Journal of Business and Economics, ISSN 2155-7950, USA
July 2013, Volume 4, No. 7, pp. 620-633
© Academic Star Publishing Company, 2013
http://www.academicstar.us

Why Logistic Regression Analyses Are More Reliable Than Multiple Regression Analyses

Jianing Fang

(Marist College, Poughkeepsie, New York, USA)

Abstract: Medical scientists have been using logistic regression analyses for their empirical research for quite some time. However, a majority of the previous accounting or finance-related empirical studies were based entirely on results from certain forms of multivariate regressions analysis. The reliability of these findings is subject to question. By definition, all forms of multiple regressions rely critically on some assumptions of the quality of the test data. One of the major problems is that most of the financial data often violate some, and in many cases, all of these assumptions. This paper will discuss the advantages and disadvantages of both the multiple regression analyses and the logistic regression analyses for empirical research. The main goal of this article is to promote the utilization of logistic regressions in addition to any applicable multivariate analysis to provide the much needed verification and reliability of empirical study results.

Key words: multiple regression; logistic regression; efficient market hypothesis

JEL code: C35

1. Introduction

For a long time, researchers have utilized various forms of multivariate analyses as a statistical tool in most accounting or finance-related empirical studies. Since these studies are based solely on the results of some form of multiple regressions, the reliability of their findings is subject to question. By definition, all forms of multiple regressions rely critically on the assumptions of linearity, constant variance, absence of special causes, normality, and independence of the test data. The problem is that most of the financial data often violate some, and in many cases, all of these assumptions.

In the medical field, researchers have utilized logistic regressions to analyze their test data for decades. Medical doctors are dealing with life and death. We accountants and financial analysts are dealing with companies' or people's wealth or livelihood—not equally important, but important enough. So, if logistic regression analysis is reliable enough for medical scientists, this statistics tool must be safe enough for counting beans too. Therefore, taking guidance from medical researchers, the main goal of this study is to promote the utilization of logistic regressions, in addition to any applicable multiple regression analysis, to provide the much needed reliability of empirical study results. This study will use an example with actual market closing indices on which the author conducted dual tests on the same set of test data by running both the multivariate regressions and logistic regressions. While the multiple regression results provide the necessary statistics as well as reference or

Jianing Fang, DPS, Associate Professor of Accounting, Marist College; research areas: XBRL, IFRS, tax, fair trade, research methods. E-mail: jfang@ambras.com and jianing.fang@marist.edu.

comparison with prior studies, the logistic regression results confirm the reliability of the empirical findings.

2. Literature Survey

Long before the debate on the efficiency of the global security markets, the efficient market hypothesis (EMH) was one of the most hotly contested topics among academic and finance professionals worldwide. By now the literature is so broad that it is practically impossible to review all of the relevant literature in this article. Instead, only the papers that are the most relevant to this study are discussed. The debate of market efficiency was started in 1900 by Bachelier. By analyzing the movements of commodity prices, Bachelier presents convincing evidence that commodity speculation in France is a “fair game” because the price movements follow the pattern of a “random walk”. This means that neither buyers nor sellers could expect to make significant profits.

The term “random walk” was first used, not in the field of finance, but in the science journal Nature. Pearson and Rayleigh (1905) describe an interesting but perplexing question: if a drunk were lost in a vacant field, what would be the most efficient search pattern to find the drunk at some time later? There would be no way to tell where he or she would end up because the drunk would be expected to wander randomly around the field without any direction or destination. Therefore, there would be no rational efficient search pattern. Osborne (1959) is credited as the first to evoke the “efficient market” controversy that has since stimulated an enormous amount of research and heated discussion (Lorie, Dodd, and Kimpton, 1985). However, the efficient market hypothesis was first proffered by Fama (1965, 1970). His work motivated the search for a viable economic theory for security investment. In his 1970 article, Fama also gave a thorough review of all the then- existing theoretical and empirical studies on the subject.

2.1 Related Anecdotal and Empirical Studies on EMH

Innumerable anecdotal and empirical studies followed the work of Fama. Fama, Fisher, Jensen and Roll (1969) showed that the equity market reacts quickly to new information. They analyze 940 stock splits from January 1927 to December 1959. By running a natural logarithm regression on the monthly returns, they find that stock splits have often been followed closely by dividend increases, indicating that these companies were enjoying abnormally good business, generally during boom periods. Thus, a stock split can be a valuable indicator that the managers of a firm are optimistic about future business, profit, and cash flow prospects. The caveat is that the authors just “assume…the usual assumptions of the linear regression” are satisfied (p. 4). The authors conclude that the results of their study support the weak form EMH because the information regarding a stock split is fully reflected in the share price very quickly.

Recently, EMH on the global scale has become a hot topic. In an attempt to explain the uniformity of the free fall in international security markets during the 1987 crash, King and Wadhwani (1990) develop a contagion model to analyze how “mistakes” or errors in valuation in one market can be transmitted to other nations’ security exchanges despite the many notable differences in terms of market control legislations and mechanisms, as well as different economic, political, social and other relevant backgrounds. They construct the model first with two markets and then expand to scenarios with many different markets for both overlapping and non-overlapping trading hours. They test the contagion model with data from security exchanges of New York, London, and Tokyo[1] for the eight-month period around the infamous black Monday—July 1987 to February 1988. King et al. find that: “the trading of stocks in one market per se affects share prices in other markets, that is, share prices

[1] The total market capitalization of these three markets represented about 80% of the world total then.

respond both to public information about economic fundamentals and to share-price changes elsewhere" (p. 29). This is because investors deduce information from share price movements in other countries for their investment decisions. Thus, their contagion model hypothesis posits that a "mistake" in one security market can be transmitted to other nations' markets.

In her dissertation, Guo (1990) provides a detailed study of the Hong Kong Stock Exchange (HKSE)—its history, government policy, public finance, monetary supply, tax system, and other relevant background. She describes the HKSE as "small" but "active". After a comprehensive review of the EMH, she tests the weak form and semi-strong form EMH on HKSE with data for 40 randomly selected stocks listed on the exchange, as well as the Hang Seng Index from November 1, 1984 to November 1, 1988. Guo conducts frequency distribution tests, independence tests, residual analysis, return correlation analysis, stability tests, range tests, and the goodness of fit tests for the distributions during different time periods and for dividend and bonus announcements, respectively. She finds that both the "random walk" model (the weak form) and semi-strong form EMH are supported, and that "the results of the tests on the Hong Kong market are similar to the results found by Fama (1970) for the U.S. stock market" (p. 45).

Guo (1990) also studies the transmission of stock market movements among Hong Kong, the United States, Japan, and the United Kingdom. She performs VAR (vector autoregression) tests on daily market indices of these four security exchanges (HIS, CRSP, NSA, and FTIO, respectively) for the 19-year testing period from 1970 to 1988. The empirical evidence shows that the U.S. market is the most influential in terms of innovation transmission.[2] Guo finds that the HKSE is highly vulnerable to external factors with the largest influence from the U.S. market.

The most recent scholarship includes an empirical analysis by Borgas (2010) who tested the daily closing values of stock market indexes for UK, France, Germany, Spain, Greece and Portugal to test the weak form EMH from January 1993 to December 2007. Using a runs test and joint variance ratio tests, her test results suggest mixed evidence on EMH among the sample stock exchanges. Guidi (2010) examines the day-of-the-week effect on stock market prices using daily closing price indices for the Italian stock market index (MIB) for the time period covering January 4, 1999 through March 5, 2009. He utilizes various forms of regression analyses such as GARCH-M and UVR. His test results do not support the weak form EMH because investors can earn above-normal returns by following the trend of the past price movement of the securities. Nevertheless, he acknowledges that his tests are only reliable "if there [are] no serial correlations among returns" (p. 21).

2.2 Logistic Regression Analysis

The foregoing not only reviews the EMH basic theory as well as the supporting and contradicting empirical tests, but also demonstrates that most of the prior studies rely only on some form of regression analysis. Empirical study results based solely on any form of multiple regressions is not reliable. By definition, all forms of multiple regressions rely critically on the assumptions of linearity, constant variance, absence of special causes, normality, and independence of the test data.[3] The problem is that most of the financial data often violate some, and in many cases, all of these assumptions. Therefore, a contribution of this study is that the author conducts dual tests on the same set of test data by running both the multiple regressions and logistic regressions. While the multiple regression results provide the necessary reference or comparison with prior studies, the logistic regression results

[2] 46.63% for U.K., 47.25% for Japan, and 37.25% for Hong Kong in a 20-day horizon. These rates changes for different test range of 2, 5, 10, and 20-day horizon.

[3] Outliers due to one-time situations have been removed from the data.

provide confirmation of the reliability of the empirical findings.

The literature on logistic regression is large and growing rapidly. Books that cover aspects of logistic regression include Cox (1970), Breslow and Day (1980), Kleinbaum, Kupper, and Morgenstern (1982), Schlesselman (1982), Mansfield (1994), and Hair, Anderson, Tatham and Black (1995), and Lachin (2008), but all except Lachin (2008) do not place their central focus on this subject. Fortunately, some of the techniques for application of this method and interpretation of the results may be found in statistical and financial studies. When the author started his research about fifteen years ago, there were some works by the medical scientists (Olivier, Blake, Steed and Salgado 1978) but hardly any from the business side. Today, there are countless studies utilizing logistic analysis in the medical field, and more and more finance and business papers are testing this reliable tool include Kotha, Rajgopal, and Venkatachalam (2004), Blumenschein, Blomquist, Johannesson, Horn, and Freeman (2008), Kolasinski and Kothari (2008), Ellingsen, Johannesson, Lilja and Zetterqvist 2009), Lampe (2011), Mayew and Venkatachalam (2012), and Avnet, Pham and Stephen (2012).

3. Economic Theories and Statistic Tools

The basic economic theory this study is based on is the contrarians' view of the ***efficient-market hypothesis***. Stocks are sold when they are considered out of favor or over-valued based on some calculation utilizing one of (or a combination of) the security analysis theories and/or techniques. The same stocks are bought because other investors take the exact opposite position, probably after running some different or similar analyses. The ***Logistic Indicator*** (Fang, 2005) is developed under the belief that stock prices are changing constantly as new information become available. But all this balancing and re-balancing takes time, especially in the global market. And a small foreign exchange often looks upon and follows the leadership of the global markets. Most scholars and practitioners believe, and the author concurs, that so far, Wall Street has been providing such leadership.

3.1 Review of Applicable Statistic Models

The statistic tools used to develop the ***Logistic Indicator*** are the applicable techniques of multivariate analysis. To define multivariate analysis is a rather taxing and complicated task. For starters, it refers to all statistical methods that simultaneously analyze multiple measurements on each individual or object under investigation. Any simultaneous analysis of more than two variables can be loosely considered multivariate analysis. Specific techniques include but are not limited to: multiple regression and multiple correlation, multiple discriminant analysis, principal components analysis and common factor analysis, multivariate analysis of variance and covariance, canonical correlation analysis, cluster analysis, multidimensional scaling conjoint analysis, and logistic regression (Dillon and Goldstein 1984; Hair, Anderson, Tatham and Black, 1995), plus the new bispectral test (Rusticelli, Ashley, Dagum and Patterson, 2008). Two of these techniques—multiple regression and logistic regression—will be used in this study for hypothesis testing.

3.2 Multiple Regression Analysis

Multiple regression analysis is a statistical technique that can be used to analyze the relationship between a single dependent variable (criterion) and several independent variables (predictors). The objective of multiple regression analysis is to use the several independent variables whose values are known to predict the single dependent value the researcher wishes to know. The reliability of the technique is critically reliant on the following assumptions:

- Linearity—the relationships between the dependent variable (Y) and its independent variables (Xs) are linear.

- Constant variance—the variance of Y is constant for all values of the Xs.
- Special causes—outliers due to one-time situations have been removed from the data.
- Normality—Xs are normally distributed when hypothesis tests and confidence limits will be used.
- Independence—Xs are uncorrelated with one another.

A more important reason for using multiple regression analysis instead of simple regression analysis is that "if the dependent variable depends on more than one independent variable, a simple regression of the dependent variable on a single independent variable may result in a biased estimate of the effect of this independent variable on the dependent variable" (Mansfield, 1994, p. 510).

When a dependent variable is a function of more than one independent variable, the observed relationship between the dependent variable and any one of the independent variables may be misleading because the observed relationship may reflect the variations in other independent variables. Since these other independent variables are totally uncontrolled, they may be varying in such a way as to make it appear that this independent variable has more or less of an effect on the dependent variable than in reality. To estimate the true effects of this independent variable on the dependent variable, we must include all the independent variables in the regression; that is, we must construct a multiple regression.

$$\mathbf{Y}= \boldsymbol{a} + \boldsymbol{b}\mathbf{X} + \boldsymbol{e} \quad \text{(simple regression)} \tag{1}$$

$$\mathbf{Y}= \boldsymbol{a} + \boldsymbol{b}_1\mathbf{X}_1 + \boldsymbol{b}_2\mathbf{X}_2 + \boldsymbol{e} \quad \text{(multiple [2] regression)} \tag{2}$$

$$\mathbf{Y}= \boldsymbol{a} + \boldsymbol{b}_1\mathbf{X}_1 + \ldots\ \boldsymbol{b}_n\mathbf{X}_n + \boldsymbol{e} \quad \text{(multiple [n] regression)} \tag{3}$$

The model underlying multiple regression analysis is essentially the same as simple regression analysis where Y is the observed value of the dependent variable (conditional mean of dependent variable) and is assumed to be a linear function of more than one independent variable ($X_{n)}$. The error term is signified by *e*. As in the case of simple regression, it is assumed that the expected value of *e* is zero, that *e* is normally distributed, and that the standard deviation of *e* is the same regardless of the value of $X_{1,}$ $X_{2,}$ or $X_{n.}$ Also, the values of *e* are assumed to be statistically independent.

3.2.1 Least-Squares Estimates of the Regression Coefficients

The first step in multiple-regression analysis is to identify the independent variables and to specify the mathematical form of the equation relating the expected value of the dependent variable to these independent variables. Same as in the case of simple regression, these constants are estimated by finding the value of each one that minimizes the sum of the squared deviations of the observed values of the dependent variable from the values of the dependent variable predicted by the regression equation (Brown, 1990).

3.2.2 Multiple Coefficient of Determination

The coefficient of determination can be used to measure how well a regression equation fits the data. When a multiple regression is calculated, the multiple coefficient of determination is used for this purpose. The mathematical presentation is:

$$\boldsymbol{R^2} = \textbf{\textit{variation explained by regression/total variation}} \tag{4}$$

Which means the R^2 measures the proportion of the total variation in the dependent variable that is explained by the regression equation. The positive square root of the multiple coefficient of determination is called the ***multiple correlation coefficient*** and is denoted by **R**. It, too, is sometimes used to measure how well a multiple-regression equation fits the data (Amick and Walberg, 1975).

3.3 Logistic Regression Analysis

Regression methods have become an integral component of any data analysis—science, medical, finance, etc.

—concerned with describing the relationship between a response variable and one or more explanatory variables. It is often the case that the outcome variable is discrete, taking on two or more possible values. Over the last two decades the logistic regression model has become, in many fields, the standard method of analysis in this situation.

It is important to understand that the goal of any analysis using this method is the same as that of any other model-building technique used in statistics: to find the best fitting and most parsimonious yet technically reasonable model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. These independent variables are often called *covariates.* The most common example of modeling, and one assumed to be familiar to most people, is the usual linear regression model where the outcome variable is assumed to be continuous.

What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is *binary* or *dichotomous.* This difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression. Thus, the techniques used in linear regression analysis will motivate the author's approach to logistic regression. In the following section, both the similarities and differences between logistic regression and linear regression will be discussed with two simple examples:

3.3.1 The Multiple Logistic Regression Model

Given a collection of *n* independent variables (*x*) which will be denoted by the vector **X'** = (*x1, x2, …, xn*). Let us denote the resulting probability for such a collection of independent variables as:

$$\mathbf{P(Y = 1|X) = \pi(X)} \tag{5}$$

(read as "the probability of **Y** equals to 1 given vector **X** is equal to 'PI' **X**)

Then the logit transformation is:

$$\boldsymbol{G}\mathbf{(X) = ln\{\pi(X)/(1\text{-}\pi(X)\}} \tag{6}$$

Therefore, the multiple logistic regression model is given by the equation:

$$\boldsymbol{G}\mathbf{(X)} = \boldsymbol{\beta}\mathbf{0} + \boldsymbol{\beta}\mathbf{1}\boldsymbol{x}\mathbf{1} + \boldsymbol{\beta}\mathbf{2}\boldsymbol{x}\mathbf{2} + \ldots + \boldsymbol{\beta}\mathbf{n}\boldsymbol{x}\mathbf{n} \tag{7}$$

And

$$\mathbf{\pi(X)} = \boldsymbol{e}^{\mathbf{g(x)}}\mathbf{/1} + \boldsymbol{e}^{\mathbf{g(x)}} \tag{8}$$

Logistic regression analysis is similar to regular multiple regression analysis except that the dependent variable (Y) is binary instead of continuous. It competes with discriminant analysis as a method for analyzing binary response variables. The linear-logistic model is:

$$\mathbf{Prob(Y=y1) = 1/(1+}\boldsymbol{Exp}\mathbf{(\text{-}(}\boldsymbol{\beta}\mathbf{0+}\boldsymbol{\beta}\mathbf{1X1+}\boldsymbol{\beta}\mathbf{pXp+\Lambda)} \tag{9}$$

$$\mathbf{Prob(Y=y2) = 1\text{-}\ Prob(Y=y1)} \tag{10}$$

The Xs are the independent variables. Y is the binary dependent variable, which has two possible values, y1 and y2. The βs are the logistic regression intercept and coefficients.

Two other statistical techniques may be used with binary response data. These are multiple regression and discriminant analysis. In a comparison study, Press and Wilson (1978) chose logistic regression as the best all around method in the two-group cases. When faced with a binary response variable and with multiple regression analysis, one approach might be to ignore the binary nature of the response. This approach was especially popular in the early days of computing when multiple regression software was the only software widely available. Unfortunately, this method suffers from a complete failure of the underlying assumptions—normality, constant variance, and independence. Even if we are willing to overlook the failures in the underlying assumptions, a practical problem exists: what do the predicted values of the dependent variable mean?

Linear discriminant analysis is the optimum method for data from two multivariate-normal populations with equal covariance matrices. However, when any of the assumptions breaks down, this method is no longer adequate either. Fortunately, logistic regression is formulated without relying on the strict existence of these assumptions. Even the once-regarded sole disadvantage of slowness of its computational procedures has been overcome by today's powerful computing technology. If missing values are found in any of the independent variables being used, the row is omitted. If only the dependent variable is missing, the row will not be used in the formation of the coefficient estimates, but a predicted value will be generated for that row—a handy and perfect feature to be utilized for forecast or to build ***The Logistic Indicator***.

3.4 Summary

In the last few pages, the author has reviewed all the applicable statistic models for and relating to this study. The review shows that each statistic model has its unique characteristics and purposes. Multiple regression analysis is designed to analyze the relationship between a single criterion and two or more predictors. Logistic regression is the best model when some of the variables are qualitative rather than quantitative or when the required assumptions for multiple regression analysis (e.g., linearity, independence, etc.) are not met.

In the following sections, the author will utilize these statistical tools, wherever applicable, to complete the main tasks for this study: (1) to examine whether international stock exchange indexes are significantly correlated or not; thus, to provide some arguments for or against the EMH in the scope of global security market, (2) to compare the test results with similar prior studies, and (3) to demonstrate how logistic regression analysis can provide more reliable test results on empirical researches.

4. Empirical Hypothesis, Sample Data, and Variables

Generally, once a company grows so large as to have excess production, capital, or other capabilities that are beyond the ability of its home market to absorb (market saturation), then it looks beyond its national borders for additional distribution channels and often further expands its operations. The activities of the giant global companies are aggregated together with the trade volume of other companies, big and small, into various totals and indices for various periods for each country or state. King, Sentana and Wadhwani (1994) assert that all of these "globalization" activities are reflected on all the security exchange indices of the world, therefore providing a major link between foreign trade and the capital markets.

Like a twin brother of global trade, the round-the-clock global security markets, as discussed in the introductory quote from Bose (1988), are not only real but also growing fast. They provide us endless investing opportunities as well as risks day and night. For the past few decades, investors, hedge funds, and money managers have been investing in foreign economies or industry sectors growing faster than those in the United States (Littauer, 1995). Another reason to own foreign stocks is to manage portfolio risk by global diversification. Foreign markets fluctuate to a different rhythm. It is impossible to find that two markets act the same. In fact, each market advances or falls depending mainly on local economic, political, social, and other relevant conditions that are peculiar to that country (Slatter, 1995; Guo, 1990). Global markets may mean global opportunities for investment, but they also mean global risks. While all the hoopla about electronic markets enabling us to buy shares in Hong Kong at breakfast-time, sell in London at lunch-time and then buy some more again at dinner-time as the Pacific exchanges open may sound exciting, they can be dangerously problematic as well. The Crash of

1987 tells the most vivid and horrible story—the Dow Jones Industrial Average, "the world's best-known stock market index" (Zweig 1986, p. 25), declined "a record 508.32 points, 22.6 percent, which prompted the chairman of the exchange, John Phelan, to call it a meltdown" (Bose 1988, p. 1).

4.1 Empirical Hypothesis

H0: The price movement of a particular stock exchange index is not correlated with any other foreign exchange indexes

HA: The price movement of a particular stock exchange index is correlated with some other foreign exchange indexes

What is the current status of efficiency for the global security market? By testing these hypotheses with a large set of empirical data, this study will show that there are strong links between foreign stock exchanges. It will show that the Hong Kong Stock Exchange (HKSE) is highly vulnerable to market movements of other major foreign stock exchanges with the greatest influence from the United States. It will show that there is a clear pattern of movement between HKSE and these related foreign stock exchange indexes. As a result, the empirical evidence should provide valid support for the argument against the semi-strong form of the EMH in global security markets.

4.2 Empirical Data

The empirical tests use daily returns from January 2, 1988 to September 30, 1998 for four international stock exchanges obtained from Morgan Stanley's electronic database. One of the main purposes of this study is to test the validity of the semi-strong form of EMH by testing the null hypothesis that the price movement of a particular stock exchange index is not associated with one or more foreign stock exchange indexes. In all of the following analyses, Hong Kong's (second-day, or 1-day lag) Hang Seng Index (HK2) is selected to be the dependent variable because Hong Kong occupies such a strategic position in the 24-hour global security market. The test data pertaining to country and period are selected for better reference and comparison of the empirical test results with the existing studies of King et al. (1990) and Guo (1990).

4.3 The Explanatory Variables

For the purpose of establishing a practical baseline or reference for this research, the author selects the same foreign stock exchange indexes that Guo (1990) used in her study[4] as the independent variables for the first empirical test:

- Japan's Nikkei Average 225 Index (JAP)
- United States' S&P 500 Stock Index (S_P)
- United Kingdom's FTSE 100 Index (BRT)

These indexes represent the largest (in terms of market capitalization) and the most active stock exchanges in Asia, Europe, and the Americas in three different time zones.

An alert reader might wonder why factors such as a country's GDP, interest rates and other variables commonly used in most security evaluation models were not included among the independent variables in this model. The reason is that this study assumes that the efficient market theory (and arguments against it), capital asset pricing theory, and other theories mentioned in the Literature Survey section above are also valid on the global scale. Therefore, all of the factors affecting security values, whether they are political or economic, sentimental or fundamental, have already been reflected in the stock prices. A security exchange index is the

[4] Guo's study uses data for the period of November 1, 1984 to November 1, 1988. My tests cover the period of January 2, 1988 to September 30,1998

representative average—weighted or un-weighted, arithmetic or geometric—of all the individual stocks traded in a particular exchange. In fact, if these variables were included among our predictors, they would have been eliminated because they are highly correlated with the existing independent variables. Furthermore, most of these factors such as GDP, interest rates, unemployment rates, etc., do not change or are not published daily. Thus, they are not applicable for our dynamic, daily-changing model.

5. Empirical Results

5.1 The Baseline Test

Following the example of Guo (1990), the author performed a multiple regression analysis on data covering the period from January 2, 1988 to September 30, 1998. As described in the last section, Hong Kong's (second-day, or 1-day lag) Hang Seng Index (HK2) is selected as the dependent variable and Japan's Nikkei Average 225 (JAP), S&P 500 Stock Index (S_P), and United Kingdom FTSE 100 (BRT) as the independent variables. Tables 1 to 4 show all the relevant statistics for the test.

Table 1 Descriptive Statistics for Baseline Test

Variable	Sample				Standard
(Market)	Count	Minimum	Maximum	Mean	Deviation
BRT	2803	-4.863E-02	5.590E-02	4.223E-04	8.421E-03
HK2	2803	-2.174E-01	1.882E-01	5.839E-04	1.679E-02
JAP	2803	-6.795E-02	1.323E-01	-7.272E-05	1.394E-02
S-P	2803	-6.865E-02	5.115E-02	5.412E-04	8.508E-03

Note: The data consist of 2803 daily returns of Hong Kong's (second-day, or 1-day lag) Hang Seng Index (HK2), Japan's Nikkei Average 225 Index (JAP), the United Kingdom's FTSE 100 Index (BRT), and the United States' S&P 500 Index (S_P), from January 2, 1988 to September 30, 1998.

Table 2 Correlation Matrix for Baseline Test

Variable				
(Market)	BRT	HK2	JAP	S_P
BRT	1.000			
HK2	0.209	1.000		
JAP	0.243	-0.019	1.000	
S_P	0.358	0.324	0.088	1.000

Note: Pearson Correlation coefficients at the 5% significance level for variables consisting of 2803 daily returns of Hong Kong's (second-day, or 1-day lag) Hang Seng Index (HK2), Japan's Nikkei Average(225) (JAP), the United Kingdom's FTSE100 (BRT), and the United States' S&P 500 Stock Index (S_P), from January 2, 1988 to September 30, 1998.

Table 1 contains all the relevant descriptive statistics for the sample variables. The sample population is 2,803 daily closing index values for each of the four sample stock exchanges. The correlation matrix is shown on Table 2. The Pearson correlation coefficient between particular pairs of indexes varies from a low of -0.019 (between HK2 and JAP) to a high of 0.358 (between BRT and S_P). However, for the purpose of this test, the only meaningful correlations are the ones between the dependent variable HK2 and each of the three independent variables—0.209 with BRT, -0.019 with JAP, and 0.324 with S_P.

Table 3 Regression Statistics for Baseline Test

Independent Variables	Regression	Standard	
(Market)	Coefficient	Error	T-Value *
Intercept	1.667E-04	2.984E-04	0.558
BRT	2.492E-01	3.889E-02	6.402
JAP	-9.070E-02	2.201E-02	-4.119
S_P	5.643E-01	3.748E-02	15.055

Note: The data consist of 2803 daily returns of Hong Kong's (second-day, or 1-day lag) Hang Seng Index (HK2), Japan's Nikkei Average 225 Index (JAP), the United Kingdom's FTSE 100 Index (BRT), and the United States' S&P 500 Index (S_P), from January 2, 1988 to September 30, 1998. *This is the t-test value for testing the hypothesis that j = 0 versus the alternative (after removing the influence of all other independent variables).

Table 3 shows the regression coefficients, standard errors and t-values[5] for the intercept and all the independent variables. Based on these coefficients, the author derives the corresponding regression equation as follows:

$$HK2 = 1.667E\text{-}04 + 2.492E\text{-}01*BRT - 9.070E\text{-}02*JAP + 5.643E\text{-}01*S\_P \qquad (11)$$

Table 4 Multicollinearity Test for Baseline Test

Variable	R-Squared Vs Other
(Market)	Independent Variables
BRT	0.173
JAP	0.059
S_P	0.128

Note: The data consist of 2803 daily returns of Japan's Nikkei Average 225 Index (JAP), the United Kingdom's FTSE 100 Index (BRT), and the United States' S&P 500 Index (S_P), from January 2, 1988 to September 30, 1998.

The multicollinearity tests (R^2 versus other independent variables) of the independent variables (shown in Table 4) is acceptably low for each of the three independent variables. From this regression equation and the data in Tables 1 to 4, we can see that the sample size is large enough to support a reliable empirical test. These statistics imply that the return (%) of the Hang Seng Index of Hong Kong (HK2) is correlated negatively with the prior-day return of Nikkei (225) Index of Japan (JAP), but it is correlated positively with that of the FTSE (100) Index of the United Kingdom (BRT) and of the S&P (500) Index of the United States (S_P). Mathematically speaking, every unit of change in HK2 is correlated with 0.2492 of BRT, -0.0907 of JAP, 0.5643 of S_P, and with a constant of 0.00017. This result is consistent with Guo's (1990) finding that the Hong Kong Stock Exchange (HKSE) is highly vulnerable to major foreign market innovation transmission with the largest influence from the United States.

5.2 The Logistic Tests

We have now found that HKSE is highly vulnerable to major foreign market innovation transmission with the largest influence from the United States. However, all the empirical tests so far violate most of the assumptions behind multiple regression analysis. Also, they still do not tell us when and how to make day-to-day investment decisions even if we turned a blind eye on these requirements. We need a practical and statistically reliable model as a guide for this important task and for verifying the test results of the multiple regression analyses. Fortunately, we can call upon Logistic Regression for the rescue. With this practical purpose in mind and

[5] This is the *t*-test value for testing the hypothesis that *j* = 0 versus the alternative (after removing the influence of all other independent variables.

based on the findings of the prior studies, the author developed ***the Logistic Indicator***—a practical barometer and security investment tool for daily stock index trading.

Using daily percentage changes of the indexes and modifying the data for the dependent variable into a binary format—0 for daily return less than 0.5 percent (triggering threshold), and 1 for daily return of 0.5 percent or more—the author runs a logistic regression analysis with BRT, JAP, and S_P (the same group of independent variables used for the baseline test discussed above). Table 5 provides all the relevant test statistics for the full sample period from January 2, 1988 to September 30, 1998.

Table 5 Logistic Regression Statistics

Independent Variables	Sample	Regression		Prob.		-- Errors (%) --		% Correctly
(Market)	Count	Coefficient	Intercept	Level *	R-Squared			Classified
--HK2 Up 0.5% or more--	2803		-8.313E-01	0.000	0.058	2.64%	28.08%	69.28%
BRT	2803	2.741E+01		0.000				
JAP	2803	-8.685E+00		0.005				
S_P	2803	5.613E+01		0.000				
--HK2 Down 0.5% or more--	2803		-9.939E-01	0.000	0.054	1.39%	25.19%	73.42%
BRT	2803	-2.971E+01		0.000				
JAP	2803	2.958E+00		0.364				
S_P	2803	-5.105E+01		0.000				

Note: The data consist of 2830 daily returns of Hong Kong's (second-day, or 1-day lag) Hang Seng Index (HK2), Japan's Nikkei Average 225 Index (JAP), the United Kingdom's FTSE 100 Index (BRT), and the United States' S&P 500 Index (S_P), from January 2, 1988 to September 30, 1998. *This is the significance level of the test. If it is less than the predefined alpha level (0.05 in these tests), the variable is statistically significant.

Examining the information in Table 5, on any up days (HK2 gains 0.5% or more), we can see that the sample count is still 2,803 observations, and the probability level[6] for the intercept and each of the independent variables ranges from 0 to 0.005 (< 0.05, all are statistically significant). The model has an $\mathbf{R}^2$ of 0.058, an α error of 2.64%, and a β error of 28.08%. The model has a pretty high rate of successful classification of 69.28%. The model's mathematical presentation is as follows:

$$P(HK2 \geq 0.5\%) = 1/(1+Exp[-\{-0.8313\}+\{27.41*BRT\}-\{8.685*JAP\}+\{56.13*S\_P\}]) \quad (12)$$

At this point, the model only predicts whether or not the Hang Seng Index will close 0.5% or more above the level of the day before. We would also like to be able to make money when the index moves down by shorting the index. Unfortunately, the available statistics software does not have the ability to run a polychotomous logistic regression. However, in the absence of software that is capable of performing a polychotomous analysis, we can use the results of a combination of two individual logistic regressions, realizing, of course, that the resulting estimates are approximations to maximum likelihood estimates (Hosmer, 1989).

Thus, the author further modifies the data for the dependent variable into a binary format, 0 for daily decreases of less than 0.5 percent, and 1 for daily decreases of 0.5 percent or more. He runs a similar procedure for the "short" prediction as described above and lists the statistics in the second portion of Table 5. In this test, the sample count stays at 2,803 observations, and the probability level for the intercept and each of the

[6] This is the significance level of the test. If it is less than the predefined alpha level (0.05 in all of these logistic regression tests), the variable is statistically significant.

independent variables remains the same except for JAP which changes from 0.005 to 0.364 (> 0.05, not statistically significant). The model has an $\mathbf{R}^2$ of 0.054, an α error of 1.39%, and a β error of 25.19%. This model has a higher (73.42%) rate of successful classification than the "up" model. The "down" model's mathematical presentation is as follows:

$$P(HK2 \le -0.5\%) = 1/(1+Exp[-\{-0.994\}-\{29.71*BRT\}+\{2.958*JAP\}-\{51.05*S\_P\}]) \quad (13)$$

Thus, similar to King and Wadhwani (1990), Guo (1990), and King, Sentana and Wadhwani (1994), all the results of the empirical tests above reject the null hypotheses of this study that the price movement of a particular stock exchange index is not correlated with any other foreign exchange index(es) and accept the alternative hypotheses that the price movement of a particular stock exchange index is correlated with one or more other foreign exchange index(es). This empirical study's results imply that there is a clear pattern and an arbitrage opportunity for a knowledgeable investor to make meaningful profits. Thus, the semi-strong form of EMH is not supported in the global security market.

This study also demonstrates how to utilize multiple regression analysis to make comparisons and to provide references with prior studies. It has also proved that by running logistic regression analysis on the same set of data, researchers can overcome the innate deficiency of multiple regression analysis—violations of assumptions of the quality of test data. As a result, this empirical study provides a set of more reliable test results.

6. Conclusion

Most of the prior accounting or finance related empirical studies were based solely on results of some forms of multivariate regressions analysis. The reliability of their findings is subject to question because, by definition, all forms of multiple regressions rely critically on some assumptions of the quality of the test data. The obvious problem is that most of the financial data often violate some, and in many cases, all of these assumptions. By running both the multivariate regressions and logistic regressions on a set of empirical data, this study shows that while the multiple regression results provide the necessary statistics, as well as reference or comparison with prior studies, the logistic regression results enhance the research by providing the much-needed confirmation of reliability of the empirical findings.

This study also shows that, today, the capital market has undeniably become part of a global game. Most foreign stock exchanges are significantly linked or correlated. Taking advantage of the linked and round-the-clock global security market, a stock index day-trading model is developed based on the contrarian's view of the efficient market hypothesis and other economic and security analysis theories. ***The Logistic Indicato***r is both flexible and dynamic and perfectly suitable for today's fast-paced, ever-changing global financing environment. With easy access to information and modern computing technology such as XBRL (Fang, 2009), the application of the model is not only simple and fast, but it also can be easily modified for different user preferences. The model can also be easily modified to utilize findings of other studies (e.g., adding more independent variables such as private information).

However, the model is newly developed; therefore, it lacks the much needed empirical data to support its reliability and applicability. Users are also cautioned to reevaluate the model whenever there is a change in the market conditions affecting the relevant exchanges used as variables in the analysis.

References:

Amick D. J. and Walberg H. J. (1975). *Introductory Multivariate Analysis*, Berkeley, California: McCutchan Publishing Corporation.

Avnet T., Pham M. and Stephen A. (Dec. 2012). "Consumers' trust in feelings as information", *Journal of Consumer Research*, Vol. 39, No. 4, pp. 720-735.

Bachelier L. (1900), Theorie de la Speculation (Gauthrei-Villars, Paris), translated into English by Cootner P., *The Random Character of Stock Market Prices*, MIT Press, Cambridge, MA., 1964.

Blumenschein K., Blomquist G., Johannesson M., Horn N. and Freeman P. (Jan., 2008). "Eliciting willingness to pay without bias: Evidence from a field experiment", *The Economic Journal*, Vol. 118, No. 525, pp. 114-137.

Borgos M. R. (2010). "Efficient market hypothesis in European stock markets", *The European Journal of Finance*, Vol. 16, No. 7, pp. 711-726

Bose M. (1988). *The Crash*, London: Bloomsbury Publishing Ltd.

Breslow N. E. and Day N. E. (1980). *Statistical Methods in Cancer Research*, Lyon, France: International Agency on Cancer.

Brown S. J. (1990). *Quantitative Methods for Financial Analysis*, Homewood, Illinois: Dow Jones-Irwin.

Cox D. R. (1970). *The Analysis of Binary Data*, London: Methuen.

Dillon W. R. and Goldstein M. (1984). *Multivariate Analysis—Methods and Applications*, New York: John Wiley & Sons, p. 587.

Ellingsen T., Johannesson M., Lilja J. and Zetterqvist H. (Jan., 2009). "Trust and truth", *The Economic Journal*, Vol. 119, No. 534, pp. 252-276

Fama E. F. (1965). "The behavior of stock-market prices", *The Journal of Business*, No. 38, pp. 34-105.

Fama E. F. (1970). "Efficient capital market: A review of theory and empirical work", *Journal of Finance*, No. 25, pp. 383-417.

Fama E. F., Fisher L., Jensen M. and Roll R. (1969). "The adjustment of stock prices to new information", *International Economic Review*, No. 10, pp. 1-21.

Fang J. (2005). "An empirical investigation of the efficient stock market hypothesis", Doctoral dissertation, Pace University, New York.

Fang J. (May 2009). "How CPAs can master XBRL", *The CPA Journal*.

Guidi F. (2010). "Day-of-the-week effect and market efficiency in the Italian stock market: An empirical analysis", *IUP Journal of Applied Finance*, Vol. 16, No. 2, pp. 5-32.

Guo E. (1990). "An empirical examination of price behavior on the Hong Kong stock market", Doctoral dissertation, Virginia Polytechnic Institute and State University.

Hair J. F., Anderson R. E., Tatham R. L. and Black W. C. (1995). *Multivariate Data Analysis with Readings*, New York: Macmillan Publishing Company.

Hosmer D. W. and Lemeshow S. (1989). *Applied Logistic Regression*, New York: John Wiley & Sons.

King M. A. and Wadhwani S. (1990). "Transmission of volatility between stock markets", *The Review of Financial Studies*, No. 3, pp. 5-33.

King M. A., Sentana E. and Wadhwani S. (Jul., 1994). "Volatility and links between national stock markets", *Econometrica*, No. 62, pp. 901-933.

Kleinbaum D. G., Kupper L. L. and Morgenstern H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*, New York: Van Nostrand Reinhold.

Kolasinski A. and Kothari S. (2008). "Investment banking and analyst objectivity: Evidence from analysts affiliated with mergers and acquisitions advisors", *The Journal of Financial and Quantitative Analysis*, Vol. 43, No. 4, pp. 817-842.

Kotha S., Rajgopal S. and Venkatachalam M. (2004). "The role of online buying experience as a competitive advantage: Evidence from third-party ratings for e-commerce firms", *The Journal of Business*, Vol. 77, No. S2, pp. S109-S133.

Lachin J. M. (2008). *Logistic Regression Models in Biostatistical Methods: The Assessment of Relative Risks*, Hoboken, USA: John Wiley & Sons, Inc.

Lampe M. (2011). "Explaining nineteenth-century bilateralism: Economic and political determinants of the Cobden-Chevalier network", *The Economic History Review*, Vol. 64, No. 2, pp. 644-668.

Littauer S. (1995). *How to Buy Stocks the Smart Way*, Chicago: Dearborn Financial Publishing.

Lorie J. H., Dodd P. and Kimpton M. H. (1985). *The Stock Market: Theories and Evidence*, Homewood, IL, USA: Richard D. Irwin.

Mansfield E. (1994). *Multivariate Analysis—Methods and Applications*, New York: John Wiley & Sons, p. 587.

Mayew W. and Venkatachalam M. (2012). "The power of voice: Managerial affective states and future firm performance", *The Journal of Finance*, Vol. 67, No. 1, pp. 1-43.

Olivier C., Blake R., Steed L. and Salgado C. (1978). "Risk of vancomycin-resistant enterococcus (VRE) bloodstream infection among patients colonized with VRE”, *Infection Control and Hospital Epidemiology*, Vol. 9, No. 34, pp. 1988-2013.

Osborne M. E. M. (1959). “Brownian motions in the stock market”, *Operations Research*, No. 7, pp. 145-173.

Pearson K. and the Right Honorable Lord Rayleigh (1905). “The problem of the random walk”, *Nature*, No. 72, pp. 294, 318, and 342.

Press S. J. and Wilson S. (1978). “Choosing between logistic regression and discriminant analysis”, *Journal of the American Statistical Association*, No. 73, pp. 699-705.

Rusticelli E., Ashley R., Dagum E. and Patterson D. (2008). “A new bispectral test for nonlinear serial dependence”, *Econometric Reviews*, Vol. 28, No. 1-3, pp. 279-293.

Schlesselman J. J. (1982). *Case-Control Studies*, New York: Oxford University Press.

Slatter J. (1995). *Straight Talk about Stock Investing*, New York: McGraw-Hill.

Zweig M. E. (1986). *Winning on Wall Street*, New York: Warner Books.