LOGISTIC REGRESSION TO DETERMINE SIGNIFICANT FACTORS

ASSOCIATED WITH SHARE PRICE CHANGE

By

HONEST MUCHABAIWA

submitted in accordance with the requirements for the degree of

MASTER OF SCIENCE

in the subject

STATISTICS

at the

UNIVERSITY OF SOUNTH AFRICA

SUPERVISOR: MS. S MUCHENGETWA

February 2013

ACKNOWLEDGEMENTS

First and foremost, I wish to thank God, the Almighty for giving me strength, discipline, determination, and his grace for allowing me to complete this study.

To my supervisor – Ms. S Muchengetwa, Thank you for your guidance, constructive criticism, understanding and patience. Without your support I would not have managed to do this.

The study would not have been complete without the assistance of Mrs A Mtezo who helped me to access data from the McGregor BFA website. Your help is greatly appreciated.

Last but not least, my gratitude goes to my wife Abigail Witcho, my daughter Nicole and my son Nathan for being patient with me during the course of my studies.

DECLARATION BY STUDENT

I declare that the submitted work has been completed by me the undersigned and that I have not used any other than permitted reference sources or materials nor engaged in any plagiarism.

All references and other sources used by me have been appropriately acknowledged in the work.

I further declare that the work has not been submitted for the purpose of academic examination, either in its original or similar form, anywhere else.

Declared on the (date): 28 February 2013

Signed: ____________________________

Name: Honest Muchabaiwa

Student Number: 46265147

ABSTRACT

This thesis investigates the factors that are associated with annual changes in the share price of Johannesburg Stock Exchange (JSE) listed companies. In this study, an increase in value of a share is when the share price of a company goes up by the end of the financial year as compared to the previous year. Secondary data that was sourced from McGregor BFA website was used. The data was from 2004 up to 2011.

Deciding which share to buy is the biggest challenge faced by both investment companies and individuals when investing on the stock exchange. This thesis uses binary logistic regression to identify the variables that are associated with share price increase.

The dependent variable was annual change in share price (ACSP) and the independent variables were assets per capital employed ratio, debt per assets ratio, debt per equity ratio, dividend yield, earnings per share, earnings yield, operating profit margin, price earnings ratio, return on assets, return on equity and return on capital employed.

Different variable selection methods were used and it was established that the backward elimination method produced the best model. It was established that the probability of success of a share is higher if the shareholders are anticipating a higher *return on capital employed,* and high *earnings/ share*. It was however, noted that the share price is negatively impacted by dividend yield and earnings yield.

Since the odds of an increase in share price is higher if there is a higher return on capital employed and high earning per share, investors and investment companies are encouraged to choose companies with high earnings per share and the best returns on capital employed.

The final model had a classification rate of 68.3% and the validation sample produced a classification rate of 65.2%.

Keywords: Logistic Regression, Binary Logistic Regression, Share Price, Stock Exchange, Akaike's Information Criterion, Wald Test, Score Test, Enter method, Stepwise Logistic Regression.

CONTENTS

List of Tables

CHAPTER 1: INTRODUCTION

Businesses have two choices when they want to raise investment capital to expand their operations. The choices are either to borrow from a bank or to issue shares (Johannesburg Stock Exchange, 2011). A share or a stock or equity is a portion of a company and its owner has a claim on that business's earnings and assets (Johannesburg Stock Exchange, 2011). A person who owns shares within a company is called a shareholder. Shareholders buy shares hoping for an increase in the share prices and thus increasing their capital in what is referred to as capital gains on their investment and they will be also hoping to receive dividends which can act as a source of income.

The shares of a company can be transferred from one shareholder to another through sale or other mechanisms, unless prohibited. Such transfers are governed by laws and regulations especially if the issuer is a public entity. The need to develop a platform for shareholders to trade their shares has resulted in the establishment of stock exchanges. A stock exchange is defined as an organisation that provides a marketplace for easy buying and selling of shares, derivatives and financial products (http://en.wikipedia.org/wiki/Stock).

Stock prices change every day as a result of market forces. This means that share prices change because of supply and demand. If more people want to buy a stock (demand) than sell it (supply), then the price moves up. Conversely, if more people wanted to sell a stock than buy it, there would be greater supply than demand, and the price would fall.

The stock exchange for South African listed companies is called the Johannesburg Stock Exchange. The stock exchange reduces the risk of trading in shares by providing a fair and transparent pricing and also policies for registered / listed companies. The environment in which the stock exchange operates has strict regulations and all listed companies have to comply with certain listing requirements.

When shareholders invest their money by buying shares on the Johannesburg stock exchange (JSE) their motive is to make money and this can only happen if the share price appreciates in value after the purchases. This means when they decide to sell the share, they will make a profit. On the other hand if a share losses values then the shareholders will make a loss when they dispose of the shares.

The current Johannesburg Stock Exchange (JSE) Limited was established as The Johannesburg Exchange & Chambers Company on the 8th of November 1887 by Benjamin Minors Woollan, a London businessman. It was established to facilitate the eruption of need to trade that was triggered by discovery of gold in the Witwatersrand in 1886. By 31 December 2012, it was the largest stock exchange in Africa and the 17th in the world with a market capitalisation of US$903billion, with US$287billion having exchanged hands on the market (http://en.wikipedia.org/wiki/List_of_stock_exchanges). There were 472 listed companies by end of December 2012.

Johannesburg stock exchange is a very competitive in comparison with other markets in the world. Unlike most of the stock exchanges in Africa which are not yet transacting electronically, the JSE is fully electronic and it uses a system called the Johannesburg Equities Trading (JET) System. With the JET system sellers of a stock will indicate the amount of shares that they will be selling and the price. Prospective buyers will also indicate the stock that they are willing to buy, the price and the quantity. As soon as there is a match on the selling price that the seller is willing to sell for and the price at which the buyer is willing to pay then a trade is automatically executed. The trades are conducted in real time.

The Table 1.1 below shows the position of the JSE on the Top 20 world stock exchanges as at 31 December 2012. A total of 287 billion United States dollars' worth of trades was conducted in 2012 alone. This translates to more than one billion United states traded daily since the stock exchange opens from Monday to Friday excluding public holidays and the year 2012 had 250 such days in South Africa.

Table 1.1: Top 20 Stock Exchanges in the World by Market Capitalisation

Rank	Stock Exchange	Economy	Headquarters	Market Capitalisation (US$bn)	2012 Annual Trade Value (US$bn)
1	NYSE Euronext	United States/Europe	New York City	14,085	12,693
2	NASDAQ OMX Group	United States/Europe	New York City	4,582	8,914
3	Tokyo Stock Exchange	Japan	Tokyo	3,478	2,866
4	London Stock Exchange	United Kingdom	London	3,396	1,890
5	Hong Kong Stock Exchange	Hong Kong	Hong Kong	2,831	913
6	Shanghai Stock Exchange	China	Shanghai	2,547	2,176
7	TMX Group	Canada	Toronto	2,058	1,121
8	Deutsche Börse	Germany	Frankfurt	1,486	1,101
9	Australian Securities Exchange	Australia	Sydney	1,386	800
10	Bombay Stock Exchange	India	Mumbai	1,263	93
11	National Stock Exchange of India	India	Mumbai	1,234	442
12	SIX Swiss Exchange	Switzerland	Zurich	1,233	502
13	BM&F Bovespa	Brazil	São Paulo	1,227	751
14	Korea Exchange	South Korea	Seoul	1,179	1,297
15	Shenzhen Stock Exchange	China	Shenzhen	1,150	2,007
16	BME Spanish Exchanges	Spain	Madrid	995	731
17	JSE Limited	South Africa	Johannesburg	903	287
18	Moscow Exchange	Russia	Moscow	825	300
19	Singapore Exchange	Singapore	Singapore	765	215
20	Taiwan Stock Exchange	Taiwan	Taipei	735	572

Source: http://en.wikipedia.org/wiki/List_of_stock_exchanges

A share/stock price is the reigning price at which a specific share can be sold or bought on the stock exchange. There are a number of factors that affect the price of a share. According to the Johannesburg Stock Exchange (2011), besides supply and demand, the price of a share is affected by the following;

- The share price of a profitable company will be more valuable because more investors will be viewing them as a worthwhile investment.
- The share price is also influenced by economic and political events.

Numerous scientific attempts have been made to try and accurately predict stock price movement but no single method have been discovered to date (Schumaker and Chen, 2006). According to Senol (2008) there is no method that has been found to precisely predict the stock price behaviour. He also wrote that high rate of uncertainty and volatility that is associated with share price renders trading in stocks a very higher risk as compared to any other investment area. This makes stock price behaviour difficult to predict.

Senol (2008) indicated that conventional methods, have been applied to stock price prediction but they have either partially succeeded or failed completely to deal with the non-linear and multifaceted behaviour of stock prices. Lawrence (1997) used neural networks to forecast stock market prices whilst Sharma (2011) used regression analysis to predict the stock prices. On the other hand Al-Dini, Dehavi, Zarezadeh, Armesh, Manafi, and Zraezadehand (2011) used Fuzzy Regression to determine the relationship between financial variables and stock price.

Khan, Aamir, Qayyum, Nasir, and Khan (2011) used multiple regression to assess the variables that impact on Stock price. Azam and Kumar (2011), also applied multiple regression analysis to predict the relationship between stock prices and influencing variables.

From one reporting period of a company to the next (a financial year), the share price may go up, remain constant or go down. Investors are interested in an increase in share price as that means a growth in their wealth. A constant share price is as good as a decline in the share price for shareholders as they would not have realized any gain on their investment. Thus, in this research the success of a share price is when the share price increases in value whilst a failure is when the share price goes down or remained constant.

The purpose of this study is to devise a method of predicting the annual change in share price (ACSP) of JSE listed companies hence enabling prospective investors to invest their money in shares that are more likely to appreciate in value. ACSP is given by;

$$ACSP = \begin{cases} 1 = success \quad if\ annual\ change\ in\ share\ price\ is\ positive \\ 0 = failure \quad if\ annual\ change\ in\ share\ price\ is\ negative \\ \qquad\qquad\qquad or\ constant \end{cases}$$

Thus, the objectives are;

- To fit a logistic model to the annual change in share price
- To determine the adequacy of the fitted model, and
- To compare the results of binary logistic regression using stepwise backward elimination, stepwise forward selection and a method of entering all independent variables at once.

Logistic regression is the most popular regression technique that is used for modeling categorical dependent variables (Kleinbaum, Kupper, Nizam and Muller, 2008).

This thesis utilises logistic regression to find the variables that determine the ACSP at the Johannesburg Stock Exchange (JSE). Logistic regression was chosen because the researcher is interested in the annual change in share price as the dependent variable (either success or failure). The results will help investors to make informed decisions based on the odds of an annual increase in share and the odds of an annual decrease or static share price. In this research the success of a share is when it appreciates value and a failure is when a share loses value or does not change in value.

In Chapter 2 of the thesis, a literature review is presented. The literature review has three sections. First, a brief review of the methods that have been used in the past to predict share prices and the variables that were used, the second section has definitions of the variables that will be used to determine share price. The theory of Logistic regression, its application to share price and, the steps of carrying out stepwise binary logistic regression procedures and the measures that are used to determine significance of variables for inclusion or exclusion in a model are presented in Chapter 3. Research design, variables used and the sample size are discussed in Chapter 4. Analysis and discussion of results will make up Chapter 5 and the summary, conclusion and recommendations will be presented in Chapter 6.

CHAPTER 2: SHARE PRICE AND ASSOCIATED VARIABLES

2.1: Introduction

This chapter presents the literature associated with share price changes and the researches done so far on share price determination. The key factors associated with share price change will be discussed and past results that validate the association between the factors and share prices will be presented. Terminology associated with share price will also be defined.

2.2: Variables Associated with change in Share Price

According to Lawrence (1997) analysts either use technical analysis or fundamental analysis to determine the future value of a stock. Technical analysis uses the assumption that share prices move in trends influenced by the continuously changing attitudes of investors. Technical analysis use movements in share price and trends in the volume of shares traded to predict stock price. This method utilises charts to forecast future stock price movements. It is based on the assumption that future market direction can be determined by examining historical prices as history has a tendency of repeating itself.

Fundamental analysis on the other hand is dependent on in-depth analysis of a company's financial performance and profitability to establish the share price. Lawrence (1997), postulated that by studying a company's competition, the overall economic conditions, its management and other factors, one can establish the expected returns and the actual value of shares. Fundamental analysis is based on the assumption that a firm's current share price and its future price is dependent on its intrinsic value and expected return on investment.

According to Matthew and Odularu (2009), if a company declares a good bonus and dividends for its shareholders, this will also lead to an increase in its share price. Matthew and Odularu (2009) further postulated that investors will be attracted if a good dividend and bonus history is maintained and this will lead to an increase in the value of the market capitalisation of the company. As a result, more funds would be at the company's disposal for growth purposes and this will then lead to an increase in its turnover in an ever-flowing cycle.

Khan, Aamir, Qayyum, Nasir, and Khan (2011) used multiple regression to assess the relationship between *stock price* and *dividend yield*, *profit after tax, earnings per share, retention ratio* and *return on equity.* They regressed the dependent variable (market price of shares) against *retention ratio* and *dividend yield* after with three other control variables namely *earnings per Shares, Profit after Tax* and *Return on Equity* to assess their effect on *Stock Prices*.

Their results revealed that *earnings per share*, *dividends* and *profit after tax* had a significant positive relationship to stock price at the Karachi Stock Exchange. However, r*etention ratio* and *return on equity* were not significant contributors to *stock price*. *Dividends* were the major determinants of the *share price*.

Nishat and Irfan (2003) used cross-sectional regression analysis to explore the relationship between *stock price volatility* and *dividend policy* and *firm size*. Their conclusion was that *dividend yield*, *pay-out ratio* and *firm size* were the determinants of *stock price*.

Midan (1991) used multiple regression to establish the determinants of changes in stock prices of Kuwaiti companies. The results revealed that the Kuwaiti *stock prices* were mainly driven by *earnings per share*, and to a lesser extent by the degree of financial leverage. Madan suggested that further research be carried out since the sample that was used for the research was small.

Al-Dini, Dehavi, Zarezadeh, Armesh, Manafi, and Zraezadehand (2011) used fuzzy regression to determine the relationship between financial variables and *stock price*.

Their findings were that there is a relationship between *dividends per share*, *earning per share*, and *price to earnings* variables and *stock price*. They found a positive relationship between *earning per share* and *stock price*, a negative relationship between *dividends per share* (DPS) and Iran Khodro's *stock price*, and also a negative relationship between *price to earnings ratio* and *stock price*. It is predicted that the more the ratio amount decreases, the more the *stock price* increases.

Azam and Kumar (2011), applied multiple regression analysis to predict the relationship between influencing variables and *stock prices*. Their findings were that stock price was positively related to *dividend yield*, *earnings per share*, *foreign direct investments* and *gross domestic product growth rate*.

According to D'Amato (2010) investors make use of a number of factors to determine the financial health of a listed company. They can use profit and loss, cash flow statements and balance sheets that can be summarised in the form of financial ratios. Financial ratios compare one financial figure with another financial figure and they are known to be associated with share price changes. Financial ratio analysis looks at a firm's financial statements, its management, the health and position in the competitive environment to determine a share price value.

Majority of the variables that were found to be associated with *stock price* changes in past research such as *earnings per share*, *earnings yield*, and *return on assets* are financial ratios. Thus, in this research financial ratios will be used as the independent variables. Some of the important ratios are defined below:

Earnings per share (EPS) ratio measures earnings in relation to every share on issue. The formula is given by

$$\text{EPS} = \frac{\text{Net Income}}{\text{Average Weighted Number of Shares}} \; x\ 100$$

EPS indicates how much each share earned and the higher the EPS, the more likely will the share price go up.

Earnings yield (EY) is Earnings per share expressed as a percentage of the current share price. This is calculated as:

$$\text{EY} = \frac{\text{EPS}}{\text{Share Price}} \text{X } 100$$

The higher the earnings yield, the more likely will the share price go up.

Price to earnings ratio (PE) indicates the number of times the share price covers the earnings per share over a 12 month period. It is calculated as:

$$PE = \frac{Share\ Price}{Earnings\ per\ share}$$

It can be interpreted as how much an investor pays for every rand that the company earns. According to D'Amato (2010), *earnings per share ratio is* widely used by most investors to assess a company's value. The higher the value the more likely the share price will go up because the investors will be seeing value in the company.

Return on assets, affectionately known as ROA, is a measurement of management performance. It indicates how well a corporation utilises its assets to generate revenue. A higher ROA signifies a higher level of management performance. The ROA is calculated using the formula:

$$ROA = \frac{Net\ Income}{Average\ Total\ Assets} \text{ x } 100$$

Return on equity (ROE) is a measurement of management performance which indicates how well a company has used the capital from its shareholders to generate profits. A higher ROE signifies a higher level of management performance. It is calculated using the formula:

$$ROE = \frac{Net\ Income}{Average\ Shareholder's\ Equity} \text{ X } 100$$

Dividend yield (DY) is a calculation of all the dividends paid in a calendar year expressed as a percentage of a company's current share price. It is given by the formula:

$$DY = \frac{Full\ Year\ Divident}{Share\ Price} \text{ x } 100$$

The higher the dividend yield the more attractive the share and increasing demand and hence the share price.

Debt to equity ratio (DE) gives an indication of a corporation's capital structure and shows if a corporation is more reliant on debt or shareholder capital (equity) to finance assets and activities. The formula is given by:

$$DE = \frac{Total\ Debt}{Shareholder's\ Equity}$$

A higher ratio indicates greater risk as greater debt can result in unstable earnings due to extra interest expense as well as increased susceptibility to business downturns (D'Amato, 2010).

Debt to assets ratio provides the relationship between a company's debts and assets. The formula is:

$$DE = \frac{Total\ Debt}{Total\ Assets}$$

A value close to zero is normally satisfactory, because it shows that more assets are paid for without having to borrow money. Creditors have first claim on a firm's assets in the event of forced liquidation and thus the lower the debt to assets ratio, the more attractive the share to the investor.

Return on capital employed (ROCE) is also a measurement of management performance. It indicates how well a company is utilising its capital to generate profits. The formula for calculating ROCE is:

$$ROCE = \frac{Profit\ before\ interest\ and\ Tax}{Capital\ Employed} X\ 100$$

Operating profit margin (OPM) is a ratio of operating profit to sales or turnover. It is calculated by:

$$OPM = \frac{Profit\ before\ interest\ and\ Tax}{Turnover} X\ 100$$

A high operating profit margin is either due to high sales prices or low costs and is normally good news as it suggests good company performance and hence attractive to investor thus associated with increase in share prices.

Assets to capital employed ratio shows the proportion of assets in the capital employed. The ratio is calculated as:

$$Assets\ to\ Capital\ Employed\ ratio = \frac{Assets}{Capital\ Employed}$$

A company's capital employed is divided into assets and working capital. A high asset to capital employed ratio denotes the heavy investment in assets and insufficient working capital.

2.3: Summary

Past research indicated that the share price is mainly affected by financial ratios which measure the performance of the management and the performance of the company at large. The variables that were outstanding in predicting the *share price* in almost all the researches that were carried out prior to this research are *dividends*, and *earnings per share*. There are other financial ratios that came out once or twice in the statistical researches conducted over the years. In this research all the financial ratios will be used as independent variables against a categorical variable annual change in share price (success or failure). In such a case where a variable with binary responses is used as the dependent variable against metric independent variables, multiple linear regression that was used by most researchers will not be appropriate and thus binary logistic regression will be used for the research.

CHAPTER 3: LOGISTIC REGRESSION

3.1: Introduction

This chapter presents the theory of logistic regression, make an account of how logistic regression differs from conventional regression. The history of logistic regression, its' application to share price is also discussed. Model fit statistics such as deviance, the likelihood ratio, Wald test and score test which are used to assess the significance of individual coefficients for inclusion or exclusion in a model in stepwise logistic regression were discussed.

3.2: Logistic Function and Logistic Regression

According to Al-Ghamdi (2001), regression methods are widely used for analysing the relationship between a dependent variable and one or more independent variables. The most popular regression method is linear regression using the method of least squares also referred to as conventional regression analysis (CRA). It is however applicable if the dependent variable is continuous, independent and identically distributed (iid) only. In cases where the dependent variable is categorical, conventional regression analysis is not appropriate.

The most significant reasons why CRA cannot be used when there is a dichotomous dependent variable are:

1. The dependent variable in CRA should be continuous, and

2. The dependent variable in CRA can take negative values.

3. The dependent variable in CRA should be normally distributed

4. The error terms in CRA should be independent and identically distributed

These CRA assumptions are not satisfied in cases where the dependent variable is categorical. In such cases logistic regression analysis (LRA) is applied (Dayton, 1992).

Logistic regression, like least squares regression, is a statistical technique that is used to explore the relationship between a dependent variable and at least one independent variable. The difference is that, linear regression is used when the dependent variable is continuous, while logistic regression techniques are used with categorical dependent variables.

Logistic regression, like any other model building technique in statistics is aimed at finding the best fitting and most economical and yet sensible model to assess the relationship between a response variables and at least one independent variables. It differs from the linear regression in that, it can be applied when the dependent variable is categorical and that it does not require rigorous assumptions to be met (Al-Ghamdi, 2001).

3.3: Binary Logistic Regression

Binary Logistic regression is a prognostic model that is fitted where there is a dichotomous/binary dependent variable like in this instance where the researcher is interested in whether there was an increase in stock price or not. Usually, the categories are coded as “0" and "1" as it results is a straightforward interpretation. Normally the category of interest also affectionately referred to the case is typically coded as "1" and the other group is also known as a "non case" as "0" (http://en.wikipedia.org/wiki/Logistic_regression). In this research an increase in the share price, “case”, will be denoted by a 1 and if the price remained the same or declined “non case” will be denoted by 0 (Prempeh, 2009).

3.4: Logistic Regression Model

According to Harrell (2001), the formula for a logistic regression model is given by;

$$\pi\ (\boldsymbol{x_i}) = \mathrm{P}(y_i = 1 \vdots \boldsymbol{x_i})$$

$$= [1 + \exp(-\boldsymbol{X^T\beta})]^{-1}$$

where, $y_i = \begin{cases} 1\ if\ a\ share\ price\ increases \\ 0\ if\ a\ share\ price\ decreases \\ \quad or\ remains\ constant \end{cases} \quad i = 1, 2, \dots, n$

$\boldsymbol{X^T\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1}\, x_{p-1}$

$$\boldsymbol{\beta_{pX1}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad , \quad \boldsymbol{X_{pX1}} = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_{p-1} \end{bmatrix} \ , \ \boldsymbol{X_{i_{pX1}}} = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix}$$

$x_1, x_2, \dots, x_k$ are the independent variables.

β_0 is the coefficient of the constant term

$\beta_1, \beta_2, \ldots.., \beta_{p-1}$ are the coefficients of the p independent variables

π $(\boldsymbol{x_i})$ is the probability of an event that depends on p-independent variables.

Since π $(\boldsymbol{x_i}) = [1 + \exp(-\boldsymbol{X^T\beta})]^{-1}$

$$= \frac{1}{1 + \exp(-\boldsymbol{X^T\beta})}$$

$$\Rightarrow 1 - \pi\ (\boldsymbol{x_i}) = 1 - \frac{1}{1+\exp(-\boldsymbol{X\beta})}$$

$$= \frac{[1 + \exp(-\boldsymbol{X^T\beta})] - 1}{1 + \exp(-\boldsymbol{X^T\beta})}$$

$$= \frac{\exp(-\boldsymbol{X^T\beta})}{1 + \exp(-\boldsymbol{X^T\beta})}$$

$$\Rightarrow \frac{\pi(\boldsymbol{x_i})}{1-\pi\ (\boldsymbol{x_i})} = [\exp(-\boldsymbol{X^T\beta})]^{-1}$$

Thus, $\text{In}\left(\frac{\pi(x_i)}{1-\pi\ (x_i)}\right) = logit[\pi\ (\boldsymbol{x_i})]$

$$= \boldsymbol{X^T\beta}$$

According to Kleinbaum, Kupper, Nizam and Muller (2008), logistic regression quantifies the relationship between the dichotomous dependent variable and the predictors using odds ratios. Odds ratio is the probability that an event will occur divided by the probability that the event will not happen. In this study the odds ratio is the probability that a share price will appreciate in value annually divided by the probability that the share price will not appreciate in value.

Odds are calculated using the formula;

$$\text{Odds} = \frac{\text{P(Case)}}{\text{P(Non case)}}$$

$$= \frac{\pi(x)}{1 - \pi(x)}$$

$$= [\exp(-\boldsymbol{X^T\beta})]^{-1}$$

where, $\pi(x)$ is the probability of success (case) and $1 - \pi(x)$ is the probability of failure (non case).

The odds ratio (OR) which is meant to indicate whether the odds of a success (case) are equally likely to the odds of failure is given by

$$\text{Odds Ratio} = \frac{\text{Odds of Case}}{\text{Odds of Non case}}$$

An odds ratio of one is an indication that the odds of a success (case) outcome are equally likely for to the odds of a failure (non-case) (http://en.wikipedia.org/wiki/Logistic_regression). The odds ratio has a minimum value of zero but have no upper limit. A value less than one indicate that the case is not likely to prevail under those circumstances and a value greater than one indicates a high likelihood for belonging to the group. The further the odds ratio is from one, the stronger the relationship.

Rearranging, the resultant will be

$$\frac{\pi(x)}{1 - \pi(x)} = [\exp(-\boldsymbol{X^T\beta})]^{-1}$$

$$= odds$$

Taking the natural logarithm of both sides:

$$\text{In}\left[\frac{\pi(x)}{1-\pi(x)}\right] = -[-\boldsymbol{X^T\beta}]$$

$$= \boldsymbol{X^T\beta}$$

$$\text{In(Odds)} = \text{Logit (y)}$$

$$= \text{In}\left[\frac{\pi(\text{x})}{1-\pi(\text{x})}\right]$$

$$= \boldsymbol{X^T\beta}$$

Where, Logit (y) is the natural logarithm of the odds of outcome,

The coefficients $\boldsymbol{\beta} = [\beta_0 , \beta_1, \beta_2 , \dots , \beta_k]^T$ are estimated using the maximum likelihood (ML) method:

$$\text{G(x)} = \text{In}\left[\frac{\pi(\text{x})}{1-\pi(\text{x})}\right] = \boldsymbol{X^T\beta}$$

The transformation $G(\text{x})$ is referred to as the logit transformation:

According to Al-Ghamdi (2001), the logit transformation, G(x) is important because it has a lot of the desirable properties of a linear regression model. The logit transformation, is linear in its parameters, may range from $-\infty$ to $+\infty$ depending on the range of $\boldsymbol{X}$. The inverse of the logit transformation can only take values 0 or 1.

3.5: Assumptions of Logistic Regression

Logistic regression is not dependent on stringent assumptions to be met as compared to linear regression. The fact that logistic regression analysis does not require a lot of assumptions renders it more preferable in some instances to other methods. The following details how it differs from other techniques:

- The error terms are with a mean of zero and a variance of $\pi(\text{x})[1-\pi(\text{x})]$. (Hosmer and Lemeshow, 2000).

- The conditional mean of the regression equation is greater than or equal to 0 and less than or equal to 1.
- The same principles used when conducting linear regression also apply but the difference is only that the equation will be modelling the log odds and not the actual relationship among variables.

3.5.1: Model Estimation

According to Kutner, Nachtsheim, Neter, and Li (2005), since the dependent variable is dependent and can take values 1 and 0 with probabilities $\pi\ (\boldsymbol{x_i})$ and $1 - \pi\ (\boldsymbol{x_i})$ respectively, Y follows a Bernoulli distribution with $E(Y) = \pi\ (\boldsymbol{x_i})$.

Thus, $Y_i = \pi\ (\boldsymbol{x_i}) + \varepsilon_i$.

$$\mathrm{E}(Y_i) = \pi\ (\boldsymbol{x_i})$$

$$= [1 + \exp(-\boldsymbol{X^T\beta})]^{-1}$$

$$= \frac{1}{1 + \exp(-\boldsymbol{X^T\beta})}$$

$$\mathrm{P}(Y_i = 1) = \pi\ (\boldsymbol{x_i})$$

$$\mathrm{P}(Y_i = 0) = 1 - \pi\ (\boldsymbol{x_i})$$

The probability density function can be presented as

$f_1(Y_i) = \pi\ (\boldsymbol{x_i})^{Y_i}[1 - \pi\ (\boldsymbol{x_i})]^{1-Y_i}$ for $Y_i = 0, 1, 2, \dots, n$

The Y_i's are assumed to be independent and thus, the joint probability function is given by

$$g(Y_1, \dots, Y_n) = l(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_1(Y_i)$$

$$= \prod_{i=1}^{n} \pi\ (\boldsymbol{x_i})^{Y_i}[1 - \pi\ (\boldsymbol{x_i})]^{1-Y_i}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters.

Working with logarithms is much easier in this case (Kutner, Nachtsheim, Neter, and Li, 2005). Taking natural logarithms of both sides we have:

$$In[g(Y_1, \dots, Y_n)] = In\big(l(\boldsymbol{\beta})\big) = \textstyle\sum_{i=1}^{n}[Y_i In\ \pi\ (\boldsymbol{x_i}) + (1 - Y_i) In\ (1 - \pi\ (\boldsymbol{x_i}))]$$

$$L(\beta) = In(l(\beta)) = \sum_{i=1}^{n}\{Y_i In[\pi(x_i)] + (1 - Y_i) In[1 - \pi(x_i)]\}$$

$$= \sum_{i=1}^{n}\{Y_i In\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] + \sum_{i=1}^{n} In[1 - \pi(x_i)]$$

since $\text{In}\left[\frac{\pi(\text{x})}{1-\pi(\text{x})}\right] = \boldsymbol{X^T\beta}$

$$\Longrightarrow L(\beta) = In(l(\beta)) = \sum_{i=1}^{n} Y_i(\boldsymbol{X^T\beta}) - \sum_{i=1}^{n} In[1 + exp(\boldsymbol{X^T\beta})]$$

The maximum likelihood of $\boldsymbol{\beta}$ is obtained by maximising the $L(\beta) = In(l(\beta)) = \sum_{i=1}^{n} Y_i(\boldsymbol{X^T\beta}) - \sum_{i=1}^{n} In[1 + exp(\boldsymbol{X^T\beta})]$ with respect to $\boldsymbol{\beta}$. The process yields equations that are nonlinear in $\boldsymbol{\beta}$ and hence the estimates are obtained by numerical methods (Kutner, Nachtsheim, Neter, and Li, 2005).

3.5.2: Model Diagnostics

After estimating the Logistic regression model parameters using the maximum likelihood estimator, there is a need to assess the significance of the variables with regards to predicting the response variable. There are a number of statistics that can be used to carry out the assessment and these include deviance, likelihood ratio, Wald Test and Score Test (Harrell, 2001). These tests are discussed in the sections below.

Deviance

The observed values of the dependent variable must be compared with the estimated values obtained from models with and without the variable in question. This comparison is based on the log-likelihood function;

$\sum_{i=1}^{n}\{y_i In[\pi(x_i)] + (1-y_i)In[1-\pi(x_i)]\}$.

A comparison has to be made between a saturated model and the current model where a saturation model is one that contains as many parameters as the number of data points and the current model is the one that contains only the variables being assessed. The comparison of the current to saturated model is based on the likelihood ratio:

$$D = -2In[\frac{Likelihood\ of\ the\ current\ model}{Likelihood\ of\ the\ saturated\ model}]$$

Using the two equations above, the test statistic can be obtained to be

$$D = -2\sum_{i=1}^{n}[y_i In(\frac{\pi(x_i)}{y_i}) + (1-y_i)In(\frac{1-\pi(x_i)}{1-y_i})]$$

According to Hosmer and Lemeshow (2000), the statistic D, is called the deviance, and it plays an essential role in the assessment of goodness of fit of the model. The deviance plays the same role in logistic regression as the residual sum of squares plays in linear regression (Hosmer and Lemeshow, 2000).

Deviance (D) follows a chi-square distribution with q- degrees of freedom, where q is the number of covariates in the equation. It tests the hypothesis:

H_0 : All the coefficients of the parameters in the saturated model and not in the current model are equal to zero

H_1 : Not all the coefficients of the parameters in the saturated model and not in the current model are equal to zero

A p-value greater than 0.05 (the significance level) is an indication that at least one coefficient is non-zero (Abdelrahman, 2010). According to Agresti (2007), large deviance values and p-values less than 0.05 are an indication of lack of fit of the current model.

$\mathbf{R^2}$ for Logistic Regression

Unlike when using liner regression where the r-square measures the amount of variation in the dependent variable that is explained by the independent variables, in logistic regression there is controversy regarding the relevance of r-square measures in assessing the predictive power of a model (Harrell, 2001). The R^2 for Logistic regression is estimated by the Cox and Snell R^2 computed as ;

$$Cox\ \&\ Snell\ Pseudo\ R^2 = [\frac{-LL_0 - LL_k}{-LL_0}]^{n/2}$$

where, LL_0 is the loglikelihood of the null model and LL_k is the loglikelihood of the current model. This value cannot reach 1 and Nagelkerke improved it to reach 1. The improved R^2 is given by :

$$\text{Nagelkerke Pseudo } R^2 = \frac{[\frac{-LL_0 - LL_k}{-LL_0}]^{n/2}}{1 - [-2LL_0]^{n/2}}$$

where, LL_0 is the loglikelihood of the null model and LL_k is the loglikelihood of the current model (Hosmer and Lemeshow, 2000).

According to Hosmer and Lemeshow (2000), unlike in linear regression the R^2 for logistic regression is only used to compare competing models that are used for the same data. A value of 1 is an indication of a perfect fit whilst a value of zero is an indication that there is no relationship. The higher the value the better fit the model.

Likelihood Ratio Test

The Likelihood ratio test tests the significance of all the variables included in logistic regression model. The statistic is given by:

$$-2\log\left(\frac{\mathrm{L}_0}{\mathrm{L}_1}\right) = -2[\mathrm{Log}(\mathrm{L}_1) - \mathrm{Log}(\mathrm{L}_1)] = -2(\mathrm{L}_0 - \mathrm{L}_1)$$

where L_0, is the maximum value for the likelihood function of a simple model and L_1, is the maximum value for the likelihood function of a full model.

The full model will be having all the parameters of interest and the simple model has one variable dropped (Hosmer and Lemeshow, 2000). The likelihood ratio tests the following hypothesis:

H_0 : The dropped variables are not a significant contributor to predicting the dependent variables (that is, $\beta_i = 0$)

H_1 : The dropped variables are important to predicting the dependent variables ($\beta_i \neq 0$).

According to Prempeh, (2009) the likelihood-ratio test is chi-square distributed and if test is significant then the dropped variable will be a significant predictor in the equation whilst on the other hand if the test is not significant then the variable is considered to be unimportant and thus will be excluded from the model.

The Log-likelihood ratio is the difference between the deviance of the null model (model with just the constant) and a model after adding independent variable(s). $Loglikelihood\ Ratio\ = D_{Null} - D_{p-1}$

Where D_{Null} is the deviance of the null model and D_{p-1} is the deviance of a model with $p - 1$ parameters.

Omnibus Test of Model Coefficients

Like the likelihood ratio test statistic, the omnibus test statistic is a measure of the overall model fit. It tests the hypothesis that:

H_0 : All the coefficients of independent variables are equal to zero.

H_1 : There is at least one coefficient of an independent variable that is not equal to zero.

The omnibus test statistic is equivalent to the F-test in linear regression (Lawrence, Gamst, and Guarino, 2006). The null hypothesis is rejected when the p-value of the omnibus test statistic of less than 0.05 (significance level). A significant test statistic implies that the logistic regression can be used to model the data.

Hosmer – Lemshow Goodness of fit test

The Hosmer-Lemeshow goodness-of-fit statistic is another test used to assess the model fit. The test compares the predicted values against the actual values of the dependent variable. The method is similar to the chi-square goodness of fit. The Hosmer-Lemeshow test involves grouping the sample into g groups based on the percentiles of estimated probability (Hosmer and Lemeshow, 2000). The method uses $g = 10$ groups where the first group contains $n'_1 = \frac{n}{10}$ subjects with the lowest probabilities and the last group made up of $n'_{10} = \frac{n}{10}$ subjects with the largest probabilities.

The Hosmer-Lemeshow test is calculated using the formula;

$$\text{Hosmer} - \text{Lemeshow test} \; = \sum_{i=1}^{g} \frac{(O_i - n'_i \hat{\pi}_i{}^2)}{n'_i \hat{\pi}_i (1 - \hat{\pi}_i)}$$

where, n'_i represents the number of observations in the i^{th} group,

$\mathrm{O_i}$ is the observed outcomes in group i, given by: $\mathrm{O_i} = \sum_{j=1}^{c_i} y_j$

C_i denotes the number of covariate patterns in the in the i^{th} group

$\hat{\pi}_\mathrm{i}$ is the estimated probability that an event outcome for group i, and g is the number of groups.

The statistic follows a chi-square distribution with $g - 2$ degrees of freedom (Hosmer and Lemeshow, 2000). A good fit model will have a small Hosmer-Lemeshow test statistic and a p-value that is greater than 0.05 (the significance level).

Classification tables

A Classification table gauges the predictive accuracy of a multivariate logistic regression model. The method involves cross classifying the dependent variable y with the categorical variable emanating from the fitted logistic probabilities ($\hat{y}$). The percentage of successes that have been correctly classified as success is called sensitivity of the model, whilst the percentage of failures that have been correctly classified is called specificity of the model. The failures that are incorrectly classified as success are referred to as false positive and the success that are incorrectly classified as failures are referred to as false negatives (Sharma, 1996). A typical classification tables is as shown below;

Table 3.1: Classification Table

		Predicted		
		Change in Share Price		Percentage Correct
		No Increase (failure)	Increase (success)	
Change in Share Price	No Increase (failure)	a	b	$\frac{a}{a+b}(100)$
	Increase (success)	c	d	$\frac{d}{c+d}(100)$
Overall Percentage				$\frac{a+d}{a+b+c+d}(100)$

In table 3.1, the ratio $\frac{a}{a+b}(100)$ is the specificity of the model, and $\frac{d}{c+d}(100)$ is the sensitivity of the model.

Higher specificity and sensitivity are an indication of a good fit of the model. The classification table will be used for data validation. According to Kutner, Nachtsheim, Neter, and Li (2005) if a model fitting sample produces the same prediction error rate as the validation sample then the fitted model will be reliable.

Akaike's Information Criterion (AIC)

Akaike's Information Criterion (AIC) measures the relative value of a statistical model for a given set of data. The AIC can be used to select the best model. AIC is useless when it is used in isolation as it does not test any hypothesis but can only compare different models. The formula for calculating AIC is:

$$-2L(\beta) + 2(p),$$

where p is the number of parameters in the model plus 1 and L is the log-likelihood of the model given the data.

AIC rewards goodness of fit and penalises and for over fitting. A model with the lowest AIC value will be the most preferable model.

Wald Test

The Wald statistic is another test that can be used to assess the significance of individual logistic regression coefficients. The formula for computing the Wald statistic is;

$$W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)},$$

where, $\hat{\beta}_i$ is the estimate of the coefficient of the independent variable x_i and $SE(\hat{\beta}_i)$ is the standard error of $\hat{\beta}_i$. The squared value of the Wald statistics as indicated below is chi-square distributed with one degree of freedom (Rana, Midi, and Sarkar, 2010).

$$W^2 = \frac{\hat{\beta}_i^{\,2}}{[SE(\hat{\beta}_i)]^2}.$$

The Wald Statistics tests the following hypotheses:

$H_0 : \beta_i = 0, \text{for } i = 1, 2, \dots, p,$ and,

$H_1 : \beta_i \neq 0, \text{for } i = 1, 2, \dots, p.$

The Wald statistic is chi-square distributed with 1 degree of freedom. The null hypothesis is rejected if the p-value of the test is less than 0.05 (significance level). A coefficient with a p-value of the Wald statistic less than 0.05 implies that the variable is important in the model.

Score Test

Score test is one method of assessing the importance of individual independent variables that does not require the calculation of the maximum likelihood estimates of coefficients. According to Thompson (2009) the score test is computed by finding the first and second derivatives of the log likelihood function.

The statistic to test the hypothesis:

$H_0: \beta_k = 0$, and

$H_1: \beta_k \neq 0$, is given by;

$$S(\beta) = \frac{U(\beta_k)^2}{I(\beta_k)},$$

where,

$$U(\beta_k) = \frac{\partial L(\beta_k/x)}{\partial \beta}$$

and

$$I(\beta) = \frac{-\partial^2 L(\beta/x)}{\partial \beta^2}$$

where, L is the log-likelihood function depending on a univariate β and x is the data.

The score test follows a chi-square distribution with one degree of freedom. With the score test, the null hypothesis will be rejected if the p-value of the test is less than 0.05 (the significance level). A coefficient with a p-value of the Score statistic less than 0.05 implies that the variable is important in the model.

Residuals Analysis

Residual analysis in any model is done to assess how best the model fits the data. In logistic regression, the model is of the form

$$\mathrm{y} = \pi\,(\mathrm{x}) + \varepsilon$$

and y can only take values ‘1’ or ‘0’. This implies that

$$\varepsilon = 1 - \hat{\pi}(\mathrm{x}_i)\ for\ Y_i = 1, \text{ and}$$

$$\varepsilon = -\hat{\pi}(x_i)\ for\ Y_i = 0$$

This means that the residuals' distribution under the assumptions of the fitted model is correct is not known (Kutner, Nachtsheim, Neter, and Li, 2005). Thus, the estimated error variance is given by:

$$V(Y|X = x) = \hat{\pi}(x)(1 - \hat{\pi}(x))$$

Dividing the ordinary residual by the estimated standard error Y_i gives the Pearson residual:

$$pr_i = \frac{\hat{\varepsilon}}{\sqrt{\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))}}$$

$$= \frac{Y_i - \hat{\pi}(x_i)}{\sqrt{\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))}}$$

The Pearson residuals do not have unit variance and they are standardized by their estimated standard deviation to produce Studentised Pearson residuals. The Studentised Pearson residuals is calculated as;

$$spr_i = \frac{Y_i - \hat{\pi}(x_i)}{\sqrt{\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))(1 - \hat{h}_{ii})}}$$

$$= \frac{pr_i}{\sqrt{1 - \hat{h}_{ii}}},$$

where $\hat{h}_{ii}$ is the i^th diagonal element of the nxn the matrix,

$$\boldsymbol{H} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{W}}^{1/2}$$

where $\widehat{W}$, is a diagonal matrix with elements $\hat{\pi}(x_i)(1 - \hat{\pi}(x_i)$

$\boldsymbol{X}$ is an the nxp design matrix, $\begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$

Studentised Pearson residuals are valuable in identifying outliers or influential observations and they follow a standard normal distribution for large n (Hosmer and Lemeshow, 2000).

One other residual that is used in logistic regression is the deviance residuals. These residuals are used to identify potential outliers in the model. It is computed as follows;

$$\mathrm{dr_i} = \mathrm{sign}(\mathrm{Y_i} - \hat{\pi}\,(\mathrm{x_i})\{-2[\mathrm{Y_i In}(\hat{\pi}\,(\mathrm{x_i}) + (1 - \mathrm{Y_i})\mathrm{In}(1 - \hat{\pi}\,(\mathrm{x_i})]\}^{1/2}$$

According to Mekonnen, (2011) cases with absolute deviance and standardized residual values greater than 3 may signify a lack of fit.

Cooks distance

Within the package that was used for analysis (SPSS), there is a statistic called the Cook's distance. It quantifies the influence of an observation to the model (that is whether a case is an influential outlier or not). The value of the Cook's distance is a function of the observation's leverage and of the magnitude of its standardised residual. According to (Hosmer and Lemeshow, 2000), the Cook's Distance for logistic regression is estimated by:

$$\Delta\hat{\boldsymbol{\beta}}_j = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-j)})^T \widehat{\boldsymbol{W}}\,(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-j)})^T$$

Where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(-j)}$ are the maximum likelihood estimates for the model with and without the j^{th} observation.

$\widehat{\boldsymbol{W}}$, is a diagonal matrix with elements $\widehat{\pi}\,(\mathrm{x}_i)(1-\widehat{\pi}\,(\mathrm{x}_i)$, and

$\boldsymbol{X}$ is an the nxp design matrix, $\begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$

Observations with standardised residuals greater than 3 and Cook's distance greater than 1 are considered to be influential outliers (Mekonnen, 2011).

3.6: Stepwise Logistic Regression

According to Cramer (2002), the logistic function was invented in the 19th century for the description of populations and the course of autocatalytic chemical reactions. Verhulst published three papers between 1838 and 1847 showing how logistic models agreed very well with the course of the populations of France, Belgium, Essex, and Russia for periods up to 1833. The logistic function was rediscovered in 1920 by Pearl and Reed in modelling the population of the United States for the period 1790 to 1910 (Cramer, 2002). It is believed that Pearl and Reed had no prior knowledge of Verhulst’s work. Today logistic regression is applied in almost every field containing population or categorical response variables such as wildlife, fishing, ecology, epidemiology, plant biology, and public health (Liu, 2009).

Stepwise logistic regression is a systematic method of identifying variables for inclusion or exclusion from a model in a statistical chronological manner. There are mainly two versions of stepwise logistic regression namely forward selection and backward elimination (Kutner, Nachtsheim, Neter, and Li, 2005).

The forward selection method starts with a null or basic model (which includes only the constant, β_0) and adds significant variables to the model. On the other hand, backward elimination method starts with the full model (one including all the possible explanatory variables) and removes insignificant variables from the model (Sarkar, Midi, and Rana 2010).

Sarkar, Midi, and Rana (2010) indicated in their paper that the selection of variables to be included or excluded is a vital consideration when fitting logistic regression models. There is a need to include variables that will result in a model that can be used to make precise predictions at the same time avoiding over-fitting the data. The process of choosing which variables to include in the model is laborious and often not feasible in cases where there are a lot of independent variables. Stepwise regression overcomes such challenges by automating the variable by applying chronological methods.

Stepwise logistic regression is widely used in cases were there many independent variables and it uses a sequence of likelihood ratio test, score test or Wald test to determine the inclusion or exclusion of variables into the model. It can be emphasised that there is no one size fit all model which can be applied in all cases and thus there is a need to apply two or more models to the same study for comparison purposes.

Stepwise Forward Selection (Conditional): This is a stepwise selection technique with starts with a null model and then include more variables one at a time with the test of significance of the new variables being added onto the model assessed using the score statistic. The variable with the most significant score statistic is added to the model first and this process is continued until there is no significant variable left outside the model. The cut-off for significance is p-value = 0.05.

After each variable is added the computer also scrutinises if there is any variable that should be removed. The evaluation of variables for removal from the model is done using the using the probability of the likelihood ratio statistic of conditional parameter estimates.

Stepwise Forward Selection (Likelihood Ratio): This is a stepwise selection technique with starts with a null model and then include more variables one at a time with the test of significance of the new variables being added onto the model assessed using the score statistic. The variable with the most significant score statistic is added to the model first and this process is continued until there is no significant variable left outside the model. The cut-off for significance is p-value = 0.05.

After each variable is added the computer also scrutinises if there is any variable that should be removed. The evaluation of variables for removal from the model is done using the likelihood ratio statistic of conditional parameter estimates. This involves the comparison of the current model to the model after the removal of the variable. If the removal of the variable results in a better fitting model, then the variable is removed otherwise it is kept in the model.

Stepwise Forward Selection (Wald): This is a stepwise selection technique which starts with a null model and the significance of values to be included is tested using the score statistic, and exclusion of undesirable variables is based on the probability of the Wald statistic. Any variable having a significant value of wald statistic is eliminated (significant values are those with values >0.1).

Backward Elimination (Conditional): This is a stepwise selection process which starts with a full model (with all variables) and the variables are excluded from the model using the probability of the likelihood ratio statistic of conditional parameter estimates.

Backward Elimination (Likelihood Ratio): It is a stepwise selection method that starts with a full model and variables are excluded using the probability of the likelihood ratio statistic based on the maximum partial likelihood estimates. This involves the comparison of the current model to the model after the removal of the variable. If the removal of the variable results in a better fitting model, then the variable is removed otherwise it is kept in the model.

Backward Elimination (Wald): This is a stepwise selection method starting with a full model and the insignificant variables are excluded using the probability of the Wald statistic. Any variable having a significant value of wald statistic is eliminated (significant values are those with values >0.1).

Enter: The enter method is a technique for variable selection that involves including all variables at a single step and thus, there is no exclusion involved.

In this research, the Enter, Stepwise backward elimination likelihood Ratio and the stepwise forward selection likelihood ratio methods were used. This is because all the backward selection methods produce the same results and the forward selection methods produce the same results.

3.7: Interpretation of Results

The directionality of the relationship can be determined directly from the logistic coefficients, where the signs (positive or negative) represent the type of relationship between independent and dependent variable. On the other hand the magnitude of the relationship is best determined with the exponentiated coefficient, where the percentage change in the dependent variable (the odds value) is shown by the calculation $Exp(\beta_i)$.

$$= \frac{\hat{\pi}(x_i)}{1 - \hat{\pi}(x_i)} = Exp(\hat{\beta}_i)$$

Where, $\hat{\pi}(x)$ is the probability of success (case) and $1 - \hat{\pi}(x)$ is the probability of failure (non case).

A value less than one indicate that an increase in the independent variable holding other variables constant will result in the outcome less likely to occur whilst a value greater than one indicates that an increase in the independent variable holding other variables constant will result in a high likelihood of occurrence of the outcome. The further the odds ratio is from one, the stronger the relationship. Thus;

when, $\beta_i > 0$, then $\exp(\beta_i) > 1$, implying an increase in odds of success and,

when, $\beta_i < 0$, then $\exp(\beta_i) < 1$, implying a decrease in odds of success

3.8: Summary

This chapter made an account of how logistic regression differs from conventional regression. Statistics such as deviance, Likelihood ratio, Wald Test and Score Test which are used assess the significance of individual coefficients for inclusion or exclusion in a model when carrying out stepwise logistic regression were discussed.

CHAPTER 4: MATERIALS AND METHODS

4.1 Introduction

The chapter presents the variables used in the binary logistic regression, the source of the data, sample size and the software that was used for analysis. Model estimation and validation is also discussed in this chapter. The data was analysed using the IBM SPSS (originally, Statistical Package for the Social Sciences) and now called Statistical Product and Service Solutions, version 20.

4.2 Motivation

This study provides statistical methods that can be used by prospective investors to decide the best shares to invest their money into and also help current shareholders to realign their investment into shares that have higher odds of appreciating in value in future.

Some research work has been initiated in predicting the Share price by a number of researchers. Multiple regression, time series, fuzzy logic, and artificial neural network approach were the most common models used but none of the researchers have used logistic regression. According to Senol (2008), none of the methods explored could accurately forecast the share price behaviour.

Investors are interested in shares that have high odds of appreciating in value, thus in this study logistic regression will be used to determine the parameters that will enhance a share's chances of appreciation value. Logistic regression was chosen because the dependent variable (success or failure of a share price) is binary and non-metric).

4.3 Research Design

The variables dependent and independent variables used in this study, their description and the sample sizes used are outlined below:

4.3.1. Selection of dependent and independent variables

The variables used in the binary logistic regression are summarised in the table 4.1.

Table 4.1: Variables Used

Variable	Dependent/ Independent	Variable Type
Change in Share Price	Dependent	Binary
Assets/Capital Employed	Independent	Metric
Debt/Assets ratio	Independent	Metric
Debt/Equity ratio	Independent	Metric
Dividend Yield%	Independent	Metric
Earnings/ Share(C)	Independent	Metric
Earnings Yield%	Independent	Metric
Operating Profit Margin%	Independent	Metric
Price/ Earnings	Independent	Metric
Return On Assets%	Independent	Metric
Return on Equity%	Independent	Metric
Return on Capital Employed	Independent	Metric

4.3.2. Sample Size

The sample was made up of data from the annual results (financial indicators) and changes in share prices of 472 companies listed on the JSE for the period 2004 to 2011. The secondary data was downloaded from the McGregor BFA website. If a company published its results for the 9 years under review, then it would add 8 cases to the dataset as the researcher is interested in the annual changes in the share price. The change between 2004 and 2005 is a case, then change between 2005 and 2006 will be a different case. Thus, the sample size was supposed to be 8 x 472 companies listed on the Johannesburg stock exchange = 3776 records before cleaning the data. Due to some irregularities such as missing values or incomplete records which were removed from the sample and the fact that some of the companies were listed after 2004, the cleaned data had 1818 records. The sample of 1818 records was big enough since the required sample size for logistic regression is at least 400 cases (Hair, Black, Babin and Anderson, 2010).

The data set was checked for accuracy, integrity, completeness, validity, consistency, uniformity, density and uniqueness. The data set was also split into 60% for model building and 40% for model validation. Thus, the 1818 records were split into 1092 records for model fitting and the other 726 cases for model validation to assess the external validity and practical significance of the model.

4.4 Assumptions

Binary logistic regression is only applied in cases where the dependent variable is dichotomous. This assumption was met because the data was coded as

$$ACSP = \begin{cases} 1 = success \quad if\ annual\ change\ in\ share\ price\ is\ positive \\ 0 = failure \quad if\ annual\ change\ in\ share\ price\ is\ negative \\ \qquad\qquad or\ constant \end{cases}$$

The independent variables can take any form and in this case the independent variables were metric.

The requirement that the sample size should be at least 400 (Hair, Black, Babin and Anderson, 2010) was met since the sample size was 1818 records.

4.5 Model Estimation and Diagnostics

Three methods of model fitting were used for fitting binary logistic regression to establish the variables that are associated with changes in share price. The three methods of model fitting were the Enter method, forward conditional selection, and backward stepwise conditional elimination method. A comparison of the models to determine the best method of model fitting was also conducted.

When using the enter method of model fitting the following steps were followed. $Analyze \rightarrow Regression \rightarrow Binary\ Logistic \rightarrow Enter\ Method$

For the forward conditional selection method, the following steps were followed. $Analyze \rightarrow Regression \rightarrow Binary\ Logistic \rightarrow forward: Conditional\ Method$

For the backward stepwise conditional elimination method, the following steps were followed.

$Analyze \rightarrow Regression \rightarrow Binary\ Logistic \rightarrow Backward: Conditional\ Method$

For all the three model fitting methods ACSP was selected as the dependent variable and the other 11 variables namely: *Assets/Capital Employed*, *Debt/Assets ratio*, *Debt/Equity ratio*, *Dividend Yield%*, *Earnings/ Share(C)*, *Earnings Yield%*, *Operating Profit Margin%*, *Price/ Earnings*, *Return on Assets%*, *Return on Equity%*, and *Return on Capital Employed* were selected as the independent variables. On the save tab, under residuals, standardised and deviance were selected and under influence, cook's was selected as well.

4.6 Adequacy of the Model

A number of statistics were used to assess how the model was fitting the data. The deviance was to assess the goodness of fit of the model. In cases were the deviance had a p-value greater than 0.05, it was concluded that there were some variables in the model that are important in predicting the change in share price (Hosmer and Lemeshow, 2000).

The R^2 was used to compare different models which were using the same data. A model with the highest R^2 value for the data was considered to be the best model because the higher the value the better fit the model is for the data. The R^2 was however not used in isolation, the Akaike's Information Criterion (AIC) was the main model comparing statistic. The model fitting criteria producing the lowest AIC value was considered to be the best method.

Likelihood Ratio Test was used to check whether the variables added to a model were significant in predicting the change in share price. In cases where the p-value of the likelihood ratio test was less than 0.05, all added to the model were considered to be important in predicting the change annual share price.

The omnibus test statistic was used to assess whether there was a linear relationship between the probability of success or failure and the independent variables. An omnibus test statistic p-value less than 0.05 implied that the logistic regression could be used to model the data.

The Hosmer-Lemeshow goodness-of-fit statistic was another test that was used to assess the model fit. The test compares the predicted values against the actual values of the dependent variable. The method is similar to the chi-square goodness of fit. A very small Hosmer-Lemeshow test statistic is desirable and a p-value greater than 0.05 indicates that the model was acceptable.

The Wald statistic was used to assess the importance on individual independent variables in predicting the probability of success or failure of a share price. A coefficient with a Wald statistic p- less than 0.05 implies that the variable is important in the model and those variables with p-values greater than 0.05 were considered to be unimportant.

Observations with modulus of the standardised residuals that were greater than 3 and the cook's distance greater than 1 were considered to be influential outliers and hence excluded from the data and the model refitted without the influential outliers.

The excluded influential observations that were identified when the enter method was applied are shown in Table 4.2.

Table 4.2: Influential Outliers using the Enter Method

Company Serial Number	Deviance Residuals	Standardised Residual	Cook's Distance
232	5.42904	-21.39432	2.91360
544	5.21002	885.44362	5.25920
909	-3.50098	1585.45160	6.95583
914	-3.60931	-25.94668	2.75682

The excluded influential observations that were identified when the forward selection method was applied are shown in Table 4.3.

Table 4.3: Influential Outliers using the Forward Selection Method

Company Serial Number	Deviance Residuals	Standardised Residual	Cook's Distance
544	5.74856	3871.85845	6.06119
909	5.42519	1568.95217	4.84085
914	-4.14570	-73.45142	1.39844
232	-3.55439	-23.51197	3.03731

The excluded influential observations that were identified when the backward elimination method was applied are shown in Table 4.4.

Table 4.4: Influential Outliers using the Backward Elimination Method

Company Serial Number	Deviance Residuals	Standardised Residual	Cook's Distance
909	5.43364	1605.39032	6.07521
544	5.12830	716.84713	4.84986
232	-3.55439	-23.51197	3.03731
914	-3.90426	-45.17595	1.42466

The same observations were identified as influential outliers in all the three model fitting methods.

The removal of influential observations resulted in an improvement of the model fit for the model with all variables (Enter Model). The omnibus tests improved from 103.085 before removing outliers to 163.778 after the removal of outliers. The -2 Log likelihood (goodness of fit test) value for the current model improved from 1378.221 to 1314.694 whilst the Cox & Snell R Square and the Nagelkerke R Square improved by 5% and 6.7% respectively.

The Hosmer-Lemeshow test statistic improved from 39.931 before removing outliers to 19.896 after the removal. The overall full model correct classification was improved from 66.5% to 68.3% after removal of influential outliers. All these improvements signify an improvement in the model fit after the removal of the influential outliers. The changes are shown in Table 4.5

Table 4.5: Diagnostics after removing Influential Outliers (Enter Method

	Before Removing Outliers	After Removing Outliers
Omnibus Tests of Model Coefficients	103.085	163.778
-2 Log likelihood	1378.221	1314.694
Cox & Snell R Square	0.09	0.14
Nagelkerke R Square	0.121	0.188
Hosmer and Lemeshow Test	39.931	19.896
Predicted Power	66.5	68.3

When using the forward selection the removal of influential observations resulted in an improvement of the model fit for the model. The omnibus tests improved from 83.8 before removing outliers to 146.957 after the removal of outliers. The -2 Log likelihood (goodness of fit test) value for the current model improved from 1397.506 to 1331.516 whilst the Cox & Snell R Square and the Nagelkerke R Square improved by 5.2% and 7.1% respectively.

The Hosmer-Lemeshow test statistic improved from 34.524 before removing outliers to 16.026 after the removal. The overall full model correct classification was improved from 66.2% to 67.2% after removal of influential outliers. All these improvements signify an improvement in the model fit after the removal of the influential outliers. The changes are shown in Table 4.6.

Table 4.6: Diagnostics after removing Influential Outliers (Forward Selection Method

	Before Removing Outliers	After Removing Outliers
Omnibus Tests of Model Coefficients	83.8	146.957
-2 Log likelihood	1397.506	1331.516
Cox & Snell R Square	0.074	0.126
Nagelkerke R Square	0.099	0.17
Hosmer and Lemeshow Test	34.524	16.026
Predicted Power	66.2	67.2

On application of the backward elimination method, the removal of the influential outliers resulted in the omnibus tests improving from 95.82 to 155.14. The Cox & Snell R-Square improved from 8.4% before the removal of outliers to 13.3% after the removal of outliers and the Nagelkerke R Square also improved from 11.3% to 17.9% respectively. The Hosmer-Lemeshow test statistic improved from 36.321 before removing outliers to 26.499 after the removal. The overall full model correct classification improved from 66.7% to 68.3% after removal of influential outliers. All these improvements signify an improvement in the model fit after the removal of the influential outliers. The changes are shown in Table 4.7.

Table 4.7: Diagnostics after removing Influential Outliers (Backward Elimination Method)

	Before Removing Outliers	After Removing Outliers
Omnibus Tests of Model Coefficients	95.82	155.14
-2 Log likelihood	1385.49	1323.34
Cox & Snell R Square	0.08	0.13
Nagelkerke R Square	0.11	0.18
Hosmer and Lemeshow Test	36.32	26.50
Predicted Power	66.70	68.30

4.7 Validation of Results

The final step is validation of the results. At this stage the validation sample will be used to assess the external validity and practical significance of the model. The predictive power of the fitted model is assessed by comparing the correct classification percentage for the two samples. If the model produces almost the same classification accuracy for the model fitting sample and the validation sample then the models is said to be accurate/ valid.

4.8 Summary

The variables used in the binary logistic regression, the source of the data, the sample size, assumptions of the model, model estimation and diagnostics, adequacy of the model and how the results were validated was discussed in this chapter. The next chapter will present the results and findings.

CHAPTER 5: ANALYSIS AND DISCUSSION OF RESULTS

5.1: Introduction

The results of the study are presented in this chapter. Three methods of model fitting were used for fitting multivariable binary logistic regression to establish the variables that are associated with changes in share price. The three methods of model fitting were the Enter method, forward conditional selection, and backward stepwise conditional elimination method. A comparison of the models to determine the best method of model fitting was also conducted using AIC.

5.2: Logistic Regression with all variables (The Enter Method)

5.2.1. Omnibus Tests of Model Coefficients

The enter method of model fitting which involves the entering of all variables at the same step. The results in Table 5.1 show the model chi-square and the significance levels for test of the null hypothesis that all the coefficients are equal to zero.

Table 5.1: Omnibus Tests of Model Coefficients

Omnibus Tests of Model Coefficients				
Model		**Chi-square**	**df**	**Sig.**
Enter	Step	163.778	11	.000
	Block	163.778	11	.000
	Model	163.778	11	.000

The model chi-square value which is the difference between the null model and the current (full) (chi-square values =163.778), the null hypothesis is rejected since the p-value (sig. value in Table 5.1) is less than 0.05 (significance level), implying that the addition of the independent variables improved the predictive power of the model. The block and the step vales are equal to the model values since all values were entered at the same time.

5.2.2. Model Summary

Model summary have values shown in Table 5.2 indicate how good the model fits the data. The -2 Log likelihood (goodness of fit test) value for the current model is 1314.694 and that of the null model was 1334.590, a decrease of 19.896 indicating an improvement in the model after the addition of the independent variables. This implies that the addition of the variables fitted in the model improved the prediction power of the model.

Table 5.2: Model Summary

Model Summary				
Model		-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
Enter	Null Model	1334.590		
	Final Model	1314.694	.140	.188

The Cox & Snell R Square which is an attempt to provide a logistic regression equivalent to the coefficient of determination in multiple regression, hence the name pseudo-R statistic. This value was low at 14% implying a poor fit. The Nagelkerke R Square which adjusts the Cox & Snell R-square so that it ranges from ‘0’ to ‘1’ was 18.8%. These values were low signifying a poor fit of the model but there is caution when using these values because they do not explain the amount of variation accounted for by the model as does the R-square in multiple regression (Hosmer and Lemeshow, 2000).

5.2.3. Hosmer and Lemeshow Test

The Hosmer-Lemeshow test shown in Table 5.3 explores whether the predicted probabilities are the same as the observed probabilities. An overall goodness of fit of the model is indicated by p-values > 0.05 (Hosmer and Lemeshow, 2000). This model produced a significant difference between the observed and predicted probabilities indicating a poor model fit.

Table 5.3: Hosmer and Lemeshow Test

Hosmer and Lemeshow Test			
Model	**Chi-square**	**df**	**Sig.**
Enter	19.896	8	.011

5.2.4. Interpretation of the Model

The fitted model using the enter method is in Table 5.4:

$$\text{In}\left[\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right] = 0.22517 - 0.02141x_1 - 0.12783x_2 + 0.00741x_3 - 0.06833x_4 + 0.00033x_5 - 0.00103x_6 + 0.00005x_7 + 0.00036x_8 - 0.00946x_9 + 0.00099x_{10} + 0.04925x_{11},$$

where x_1 is *assets/ capital employed,* x_2 *is debt/ assets ratio, ,* x_3 *is debt/ equity,* x_4 *is dividend yield,* x_5 *is earnings per share,* x_6 is *earnings yield,* x_7 is *operating profit margin,* x_8 is *price earnings,* x_9 is *return on assets,* x_{10} is *return on equity, and* x_{11} *is return on capital employed*

The coefficient of *assets/ capital employed* as shown in Table 5.4 was -0.02141, this implies that $exp(\beta)$ = $\exp(-0.02141) \approx 0.97882$. Thus, a unit increase in assets/ *capital employed* leads *to* a decline of (0.97882-1) x 100% = 2.12% in the odds of increase in *share price*. Thus, a high value of assets / capital employed is associated with a decrease in share price.

The coefficient of *debt/ assets was* -0.12783, this implies that $exp(\beta)$ = $\exp(-0.12783) \approx 0.88000$. Thus, a unit increase in *debt/ assets* leads *to* a decrease of (0.88000-1) x 100% = 12% in the odds of an increase in *share price*. Thus, a high value of Debt /Assets is associated with a decrease in the in *share price*.

The coefficient of *debt / equity was* 0.00741, this implies that $exp(\beta)$ = $\exp(0.00741) \approx 1.00744$. Thus, a unit increase in *debt / equity* leads *to* an increase of ($1.00744 - 1$) x 100% = 0.74% in the odds of increase in *share price*. Thus, a high value of *debt / equity* is associated with an increase in the in *share price*.

The coefficient of d*ividend yield* was -0.06833, this implies that $exp(\beta)$ = $\exp(-0.06833) \approx 0.93395$. Thus, a unit increase in *dividend yield* leads *to* a decrease of ($0.93395 - 1$) x 100% = 6.61% in the odds of increase in *share price*. Thus, a high value of *dividend yield* is associated with a decrease in *share price*.

The coefficient of *earnings / share* was 0.00033, this implies that $\exp(\beta)$ = $\exp(0.00033) \approx 1.00033$. Thus, a unit increase in *earnings / share* leads *to* an increase of ($1.00033 - 1$) x 100% = 0.03% in the odds of increase in *share price*.

Thus, a high value of *earnings / share* is associated with an increase in *share price*.

The coefficient of *earnings yield* was -0.00103, this implies that $exp(\beta)$ = $\exp(-0.00103) \approx 0.99897$. Thus, a unit increase in *earnings yield* leads *to* a decrease of ($0.99897 - 1$) x 100% = 0.10% in the odds of increase in *share price*. Thus, a high value in *earnings yield* is associated with a decrease in *share price*.

The coefficient of *operating profit margin* was 0.00005, this implies that $\exp(\beta)$ = $\exp(0.00005) \approx 1.00005$. Thus, a unit increase in *operating profit margin* leads *to* an increase of ($1.00005 - 1$) x 100% = 0.01% in the odds of increase in *share price*. Thus, a high value of *operating profit margin* is associated with an increase in *share price*.

The coefficient of *price earnings* was 0.00036, this implies that $\exp(\beta)$ = $\exp(0.00036) \approx 1.00036$. Thus, a unit increase in *price earnings* leads *to* an increase of ($1.00036 - 1$) x 100% = 0.04% in the odds of increase in *share price*. Thus, a high value of *price earnings* is associated with an increase in share price.

The coefficient of *return on assets* was -0.00946, this implies that $exp(\beta)$ = $\exp(-0.00946) \approx 0.99059$. Thus, a unit increase in *return on assets* leads *to* a decrease of $(0.99059 - 1)$ x 100% = 0.94% in the odds of increase in *share price*. Thus, a high value in *return on assets* is associated with a decrease in *share price*.

The coefficient of *return on equity* was 0.00099, this implies that $\exp(\beta)$ = $\exp(0.00099) \approx 1.00099$. Thus, a unit increase in *return on equity* leads *to* an increase of $(1.00099 - 1)$ x 100% = 0.10% in the odds of increase in *share price*. Thus, a high value of *return on equity* is associated with an increase in *share price.*

The coefficient of *return on capital employed* was 0.04925, this implies that $\exp(\beta)$ = $\exp(0.04925) \approx 1.05049$. Thus, a unit increase in *return on capital employed* leads *to* an increase of $(1.05049 - 1)$ x 100% = 5.05% in the odds of increase in *share price*. Thus, a high value of *return on capital employed* is associated with an increase in *share price*.

Table 5.4: Variables in the Equation

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Assets / capital employed	-.02141	.069	.095	1	.758	.97882
	Debt / assets	-.12783	.116	1.223	1	.269	.88000
	Debt / equity	.00741	.005	1.864	1	.172	1.00744
	Dividend yield	-.06833	.016	17.547	1	.000	.93395
	Earnings / share	.00033	.000	6.583	1	.010	1.00033
	Earnings yield	-.00103	.001	3.510	1	.061	.99897
	Operating profit margin	.00005	.000	.232	1	.630	1.00005
	Price earnings	.00036	.000	.578	1	.447	1.00036
	Return on assets	-.00946	.007	2.107	1	.147	.99059
	Return on equity	.00099	.000	3.995	1	.046	1.00099
	Return on capital employed	.04925	.007	45.463	1	.000	1.05049
	Constant	.22517	.157	2.065	1	.151	1.25253

a. Variable(s) entered on step 1: Assets / capital employed, Debt / assets, Debt / equity, Dividend yield, Earnings / share, Earnings yield, Operating profit margin, Price earnings, Return on assets, Return on equity, Return on capital employed.

The Wald statistics and the significance level shows that 4 out of the 11 independent variables namely; *dividend yield*, *earnings/ share, return on equity*, and *return on capital employed* were significant to the prediction of the odds of an increase in share price. This is because they had p-values values of less than 0.05 (sig. in Table 5.4).

Al-Dini, Dehavi, Zarezadeh, Armesh, Manafi, and Zraezadehand (2011), concluded that there was a negative relationship between *dividends per share* (DPS) and *share price* which supports the results found in this study.

The fact that the results showered that the change in share price is determined by *dividend yield*, *earnings per share, return on equity*, and *return on capital employed* is supported by Khan, Aamir, Qayyum, Nasir, and Khan (2011) when they found that *share price* was positively related to *earnings per share.* Their results were however different in that they concluded that dividend yield was positively related to share price yet in this research it was found to be negatively related to share price.

5.2.5. Classification Table

A classification table which indicates how well the model predicts cases to the two dependent variable categories displayed in Table 5.5. The sample was randomly split into a model fitting sample and a validation sample. The classification table was conducted for both the model fitting sample and the validation sample. The specificity, which is the proportion of the correctly classified “no increase” in share price was 36.1% (for the model fitting sample) and the sensitivity which is the proportion of the correctly classified “increase” in share price was 91.1%. The overall full model correct classification was 68.3%. The validation sample had a correct classification of 65.2%.

Table 5.5: Classification Table

Classification Table								
Observed			Predicted					
			Model Fitting Sample			Validation Sample		
			Change in Share Price		Percentage Correct	Change in Share Price		Percentage Correct
			No Increase	Increase		No Increase	Increase	
Enter	Change in Share Price	No Increase	163	288	36.1	85	197	30.1
		Increase	57	582	91.1	55	387	87.6
	Overall Percentage				**68.3**			**65.2**
a. The cut value is .500								
b. Model Fifing Sample Approximately 60% of the cases (SAMPLE) EQ 1								
c. Validation Sample cases Approximately 60% of the cases (SAMPLE) NE 1								

5.2.6. Model Validation

Based on the classification accuracy of the fitted model, for both the model fitting sample and validation sample, it was observed that the correct classification was almost the same. The classification accuracy of the validation sample was only 3.1% less than that of the model fitting sample (65.2% and 68.3% respectively) (see Table 5.5). Thus, it can be concluded that the model was valid and can be replicated.

5.3: Logistic Regression with Stepwise Forward Selection Method

5.3.1. Omnibus Tests of Model Coefficients

The stepwise forward selection method of model fitting which starts with a null model and then variables are entered one by one into the model based on their significance as measured by the score statistic, likelihood ratio statistic and deviance. The results in Table 5.6 show the model chi-square and the significance levels for test of the null hypothesis that all the coefficients are equal to zero.

Table 5.6: Omnibus Tests of Model Coefficients

Omnibus Tests of Model Coefficients				
Model		**Chi-square**	**df**	**Sig.**
Forward Stepwise (Conditional)	Step	33.475	1	.000
	Block	146.957	2	.000
	Model	146.957	2	.000

The model chi-square value which is the difference between the null model and the current (full) model value was 146.957. The null hypothesis is rejected since the significance level is less than 0.05 (significance level), implying that the addition of the independent variables improved the predictive power of the model.

5.3.2. Model Summary

Model summary have values shown in Table 5.7 indicate how good the model fits the data. The -2 Log likelihood (goodness of fit test) value for the current model is 1331.516 and that of the null model was 1347.542, a decline of 16.026 indicating an improvement in the model after the addition of the independent variables. This implies that the addition of the variables fitted in the model improved the prediction power of the models.

Table 5.7: Model Summary

Model Summary				
Model		-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
Forward Stepwise (Conditional)	Null Model	1347.542		
	Final Model	1331.516	.126	.170

The Cox & Snell R Square was low at 12.6% and the Nagelkerke R Square which adjusts the Cox & Snell R Square so that it ranges from ‘0’ to ‘1’ was 17.0%. These values were low signifying a poor fit for the model but there is caution when using these values because they do not explain the amount of variation accounted for by the model as does the R-square in multiple regression (Hosmer and Lemeshow, 2000).

5.3.3. Hosmer and Lemeshow Test

The Hosmer-Lemeshow test shown in Table 5.8 explores whether the predicted probabilities are the same as the observed probabilities. An overall goodness of fit of the model is indicated by insignificant chi-square values (p-values > 0.05). This model produced a significant difference between the observed and predicted probabilities indicating a poor model fit which is not desired.

Table 5.8: Hosmer and Lemeshow Test

Hosmer and Lemeshow Test			
Model	**Chi-square**	**df**	**Sig.**
Forward Stepwise (Conditional)	16.026	8	.042

5.3.4. Interpretation of the Model

The fitted model using the stepwise forward selection method is in Table 5.9

$\text{In}\left[\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right] = 0.16463 - 0.06302x_4 + 0.04434\, x_{11}$,

Where x_4 is *dividend yield, and* x_{11} is *return on capital employed*

The coefficient of d*ividend yield* was -0.06302, this implies that $exp(\beta)$ = $\exp(-0.06302) \approx 0.93893$. Thus, a unit increase in *dividend yield* leads *to* a decrease of $(0.93893 - 1)$ x 100% = 6.11% in the odds of increase in *share price*. Thus, a high value of *dividend yield* is associated with a decrease in share price.

The coefficient of *return on capital employed* was 0.04434, this implies that $exp(\beta)$ = $\exp(0.04434) \approx 1.04534$. Thus, a unit increase in *return on capital employed* leads *to* an increase of (1.04534-1) x 100% = 4.53% in the odds of increase in *share price*. Thus, a high value of *return on capital employed* is associated with an increase in share price.

The model retained only 2 out of the 11 independent variables namely; *dividend yield*, and *return on capital employed.* The rest of the variables were insignificant.

This result is supported by the findings that were found by Al-Dini, Dehavi, Zarezadeh, Armesh, Manafi, and Zraezadehand (2011), when the concluded that there was a negative relationship between *dividends per share* (DPS) and *share price*. This is however different from what was found by Khan, Aamir, Qayyum, Nasir, and Khan (2011), According to Matthew and Odularu (2009), Al-Dini, Dehavi, Zarezadeh, Armesh, Manafi, and Zraezadehand (2011), and Azam and Kumar (2011) when the indicated that *stock price* was positively related to *dividend yield*.

From the researches that were reviewed in this study none of them found *return on capital* to be positively correlated to *changes in share price*.

Table 5.9: Variables in the Equation

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 2[b]	Dividend yield	-.06302	.016	16.261	1	.000	0.93893
	Return on capital employed	.04434	.005	81.044	1	.000	1.04534
	Constant	.16463	.087	3.611	1	.057	1.17896
a. Variable(s) entered on step 1: Return on capital employed.							
b. Variable(s) entered on step 2: Dividend yield.							

5.3.5. Classification Table

The specificity for the forward stepwise model was 32.8% and the sensitivity was 91.5%. Overall full model correct classification was 66.2%. The validation sample had a correct classification of 67.2%. The validation sample had a correct classification of 65.3%, the results are shown in Table 5.10.

Table 5.10: Classification Table

Classification Table								
Observed			Predicted					
			Model Fitting Sample			Validation Sample		
			Change in Share Price		Percentage Correct	Change in Share Price		Percentage Correct
			No Increase	Increase		No Increase	Increase	
Forward Stepwise (Conditional)	Change in Share Price	No Increase	148	303	32.8	86	196	30.5
		Increase	54	585	91.5	55	387	87.6
	Overall Percentage				**67.2**			**65.3**
a. The cut value is .500								
b. Model Fitting Sample Approximately 60% of the cases (SAMPLE) EQ 1								
c. Validation Sample Approximately 60% of the cases (SAMPLE) NE 1								

5.3.6. Model Validation

The classification accuracy of the validation sample using the forward selection method was 1.9% less than that of the model fitting sample (65.3% and 67.2% respectively see Table 5.10). Thus, the model fitting and the validation samples produced almost the same classification accuracy and hence the model is valid.

5.4: Model Fitting Using Stepwise Backward Selection Method

5.2.1. Omnibus Tests of Model Coefficients

The stepwise selection method starts with a model with all the variables and eliminates them one by one depending on the significance of their coefficients. The results in Table 5.11 indicate the model chi-square and the p-values for test of the null hypothesis that all the coefficients are equal to zero.

Table 5.11: Omnibus Tests of Model Coefficients

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 8[a]	Step	-2.455	1	.117
	Block	155.135	4	.000
	Model	155.135	4	.000
a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.				

The model chi-square value which is the difference between the null model and the current (full) model chi-square value was (155.135). The null hypothesis is rejected since the p-value (sig. in Table 5.11) is less than 0.05 (significance level), implying that the addition of the independent variables improved the predictive power of the model.

5.2.2. Model Summary

The -2 Log likelihood values for the stepwise forward selection shown in Table 5.12 indicate how good the model fits the data. The -2 Log likelihood (goodness of fit test) value for the current model is 1323.338 and that of the null model was 1349.837, a decline of 26.499 indicating an improvement in the model after the addition of the independent variables. This implies that the addition of the variables resulted in an improvement on the model fit.

Table 5.12: Model Summary

Model Summary				
Model		-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
Backward Stepwise (Conditional)	Null Model	1349.837		
	Final Model	1323.338[a]	.133	.179

The Cox & Snell R Square which was 13.4% and the Nagelkerke R Square was 17.9% (Table 5.12). These values were low signifying a poor fit for the model. There is a caution when using these values because they do not explain the amount of variation accounted for by the model as does the R-square in multiple regression.

5.2.3. Hosmer and Lemeshow Test

The Hosmer-Lemeshow test shown in Table 5.13 explores whether the predicted probabilities are the same as the observed probabilities. An overall goodness of fit of the model is indicated by insignificant chi-square values (p-values > 0.05). This model produced a significant difference between the observed and predicted probabilities indicating a poor model fit (p-value = 0.001).

Table 5.13: Hosmer and Lemeshow Test

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
8	26.499	8	.001

5.2.4. Interpretation of the Model

The fitted model using the backward stepwise selection method is in Table 5.14.

$$\text{In}\left[\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right] = 0.11696 - 0.06826\, x_4 + 0.00032 x_5 - 0.00100\, x_6 + 0.04214 x_{11}\ ,$$

where x_4 *is dividend yield,* x_5 *is earnings per share,* x_6 is *earnings yield, and* x_{11} *is return on capital employed*

The coefficient of *earnings / share* was 0.00032, this implies that $exp(\beta)$ = $\exp(0.00032) \approx 1.00032$. Thus, a unit increase in *earnings / share,* holding other variables constant leads to an increase of $(1.00032 - 1)$ x 100% = 0.03% in the odds of increase in *share price*. Thus, a high value of *earnings / share* is associated with an increase in share price. This result is consistent with the results that were found Khan, Aamir, Qayyum, Nasir, and Khan (2011), Midan (1991), Al-Dini, Dehavi, Zarezadeh, Armesh, Manafi, and Zraezadehand (2011), and Azam and Kumar (2011). These authors concluded that there was a positive relationship between *earnings per share* and *change in share price.*

The coefficient of *return on capital employed* was 0.04214, this implies that $exp(\beta)$ = $\exp(0.04214\ ) \approx 1.04304$. Thus, a unit increase in *return on capital employed* holding other variables constant leads to an increase of $(1.04304 - 1)$ x 100% = 4.30% in the odds of increase in *share price*. Thus, a high value of *return on capital employed* is associated with an increase in *share price*. From the researches that were reviewed in this study none of them found *return on capital* to be positively correlated to *changes in share price*.

The coefficient of *earnings yield* was -0.00100, this implies that $exp(\beta)$ = $\exp(-0.00100\ ) \approx 0.99900$. Thus, a unit increase in *earnings yield* holding other variables constant leads *to* a decrease of $(0.99900 - 1)$ x 100% = 0.10% in the odds of increase in *share price*. Thus, a high value in *earnings yield* is associated with a decrease in *share price*.

The coefficient of *dividend yield* was -0.06826, this implies that $exp(\beta)$ = $\exp(-0.06826\,) \approx\ 0.93402$. Thus, a unit increase in *dividend yield* holding other variables constant leads *to* a decrease of $(0.93402 - 1)$ x 100% = 6.60% in the odds of increase in *share price*. Thus, a high value in *dividend yield* is associated with a decrease in *share price*. This result is supported by the findings that were found by Al-Dini, Dehavi, Zarezadeh, Armesh, Manafi, and Zraezadehand (2011), when the concluded that there was a negative relationship between *dividends per share* (DPS) and *share price*. This is however different from what was found by Khan, Aamir, Qayyum, Nasir, and Khan (2011), According to Matthew and Odularu (2009), Al-Dini, Dehavi, Zarezadeh, Armesh, Manafi, and Zraezadehand (2011), and Azam and Kumar (2011) when the indicated that *stock price* was positively related to *dividend yield*.

Table 5.14: Variables in the Equation

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 8[a]	Dividend yield	-.06826	.016	17.628	1	.000	.93402
	Earnings / share	.00032	.000	6.468	1	.011	1.00032
	Earnings yield	-.00100	.001	3.398	1	.065	.99900
	Return on capital employed	.04214	.005	73.620	1	.000	1.04304
	Constant	.11696	.088	1.758	1	.185	1.12407
a. Variable(s) entered on step 1: Assets / capital employed, Debt / assets, Debt / equity, Dividend yield, Earnings / share, Earnings yield, Operating profit margin, Price earnings, Return on assets, Return on equity, Return on capital employed.							

The rest of the variables were excluded from the model and thus, only *dividend yield*, *earnings / share*, *earning yield* and *return on capital employed* were significant in predicting the odds of an increase in share price. This implies that Assets / capital employed, Debt / assets, Debt / equity, Operating profit margin, Price earnings ratio, Return on assets, Return on equity were not important in predicting ACSP.

This is supported by what Khan, Aamir, Qayyum, Nasir, and Khan (2011) found when they concluded that *return on equity* was not a significant contributors to *stock price*. On the other hand this s also contradicting with what Al-Dini, Dehavi, Zarezadeh, Armesh, Manafi, and Zraezadehand (2011) found when they concluded that there was a negative relationship between *price to earnings ratio* and *stock price*.

5.2.5. Classification Table

The specificity for the backward stepwise model was 35.7% and the sensitivity was 91.4%. Overall full model correct classification was 68.3%. The validation sample had a correct classification of 65.2% (results in Table 5.15).

Table 5.15: Classification Table

Classification Table[a]								
	Observed		Predicted					
			Selected Cases[b]			UnModel Fitting Sample		
			Change in Share Price		Percentage Correct	Change in Share Price		Percentage Correct
			No Increase	Increase		No Increase	Increase	
Step 8	Change in Share Price	No Increase	161	290	35.7	89	193	31.6
		Increase	55	584	91.4	59	383	86.7
	Overall Percentage				68.3			65.2
a. The cut value is .500								
b. Model Fitting Sample Approximately 60% of the cases (SAMPLE) EQ 1								
c. Validation Sample Approximately 60% of the cases (SAMPLE) NE 1								

5.2.6. Model Validation

Based on the classification accuracy of model fitting sample and validation sample, it was observed that the correct classification were almost the same. This model had a difference of 3.1% between the model fitting sample and the validation sample as shown in Table 5.15. Thus, since there was a difference of 3.1% only between the model fitting and the validation sample, it can be concluded that the model was valid and can be replicated.

5.5: Comparison of the three Methods of Model Fitting

Table 5.16 shows the comparison of the methods of model fitting. Based on the Akaike's Information Criterion (AIC) measure which takes into consideration the log-likelihood value and the number of variables retained in the model, the backward selection method produced the best model (AIC = 1333.385 compared to 1338.694 for the enter method and model and 1337.516 for forward selection).

The variables that were commonly significant in the three models produced by all the three methods of model fitting are *dividend yield* and *return on capital employed*. *Earnings per share* was significant when the enter and the backward selection methods were used, while *return on equity* only was significant when the enter method was used. *Earnings yield* was significant with the backward selection method only.

Table 5.16: Comparison of the three Models

		Model 1	Model 2	Model 3
		Enter	**Forward Selection**	**Backward Selection**
Model Summary	-2 Log likelihood	1314.694	1331.516	1323.338
	Cox & Snell R Square	14%	13%	13%
	Nagelkerke R Square	19%	17%	18%
Classification Accuracy	Model Fitting Data	68.3%	67.2%	68.3%
	Validation Data	65.2%	65.3%	65.2%
Significant Variables		Dividend yield	Dividend yield	Dividend yield
		Earnings / share		Earnings / share
				Earnings yield
		Return on equity		
		Return on capital employed	Return on capital employed	Return on capital employed
AIC		1338.694	1337.516	1333.385

Since the backward selection method of model fitting was found to produce the best model in comparison to the other two methods, the *annual change in share price* is determined by *dividend yield*, *earnings per share*, *earnings yield* and *return on capital employed*.

5.6: Summary

In this chapter, the results were presented and three models were compared to evaluate which model produces the best results. It was found that the backward selection method of model fitting produced the best fit for the data. Based on the best model it was found that *Annual Change in Share Price (ACSP)* is determined by *dividend yield*, *earnings per share*, *earnings yield* and *return on capital employed*.

CHAPTER 6: SUMMARY, CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

A literature review of past research in Chapter 2 showed that there are a number of factors that are associated with the changes in share price as found by other researchers who applied a variety of statistical methods. The theory of Logistic regression was presented in Chapter 3, variables used, the source of data and the sample size were presented in chapter 5. Chapter 5 saw the application of Binary logistic regression to predict the odds of success of a share. This chapter presents a summary of the methodology, findings of this study, recommendations to the investors and identifies areas of further study.

6.2 Summary

A number of researches were conducted prior to this research to try and establish the determinants of share price changes. Most of the researchers were interested in the actual growth of the share which was a continuous variable and thus method such as multiple regression analysis and time series analysis were applied. The difference between this research and the rest of the researches is that the researcher is interested in helping the investor to establish the factors that lead to an increase in the value of their portfolios. So the dependent variable was change in share price. The share price either increases or does not increase which is a dichotomous variable. Binary Logistic regression was found to be the model that could be applied to such a variable as the dependent could not meet the assumptions that should be satisfied for methods like multiple regression to be fitted.

SPSS (originally, Statistical Package for Social Sciences and now called Statistical Product and Service Solutions) was used to conduct the Binary logistic regression using the backward selection method of model fittings. The backward stepwise logistic regression started with a model with all the variables and excluded the variables with insignificant coefficients until the model was at its best predictive power.

6.3 Conclusions and Findings

The objectives of the study were to

1. To fit a logistic model to the annual change in stock price
2. To determine the adequacy of the fitted model, and
3. To compare and determine the results of binary logistic regression to stepwise logistic regression, backward elimination, and the enter method of model fitting.

Fitting a logistic model to annual change in stock Price

Factors associated with annual changes in the share price of JSE listed companies using Binary logistic regression model were studied. The independent variables that were used in the model are *assets/ capital employed, debt /assets, debt /equity ratio, dividend yield, earnings /share, earnings yield, operating profit margin, price earnings, return on assets, return on equity,* and *return on capital employed*.

The analysis of the significance of the logistic coefficients was done using likelihood ratio and Wald test. It was established that the probability of success of share is higher if the shareholders are anticipating a higher *return on capital employed,* and higher *earnings/ share*. It was however, noted that the *share price* is negatively impacted by *dividend yield and earnings yield*. Thus, the higher the *dividend yield* and/ or earnings yield, the lower the likelihood of the share price to appreciate and vice versa.

To determine the adequacy of the fitted model

The mode could correctly classify 68.3% of the changes in share prices. The validation predicted 65.2% of the changes in share price. The model was considered to be valid since both the model fitting and the validation sample produced almost the same classification accuracy.

6.4 Recommendations

It is recommended that:

- Since the odds of success of share price is higher if there is a higher return on capital employed and high earning per share, investors and investment companies are encouraged to choose companies with high earnings per share and the best returns on capital employed.
- The fact that the share price is negatively impacted by Dividend yield could be due to the fact that a company would give out part of its profits as dividends and thus not ploughing it back into the business and thus not increasing the

net worth of the shares. Dividends are a good source of income to the shareholders but if an investor is interested in Capital growth, they should buy shares of companies with high earnings/ share and high returns on capital employed, and do not pay dividends.

- The annual change in share price was found to be negatively related to earnings *yield*. This is so because as the price of a share goes up at a rate higher than that of the profits after tax, then a the high share price will mean a smaller earnings yield since the earnings yield is found by dividing *earnings per share* by *share price*. Thus, investors are encouraged to buy share with low earnings yield since the share prices will be going up at a rate higher which might signify a high demand for the shares.

6.5 Areas of Further study

Areas of further study;

- The study should be carried out in a different time period (not 2004 -2011) since the data included financial ratios for the time when a global recession was experienced and hence might have influenced the observed pattern.
- Replicate the study using data from a different stock exchange such as the America's National Association of Securities Dealers Automated Quotations (NASDAQ), New York Stock Exchange also in America, Tokyo Stock Exchange in Japan or Britain's London Stock Exchange to check if the model is also applicable in those markets.

REFERENCES

Abdelrahman A.I., 2010. *Applying Logistic Regression Model to The Second Primary Cancer Data*. Ain Shams University. Egypt.

African Capital News., 2010. *More trading records fall as JSE passed 200k trades in a day (7/05/2010)*, http://www.africancapitalmarketsnews.com/470/more-trading-records-fall-as-jse-passed-200k-trades-in-a-day/ accessed on 02/02/2010.

Agresti A., 2007. Introduction to Categorical Data Analysis. Wiley & Sons Inc. Hoboken. New Jersey. USA.

Al-Dini S.H.M. Dehavi H.D. Zarezadeh E. Armesh H. Manafi M. and Zraezadehand S., 2011. *Fitting the Relationship between Financial Variables and Stock Price through Fuzzy Regression Case study: Iran Khodro Company*. International Journal of Business and Social Science. Vol. 2 No. 11.

Al-Ghamdi A.S., 2001. *Using logistic regression to estimate the influence of accident factors on accident severity*. King Saud University. Pergamon.

Assefa. R., 2011. *Factors Affecting The Practice Of Female Genital Mutilation Of Ethiopian Women.* University of Addis Ababa. Ethiopia.

Azam M. Kumar D., 2011. *Factors Influencing the Individual Investor and Stock Price Variation: Evidence from Karachi Stock Exchange*. Australian Journal of Basic and Applied Sciences. 5(12): 3040-3043, 2011.

Cramer J.C., 2002. *The Origins of Logistic Regression*, Tinbergen Institute Discussion Paper.

Dayton. C. M., 1992. *Logistic Regression Analysis*, University of Maryland.

D'Amato E., 2010. *The Top 15 Financial Ratios*. The Australian Shareholders' Association. Chatswood NSW 2057.

Hair Jr. J. F. Black W. C. Babin B. J. and Anderson R. E., 2010. *Instructor's Manual for Multivariate Data Analysis:* A Global Perspective. Pearson.

Harrell F. E, Jr., 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer.

Hosmer D. W., and Lemeshow S., (2000)., *Applied Logistic Regression, Second Edition*. John Wiley & Sons

Johannesburg Stock Exchange (JSE), 2011. *The Stock Market for Beginners.* Johannesburg, South Africa.

Khan K.I., Aamir M. Qayyum A. Nasir A. and Khan M.I., 2011. *Can Dividend Decisions Affect the Stock Prices: A Case of Dividend Paying Companies of KSE*. International Research Journal of Finance and Economics.

Kleinbaun D.G., Kupper L.L., Nizam A., and Muller K. E., 2008., *Applied Regression Analysis and other Multivariable Methods., Fourth Edition.*, Thomson Brooks/Cole.

Kutner M. H., Nachtsheim C. J., Neter J., and Li W., 2005., *Applied Linear Statistical Model, Firth Edition.* McGraw-Hill.

Lawrence R., 1997. *Using Neural Networks to Forecast Stock Market Prices.* University of Manitoba.

Lawrence S. M. Gamst G. and Guarino A. J., 2006. *Applied Multivariate Research: Design and Interpretation*. Sage Publications Inc. London.

Liu X., 2009. *A Statistical Analysis of Key Factors Influencing the Location of Biomass-using Facilities*. University of Tennessee. Knoxville.

Matthew O. and Odularu G. O., 2009. *The impact of share market capitalization on a company's performance: A case study in the Nigerian confectionary industry*. African Journal of Business Management Vol.3 (5).

Mekonnen B., 2011. *Identification and Statistical Analysis of Factors Affecting Use of Contraceptive Methods among Ethiopian Women*. Addis Ababa University. Ethiopia.

Midani A., 1991. *Determinants of Kuwaiti Stock Prices: An Empirical Investigation of Industrial Services, and Food Company Share*s. Journal of Administrative Sciences and Economics, 1991, Vol. 2, Pages 314-303

Nishat M. and Irfan C.M., 2003. *Dividend Policy and Stock Price Volatility in Pakistan*. 11th Pacific Basin Finance, Economics and Accounting Conference.

Prempeh E. A., (2009*). Comparative Study of the Logistic Regression Analysis and the Discriminant Analysis*. University of Cape Coast.

Rana S. Midi H. and Sarkar S. K., 2010. *Validation and Performance Analysis of Binary Logistic Regression Model*. University Putra. Malaysia.

Sarkar S.K. Midi H. and Rana S., 2010. *Model Selection in Logistic Regression and Performance of its Predictive Ability*. Australian Journal of Basic and Applied Sciences. Malaysia.

Schumaker R. P. and Chen H., 2006. *A Discrete Stock Price Prediction Engine Based on Financial News*. California, USA.

Senol D., 2008. *Prediction of Stock Price Direction by Artificial Neural Network Approach.* Bogaziçi University.

Sharma S., 1996. *Applied Multivariate Techniques*. Canada: John Wiley & Sons, Inc.

Sharma S., 2011. *Determinants of Equity Share Prices in India*. Researchers World' Journal of Arts, Science & Commerce.

Thompson G. K., 2009. *Logistic Regression for the Modeling of Low-Dose Radiation Effects*. Ottawa, Canada.

http://en.wikipedia.org/wiki/Stock – accessed on 18/02/2012

http://en.wikipedia.org/wiki/List_of_stock_exchanges - accessed on 18/02/201

http://en.wikipedia.org/wiki/Logistic_regression - accessed on 18/02/2013

Appendix A: SPSS Enter Method Output

Case Processing Summary			
Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1090	60.1
	Missing Cases	0	.0
	Total	1090	60.1
Unselected Cases		724	39.9
Total		1814	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding	
Original Value	Internal Value
No Increase	0
Increase	1

Block 0: Beginning Block

Classification Table[a,b]

	Observed		Predicted					
			Model Fitting Sample			Validation Sample		
			Change in Share Price		Percentage Correct	Change in Share Price		Percentage Correct
			No Increase	Increase		No Increase	Increase	
Step 0	Change in Share Price	No Increase	0	451	.0	0	282	.0
		Increase	0	639	100.0	0	442	100.0
	Overall Percentage				58.6			61.0

a. Constant is included in the model.

b. The cut value is .500

c. Model Fitting Sample Approximately 60% of the cases (SAMPLE) EQ 1

d. Validation Sample Approximately 60% of the cases (SAMPLE) NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.348	.061	32.100	1	.000	1.417

Variables not in the Equation					
			Score	df	Sig.
Step 0	Variables	Assets / capital employed	.108	1	.742
		Debt / assets	.121	1	.728
		Debt / equity	2.156	1	.142
		Dividend yield	6.249	1	.012
		Earnings / share	5.579	1	.018
		Earnings yield	1.285	1	.257
		Operating profit margin	3.546	1	.060
		Price earnings	.915	1	.339
		Return on assets	4.597	1	.032
		Return on equity	.030	1	.863
		Return on capital employed	13.826	1	.000
	Overall Statistics		49.827	11	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	163.778	11	.000
	Block	163.778	11	.000
	Model	163.778	11	.000

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1314.694[a]	.140	.188

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	19.896	8	.011

Contingency Table for Hosmer and Lemeshow Test						
		Change in Share Price = No Increase		Change in Share Price = Increase		Total
		Observed	Expected	Observed	Expected	
Step 1	1	93	89.195	16	19.805	109
	2	68	60.238	41	48.762	109
	3	54	52.004	55	56.996	109
	4	50	48.081	59	60.919	109
	5	50	45.068	59	63.932	109
	6	37	41.800	72	67.200	109
	7	24	37.837	85	71.163	109
	8	24	33.959	85	75.041	109
	9	30	27.389	79	81.611	109
	10	21	15.428	88	93.572	109

Classification Table[a]								
	Observed		Predicted					
			Selected Cases[b]			UnModel Fitting Sample		
			Change in Share Price		Percentage Correct	Change in Share Price		Percentage Correct
			No Increase	Increase		No Increase	Increase	
Step 1	Change in Share Price	No Increase	163	288	36.1	85	197	30.1
		Increase	57	582	91.1	55	387	87.6
	Overall Percentage				68.3			65.2

a. The cut value is .500

b. Model Fitting Sample Approximately 60% of the cases (SAMPLE) EQ 1

c. Validation Sample Approximately 60% of the cases (SAMPLE) NE 1

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Assets / capital employed	-.021	.069	.095	1	.758	.979
	Debt / assets	-.128	.116	1.223	1	.269	.880
	Debt / equity	.007	.005	1.864	1	.172	1.007
	Dividend yield	-.068	.016	17.547	1	.000	.934
	Earnings / share	.000	.000	6.583	1	.010	1.000
	Earnings yield	-.001	.001	3.510	1	.061	.999
	Operating profit margin	.000	.000	.232	1	.630	1.000
	Price earnings	.000	.000	.578	1	.447	1.000
	Return on assets	-.009	.007	2.107	1	.147	.991
	Return on equity	.001	.000	3.995	1	.046	1.001
	Return on capital employed	.049	.007	45.463	1	.000	1.050
	Constant	.225	.157	2.065	1	.151	1.253

a. Variable(s) entered on step 1: Assets / capital employed, Debt / assets, Debt / equity, Dividend yield, Earnings / share, Earnings yield, Operating profit margin, Price earnings, Return on assets, Return on equity, Return on capital employed.

Appendix B: Forward Selection

Case Processing Summary			
Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1090	60.1
	Missing Cases	0	.0
	Total	1090	60.1
Unselected Cases		724	39.9
Total		1814	100.0
a. If weight is in effect, see classification table for the total number of cases.			

Dependent Variable Encoding	
Original Value	Internal Value
No Increase	0
Increase	1

Block 0: Beginning Block

Classification Table[a,b]								
	Observed		Predicted					
			Model Fitting Sample			Validation Sample		
			Change in Share Price		Percentage Correct	Change in Share Price		Percentage Correct
			No Increase	Increase		No Increase	Increase	
Step 0	Change in Share Price	No Increase	0	451	.0	0	282	.0
		Increase	0	639	100.0	0	442	100.0
	Overall Percentage				58.6			61.0
a. Constant is included in the model.								
b. The cut value is .500								
c. Model Fitting Sample Approximately 60% of the cases (SAMPLE) EQ 1								
d. Validation Sample Approximately 60% of the cases (SAMPLE) NE 1								

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.348	.061	32.100	1	.000	1.417

Block 1: Method = Forward Stepwise (Conditional)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	113.482	1	.000
	Block	113.482	1	.000
	Model	113.482	1	.000
Step 2	Step	33.475	1	.000
	Block	146.957	2	.000
	Model	146.957	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1364.990[a]	.099	.133
2	1331.516[a]	.126	.170

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	15.899	8	.044
2	16.026	8	.042

Contingency Table for Hosmer and Lemeshow Test

		Change in Share Price = No Increase		Change in Share Price = Increase		Total
		Observed	Expected	Observed	Expected	
Step 1	1	89	82.266	20	26.734	109
	2	66	55.575	43	53.425	109
	3	54	50.579	55	58.421	109
	4	47	47.450	62	61.550	109
	5	46	45.001	63	63.999	109
	6	34	42.154	75	66.846	109
	7	33	39.419	76	69.581	109
	8	26	35.950	83	73.050	109
	9	31	31.433	78	77.567	109
	10	25	21.174	84	87.826	109
Step 2	1	91	88.177	18	20.823	109
	2	68	58.806	41	50.194	109
	3	53	51.328	56	57.672	109
	4	51	47.571	58	61.429	109
	5	50	44.410	59	64.590	109

	6	35	41.194	74	67.806	109
	7	26	37.714	83	71.286	109
	8	27	34.079	82	74.921	109
	9	27	29.167	82	79.833	109
	10	23	18.556	86	90.444	109

Classification Table[a]

	Observed		Predicted					
			Selected Cases[b]			UnModel Fitting Sample		
			Change in Share Price		Percentage Correct	Change in Share Price		Percentage Correct
			No Increase	Increase		No Increase	Increase	
Step 1	Change in Share Price	No Increase	131	320	29.0	75	207	26.6
		Increase	39	600	93.9	39	403	91.2
	Overall Percentage				67.1			66.0
Step 2	Change in Share Price	No Increase	148	303	32.8	86	196	30.5
		Increase	54	585	91.5	55	387	87.6
	Overall Percentage				67.2			65.3

a. The cut value is .500

b. Model Fitting Sample Approximately 60% of the cases (SAMPLE) EQ 1

c. Validation Sample Approximately 60% of the cases (SAMPLE) NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Return on capital employed	.037	.004	68.431	1	.000	1.038
	Constant	.031	.077	.159	1	.690	1.031
Step 2[b]	Dividend yield	-.063	.016	16.261	1	.000	.939
	Return on capital employed	.044	.005	81.044	1	.000	1.045
	Constant	.165	.087	3.611	1	.057	1.179

a. Variable(s) entered on step 1: Return on capital employed.

b. Variable(s) entered on step 2: Dividend yield.

Model if Term Removed[a]

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	Return on capital employed	-739.325	113.660	1	.000
Step 2	Dividend yield	-682.632	33.748	1	.000
	Return on capital employed	-735.135	138.755	1	.000

a. Based on conditional parameter estimates

Appendix C: Backward Elimination

Case Processing Summary			
Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1090	60.1
	Missing Cases	0	.0
	Total	1090	60.1
Unselected Cases		724	39.9
Total		1814	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding	
Original Value	Internal Value
No Increase	0
Increase	1

Block 0: Beginning Block

Classification Table[a,b]								
	Observed		Predicted					
			Model Fitting Sample			Validation Sample		
			Change in Share Price		Percentage Correct	Change in Share Price		Percentage Correct
			No Increase	Increase		No Increase	Increase	
Step 0	Change in Share Price	No Increase	0	451	.0	0	282	.0
		Increase	0	639	100.0	0	442	100.0
	Overall Percentage				58.6			61.0

a. Constant is included in the model.

b. The cut value is .500

c. Model Fitting Sample Approximately 60% of the cases (SAMPLE) EQ 1

d. Validation Sample Approximately 60% of the cases (SAMPLE) NE 1

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.348	.061	32.100	1	.000	1.417

Block 1: Method = Backward Stepwise (Conditional)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	163.778	11	.000
	Block	163.778	11	.000
	Model	163.778	11	.000
Step 2[a]	Step	-.094	1	.759
	Block	163.684	10	.000
	Model	163.684	10	.000
Step 3[a]	Step	-.222	1	.638
	Block	163.462	9	.000
	Model	163.462	9	.000
Step 4[a]	Step	-.865	1	.352
	Block	162.597	8	.000
	Model	162.597	8	.000
Step 5[a]	Step	-1.567	1	.211
	Block	161.030	7	.000
	Model	161.030	7	.000
Step 6[a]	Step	-2.048	1	.152
	Block	158.982	6	.000
	Model	158.982	6	.000
Step 7[a]	Step	-1.392	1	.238
	Block	157.590	5	.000
	Model	157.590	5	.000
Step 8[a]	Step	-2.455	1	.117
	Block	155.135	4	.000
	Model	155.135	4	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1314.694[a]	.140	.188
2	1314.789[a]	.139	.188
3	1315.011[b]	.139	.188
4	1315.876[b]	.139	.187
5	1317.443[a]	.137	.185
6	1319.491[a]	.136	.183
7	1320.883[a]	.135	.181
8	1323.338[a]	.133	.179

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	19.896	8	.011
2	24.112	8	.002
3	22.990	8	.003
4	23.156	8	.003
5	20.545	8	.008
6	23.222	8	.003
7	22.841	8	.004
8	26.499	8	.001

Classification Table[a]

	Observed		Predicted					
			Selected Cases[b]			UnModel Fitting Sample		
			Change in Share Price		Percentage Correct	Change in Share Price		Percentage Correct
			No Increase	Increase		No Increase	Increase	
Step 1	Change in Share Price	No Increase	163	288	36.1	85	197	30.1
		Increase	57	582	91.1	55	387	87.6
	Overall Percentage				68.3			65.2
Step 2	Change in Share Price	No Increase	165	286	36.6	85	197	30.1
		Increase	58	581	90.9	57	385	87.1
	Overall Percentage				68.4			64.9
Step 3	Change in Share Price	No Increase	165	286	36.6	86	196	30.5
		Increase	59	580	90.8	58	384	86.9
	Overall Percentage				68.3			64.9
Step 4	Change in Share Price	No Increase	164	287	36.4	85	197	30.1
		Increase	58	581	90.9	56	386	87.3
	Overall Percentage				68.3			65.1
Step 5	Change in Share Price	No Increase	164	287	36.4	87	195	30.9
		Increase	56	583	91.2	56	386	87.3
	Overall Percentage				68.5			65.3
Step 6	Change in Share Price	No Increase	166	285	36.8	90	192	31.9
		Increase	61	578	90.5	59	383	86.7
	Overall Percentage				68.3			65.3
Step 7	Change in Share Price	No Increase	165	286	36.6	90	192	31.9
		Increase	60	579	90.6	60	382	86.4
	Overall Percentage				68.3			65.2
Step 8	Change in Share Price	No Increase	161	290	35.7	89	193	31.6
		Increase	55	584	91.4	59	383	86.7
	Overall Percentage				68.3			65.2

a. The cut value is .500

b. Model Fitting Sample Approximately 60% of the cases (SAMPLE) EQ 1
c. Validation Sample Approximately 60% of the cases (SAMPLE) NE 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Assets / capital employed	-.021	.069	.095	1	.758	.979
	Debt / assets	-.128	.116	1.223	1	.269	.880
	Debt / equity	.007	.005	1.864	1	.172	1.007
	Dividend yield	-.068	.016	17.547	1	.000	.934
	Earnings / share	.000	.000	6.583	1	.010	1.000
	Earnings yield	-.001	.001	3.510	1	.061	.999
	Operating profit margin	.000	.000	.232	1	.630	1.000
	Price earnings	.000	.000	.578	1	.447	1.000
	Return on assets	-.009	.007	2.107	1	.147	.991
	Return on equity	.001	.000	3.995	1	.046	1.001
	Return on capital employed	.049	.007	45.463	1	.000	1.050
	Constant	.225	.157	2.065	1	.151	1.253
Step 2[a]	Debt / assets	-.127	.114	1.243	1	.265	.881
	Debt / equity	.007	.005	1.883	1	.170	1.007
	Dividend yield	-.069	.016	17.663	1	.000	.934
	Earnings / share	.000	.000	6.544	1	.011	1.000
	Earnings yield	-.001	.001	3.492	1	.062	.999
	Operating profit margin	.000	.000	.215	1	.643	1.000
	Price earnings	.000	.000	.575	1	.448	1.000
	Return on assets	-.009	.006	2.069	1	.150	.991
	Return on equity	.001	.000	3.933	1	.047	1.001
	Return on capital employed	.049	.007	45.687	1	.000	1.050
	Constant	.195	.123	2.516	1	.113	1.216
Step 3[a]	Debt / assets	-.109	.106	1.049	1	.306	.897
	Debt / equity	.007	.005	1.875	1	.171	1.007
	Dividend yield	-.069	.016	17.635	1	.000	.934
	Earnings / share	.000	.000	6.499	1	.011	1.000
	Earnings yield	-.001	.001	3.426	1	.064	.999
	Price earnings	.000	.000	.580	1	.446	1.000
	Return on assets	-.008	.006	1.860	1	.173	.992
	Return on equity	.001	.000	3.853	1	.050	1.001
	Return on capital employed	.048	.007	46.163	1	.000	1.050
	Constant	.177	.117	2.302	1	.129	1.193
Step 4[a]	Debt / assets	-.107	.106	1.014	1	.314	.899
	Debt / equity	.007	.005	1.859	1	.173	1.007

	Dividend yield	-.069	.016	17.645	1	.000	.934
	Earnings / share	.000	.000	6.525	1	.011	1.000
	Earnings yield	-.001	.001	3.442	1	.064	.999
	Return on assets	-.008	.006	1.807	1	.179	.992
	Return on equity	.001	.000	3.834	1	.050	1.001
	Return on capital employed	.048	.007	46.118	1	.000	1.050
	Constant	.178	.117	2.337	1	.126	1.195
Step 5[a]	Debt / equity	.007	.005	1.772	1	.183	1.007
	Dividend yield	-.068	.016	17.340	1	.000	.934
	Earnings / share	.000	.000	6.630	1	.010	1.000
	Earnings yield	-.001	.001	3.478	1	.062	.999
	Return on assets	-.007	.006	1.298	1	.255	.993
	Return on equity	.001	.000	3.710	1	.054	1.001
	Return on capital employed	.048	.007	42.120	1	.000	1.049
	Constant	.119	.093	1.616	1	.204	1.126
Step 6[a]	Debt / equity	.007	.005	1.734	1	.188	1.007
	Dividend yield	-.068	.016	17.579	1	.000	.934
	Earnings / share	.000	.000	6.484	1	.011	1.000
	Earnings yield	-.001	.001	3.412	1	.065	.999
	Return on equity	.001	.000	3.124	1	.077	1.001
	Return on capital employed	.042	.005	74.571	1	.000	1.043
	Constant	.091	.090	1.019	1	.313	1.095
Step 7[a]	Debt / equity	.006	.004	1.616	1	.204	1.006
	Dividend yield	-.068	.016	17.345	1	.000	.935
	Earnings / share	.000	.000	6.516	1	.011	1.000
	Earnings yield	-.001	.001	3.358	1	.067	.999
	Return on capital employed	.043	.005	73.970	1	.000	1.043
	Constant	.094	.090	1.103	1	.294	1.099
Step 8[a]	Dividend yield	-.068	.016	17.628	1	.000	.934
	Earnings / share	.000	.000	6.468	1	.011	1.000
	Earnings yield	-.001	.001	3.398	1	.065	.999
	Return on capital employed	.042	.005	73.620	1	.000	1.043
	Constant	.117	.088	1.758	1	.185	1.124

a. Variable(s) entered on step 1: Assets / capital employed, Debt / assets, Debt / equity, Dividend yield, Earnings / share, Earnings yield, Operating profit margin, Price earnings, Return on assets, Return on equity, Return on capital employed.