before (c.q. the type II error curve is much flatter). Finally, Figure 5 shows the relationship between the chosen type I error, its corresponding critical value, and the type II error for time series with a length of 500 observations (the maximum length in this investigation). Now there is virtually no trade-off between type I & II error. In addition, the statistical power is 91.6% and could be increased further by using time series that are longer than 500 observations.

The type II error is much less sensitive to changes in its critical value u than with short time series. Increasing the length of the time series under investigation has a tremendous

and favorable effect on the model's prediction quality. This is not the case for other criteria - such as autocorrelation measures - that are traditionally used in empirical research. For example, one might consider an autocorrelation-based discriminating statistic based on x_{ijq} = EK=I IP ($V\ln x_t$, *Vlnx—*)| for t = j *(q - 1)* + 1, j (q - 1) + 2 , j (q - 1) + j, $i = 1, 2,..., N$, $j = n_{min}, n_{min} + 1, ..., n_{max}$, and q = 1, 2,..., Mj. This discriminating statistic yields a power between 8.8% (for j = 100) and 22.73% (for j = 500) when used in the logistic regression instead of the kurtosis p-value. The power slowly improves with increasing length j, but is not suited for any real-life application.

This observation may - at least in part - explain why the empirical research literature about financial market efficiency has not come to an unambiguous final conclusion. The argument here is that autocorrelation-based criteria are inefficient measures of market inefficiency because they cannot capture the dynamical properties of the financial market in a parsimonious way. Traditional (c.q. non-robust) autocorrelation statistics are ill-suited measures of inefficiency in the presence of large returns (outliers).

On the other hand, there is strong empirical evidence - additionally supported by the Quasi Random Walk theory - that the kurtosis p-value is a near-perfect discriminating factor with type II errors that quickly converge towards zero with increasing length. It also explains the phenomenon that was observed in Figure 1: the deviation of extremes (minimum and maximum) between the simulated Random Walk and the original series diverge in the tails of the log return distribution.
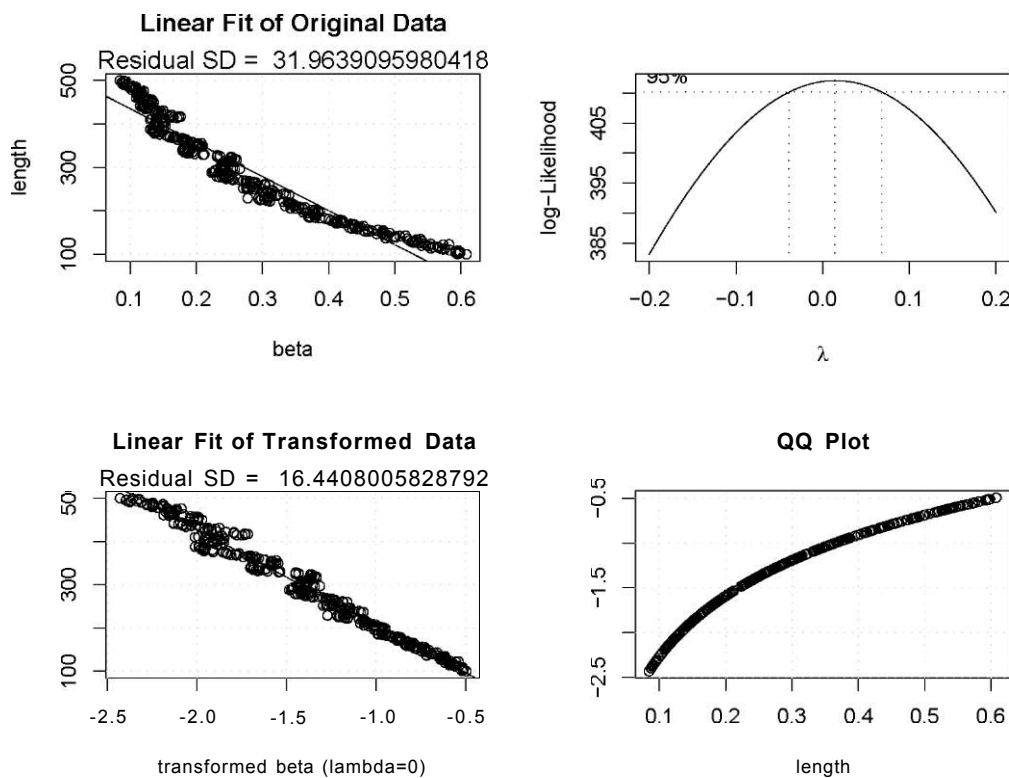

Figure 6: Relationship between length and Type II error (criterion: Kurtosis p-value).

4.3   Simple nonlinear models

Assuming a fixed Type-I error of 5% throughout, the relationship between the Type-II error Bj and the length j of the time series can be described with a simple nonlinear model j = a_1 + ßiA(Bj) with j = u_m $n,,$ $"I"$ $!)\bullet\bullet\bullet)$,$"u_{max}$ and with $\pm A_y(Bj)$ $\frac{Bj-1}{\wedge}$ (for A = 0) and A(Bj) = *log(Bj)* (when A   0). Figure 6 shows that the Maximum Likelihood estimate for A — 0, hence the loglinear model specification is the optimal Box-Cox transformation A(.) that linearizes the relationship between both variables of interest. In addition, the least squares fit is quite high ($Adj.R^2$ = 0.9798) and the parameters $\acute{c}x_1$ = -10.657 and β_1 = -216.244 highly significant. This loglinear model is an operational equation that computes the time series length that is required for any desired Type-II error B*: j^* = -10.657 - 216.244 log(B*). Suppose that the desired Type-II error B* = 15% then it is easy to obtain the required length j* = 400 (rounded to the nearest integer).

Another simple nonlinear model can be used to relate the required time series length to the optimal threshold value uj^* = a_2 + ß$_2$A(j). Figure 7 illustrates the Maximum Likelihood optimum for A — -0.66. Least Squares Estimation yields highly significant parameters and goodness of fit *(Adj.R^2 = 0.969).* The operational equation is uj* = —3.50777 + 2.50951^{j}_0 g^{+1} for j = n_{min}, n_{min} + 1, …,n_{max}. If we require a time series length of 400 observations then u$_{400}$* — 0.22.
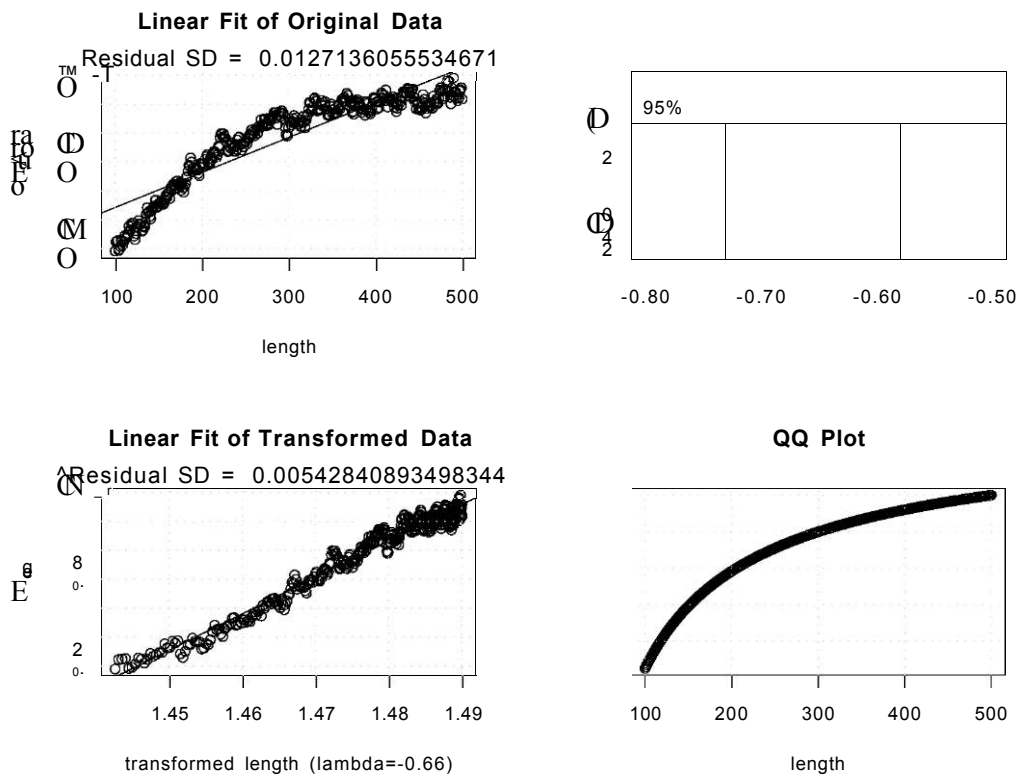

Figure 7: Relationship between threshold (w) and length (criterion: Kurtosis p-value).

The relationship displayed in Figure 6 can be recomputed for the autocorrelation-

based discriminating statistic - the results are shown in Figure 8. It is interesting to compare the discrimination performance of both statistics. A simple extrapolation of the model displayed in Figure 8 yields a required time series length of 1998 observations to obtain a desired Type II error level of 15% - this is about 5 times the length that would be required when we would use the kurtosis p-value.
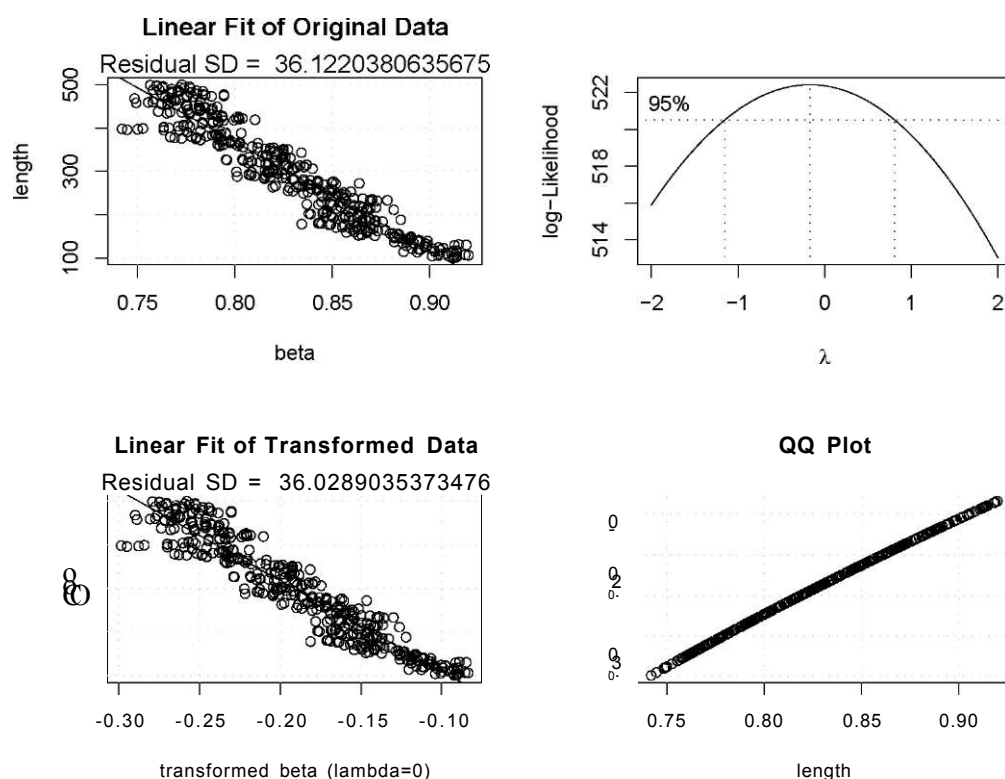

Figure 8: Relationship between length and Type II error (criterion: Autocorrelation).

5   Conclusion

The proposed benchmark model is remarkable because it is simple and promises to predict well across geographical regions and time. Therefore it is possible to compare a (hedge) fund's ability to discriminate between Random-Walk and real Stock Market time series (that are presented in a blinded experiment) with the benchmark model's performance. The objective rating of the (hedge) fund is expressed as the degree by which it outperforms the benchmark. This definition effectively allows us to compare (hedge) funds from different regions, and working under different circumstances. The applicability of this rating methodology is obviously limited to the cases where the fund under investigation is using some sort of discrimination model that helps them in some sort of stock-picking decisions.

Based on the evidence presented in this paper it is my believe that such a rating mechanism could be fairly easily implemented to systematically provide comparative, scientific

ratings about a fund's ability to make adequate (stock-picking) investment decisions with-
out the need to disclose any sensitive information to third parties.

Acknowledgements

References

[1] Airoldi, M. (2001): Correlation structure and fat tails in finance: A new mechanism. *Risk Management & Research,* Italy (Intesa-Bci Bank).

[2] Borghers, E., and Wessa, P. (2006): Descriptive statistics - skewness and kur-tosis (peakedness) for small samples - test statistic 1. *Xycoon.com (website),* http://www.xycoon.com/peakedness_small_sample_ test_l.htm, accessed 2006-12-23

[3] Borghers, E. and Wessa, P. (2006): Time series analysis - ARIMA models - non stationary time series. *Xycoon.com (website),* http://www.xycoon.com/non_stationary_time_series.htm, accessed 2006-12-23

[4] Cizeau, P., Potters, M., and Bouchaud, J.P. (2001): Correlation structure of extreme stock returns. *Quantitative Finance,* **1,** 217-222

[5] Firth, D. (1992): Bias reduction, the Jeffreys prior and GLIM. In Fahrmeir, L., Francis, B.J., Gilchrist, R., and Tutz, G. (Eds.): *Advances in GLIM and Statistical Modelling,* 91-100. New York: Springer.

[6] Firth, D. (1993): Bias reduction of maximum likelihood estimates. *Biometrika,* **80,** 27-38

[7] Statistical Computations at FreeStatistics.org (2007): Office for Research Development and Education.
http://www.freestatistics.org/blog/date/2007/Oct/18/26qptxvtt3arzpz1192711431.htm/ accessed, 2007-10-18

[8] Heinze, G. and Schemper, M. (2002): A solution to the problem of separation in logistic regression. *Statistics in Medicine,* **21,** 2409-2419.

[9] R Development Core Team (2007): R: A language and environment for statistical comput-ing, http://www.R-project.org/, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0

[10] Van der Vorst K. (2005): De identificatie van Quasi Random Walk processen, Unpub-lished Master thesis (Dutch), Lessius Dept. Business Studies, Integrated Economics Faculty, KULeuven Association

[11] Yahoo! Inc. (2007): *Yahoo! Finance (website),* http://finance.yahoo.com, accessed 2007-01.